4

STATISTICAL APPROACH TO TESTS INVOLVING PHYLOGENIES

Susan Holmes

This chapter reviews statistical testing involving phylogenies. We present both the classical framework with the use of sampling distributions involving the bootstrap and permutation tests and the Bayesian approach using posterior distributions.

We give some examples of direct tests for deciding whether the data support a given tree or trees that share a particular property, comparative analyses using tests that condition on the phylogeny being known are also discussed.

We introduce a continuous parameter space that enables one to avoid the delicate problem of comparing exponentially many possible models with a finite amount of data. This chapter contains a review of the literature on parametric tests in phylogenetics and some suggestions of non-parametric tests. We also present some open questions that have to be solved by mathematical statisticians to provide the theoretical justification of both current testing strategies and as yet underdeveloped areas of statistical testing in non-standard frameworks.

4.1 The statistical approach to phylogenetic inference

From our point of view, as statisticians, we see the phylogenetic inference as both estimation and testing problems that are set in an unusual space. In most standard statistical theory, the parameter space is either the real line $\mathbb{R}$ or an Euclidean space of higher dimension, $\mathbb{R}^d$ for instance. One notable exception for which there are a number of available statistical models and tests are ranked data. These sit in the symmetric group $\mathfrak{S}_n$ of permutations of n elements. See [61] for a book long treatment on statistics in such spaces, see [15] for some examples of data and relevant statistical analyses based on decompositions of the space, and [27] on the use of distances and their applications in that context. Of course other relevant high dimensional parameters that statisticians use are probability distributions themselves (non-parametric statistics). The authors of [16] use them to show conditions on consistency for Bayes estimates. Thus, as opposed to some authors in systematics, statisticians actually do believe that both distributions and trees can be true parameters. Although some references [4, 79, 83] do not agree with this approach, we will confer the status of true parameters to both the branching pattern or “tree topology,” that we will denote by τ and the rooted binary tree with edge lengths and n leaves denoted $\mathcal{T}_n$. The inner edge lengths are often denoted $\theta_1, \ldots, \theta_{n-2}$ and considered **nuisance** parameters. One of the difficulties in manipulating such parameters is the lack of a natural ordering of trees.

The main focus here will be the subject of hypothesis testing using phylogenies, the method chosen to estimate these phylogenies is not the focus, so that much of what is discussed is relevant whether we use maximum likelihood (MC), parsimony- or distance-based estimates. We will review the different paradigms, frequentist and Bayesian and emphasize their different approaches to the question of testing a hypothesis H_0 (either composite or simple) versus either a simple alternative H_1 or a set of alternatives $\mathcal{H}_{\mathcal{A}}$. We cannot cover many interesting aspects of the discussion between proponents of both perspectives and refer the reader to an extensive literature on the general subject of frequentist versus Bayesian approaches [6, 7, 50]. We will not go as far as a discussion of finding the best tests for each situation but will insist more on correct tests. The reader interested in the more sophisticated statistical theory of **uniformly most powerful tests** is referred to [52]. A serious attempt at applying the statistical theory of most powerful tests to model selection was made recently by [4]. We will comment on his findings, but insist that statistical tests should be able to adjust to cases where the evolutionary model is unknown or misspecified. Thus in Section 4.4 we concentrate on proposing non-parametric alternatives to existing tests.

Section 4.2 will give the statistical terminology and present some of the issues involved in statistical testing, the meaning of p-values and their comparison to Bayesian alternatives in the context of tests involving phylogenetic trees and the classical approaches to comparing tests. Section 4.3 concentrates on certain tests already in use by the community, with emphasis on their assumptions. Section 4.4 introduces a geometric interpretation of current problems in phylogeny, and proposes a non-parametric multivariate approach. Finally in the conclusion we note how many theoretical aspects of hypothesis testing remain unresolved in the phylogenetic setting. Most papers justify their results by analogy [22, 72] or by simulation [85]. To be blunt, apart from Refs. [12, 13] and [67] there are practically no statistical theorems justifying current tests used in systematics literature and this area is a wide open field for further researchers interested in the interface between multivariate statistics and geometry.

4.2 Hypotheses testing

For background on classical hypothesis tests [71] is a clear elementary introduction and [52] is an encyclopedic account.

4.2.1 *Null and alternative hypotheses*

We will consider tests of a null hypothesis H_0, usually a statement involving an unknown parameter. For example, $\mu = \mu_0$, where μ_0 is a predefined value, such as 4 for a real valued parameter (a **simple hypothesis**), or of the type H_0: $\mu \in \mathcal{M}$, with $\mathcal{M}$ a subset of the parameters, this is a **composite hypothesis**. The **alternative** is usually defined by the complementary set $\overline{\mathcal{M}}$: H_A: $\mu \in \mathcal{M}$. In the case of the Kishino–Hasegawa (KH) test [34] for instance the parameter of interest is the difference in log **likelihoods** of the two trees to be compared $\delta = \log L(D \mid \mathcal{T}_1) - \log L(D \mid \mathcal{T}_2)$ (for an extensive discussion of likelihood computations in the context of phylogenetic trees see Chapter 2, this volume). This difference δ in much of the literature, suggesting that this is the parameter of interest, however there is already slippage of the classical paradigm here since the parameter involves the data, so the definition of the exact parameter that is being tested in the KH test is unclear.

4.2.2 *Test statistics*

Suppose for the moment that H_0 is simple. Given some observed Data $D = \{x_1, x_2, \ldots, x_n\}$, it is often impossible to test the hypothesis directly by asking whether the p-value $P(D \mid H_0)$ is small, so we will use some feature of the data, or test statistic S such that the distribution of this test statistic under the null hypothesis (the **null sampling distribution**) is known. Thus, if the observed value of S is denoted s, $P(s \mid H_0)$ can be computed. We call $P(D \mid H_0)$ as it varies with the data D the sampling distribution, the quantity $P(D \mid H)$ as a function of H is called the **likelihood** of H for the data D.

Some authors [4] identify trees with distributions, this is possible supposing a fixed Markovian evolutionary model and verification of certain identifiability constraints [12]. Thus, the parameters of interest become the distributions and a test for whether the k topologies forming $\mathcal{M}_k = \{\tau_1, \tau_2, \ldots, \tau_k\}$ are equidistant from topology h is stated using the Kullback–Leibler distance between distributions [4].

In this survey, we also encourage the use of a distance between trees, but have tried to enlarge our outlook to encompass more general evolutionary models so that we no longer have the identification between trees and distributions. Not all test statistics are created equal, and in the case of the bootstrap it is always better to have a **pivotal test statistic** [23], that is a statistic whose distribution does not depend on unknown parameters. For this reason, it is preferable to centre and rescale the statistic so that the null distribution is centred at 0 and has a known variance, at least asymptotically.

4.2.3 *Significance and power*

Statisticians take into account two kinds of error:

Type I error or Significance This is the probability of rejecting a hypothesis when in fact it is true.

Type II error or (1-Power) This is the probability of not rejecting a hypothesis that is in fact false.

Usually the type I error is fixed at a given level, say 0.05 or 0.01 and then we might explore ways of making the type II error as small as possible, this is equivalent to maximizing what is known as the power function: the probability of rejecting the null hypothesis H_0 given that the alternative is true $P(\text{reject}H_0 \mid H_A)$. We often use the rejection region R to denote the values of the test statistic s that lead to rejection, for a one-sided test H_A: $\mu > \mu_0$ at the 5% level the rejection region will be given by a half line of the form $[c_{0.95}, +\infty)$, where $c_{0.95}$ is the 95th percentile of the distribution of the test statistic under the null hypothesis.

The power of the test depends on the alternative H_A which can sometimes be defined as $\mu \in \overline{\mathcal{M}}$, then the power function written as a function of the rejection region is

$$P(S(D) \in R \mid \mu \in \overline{\mathcal{M}}).$$

Trying to find tests that are powerful against all alternatives (Uniformly Most Powerful, UMP) is not realistic unless we can use parametric distributions such as exponential families for which there is a well understood theory [52]. In the absence of analytical forms for the power functions, authors [4] are reduced to using simulation studies to compute the power function. In general the power will be a function of many things: the variability of the sampling distribution, the difference between the true parameter and the null hypothesis. In the case of trees, a power curve is only possible if we can quantify this difference with a distance between trees. Aris-Brosou [4] uses the Kullback–Leibler distance. As a substitute for the more general non-parametric setup, we suggest using a geometrically defined distance.

Parametric tests use a specific distributional form of the data, non-parametric tests are valid no matter what the distribution of the data are. Tests are said to be robust when their conclusions remain approximately valid even if the distributional assumptions are violated. Ref. [4] shows in his careful power function simulations that the tests he describes are not robust.

Classical statistical theory (in particular the Neyman Pearson lemma) ensures that the most powerful test for testing one simple hypothesis H_0 versus another H_A is the likelihood ratio test based on the test statistic $S = P(D \mid H_0)/P(D \mid H_A)$.

Frequentists define the p-value of a test as the probability

$$P(S(D) \in \mathcal{S} \mid H_0),$$

where $\mathcal{S}$ is the random region constructed as the values of the statistic “as extreme as” the observed statistic $S(D)$, the definition of the region $\mathcal{S}$ depends also on the alternative hypothesis H_A, for instance for a real valued test statistic S and a two-sided alternative, $\mathcal{S}$ will be the union of two disjoint half lines bounded by what are called the **critical points**, for a one-sided alternative, $\mathcal{S}$ will only be a half line. If we prespecify a type I error to be α, we can define a **rejection region** R_α for the statistic $S(D)$ such that

$$P(S(D) \in R_\alpha \mid H_0) = \alpha.$$

We reject the null hypothesis H_0 if the observed statistic S is in the rejection region. This makes the link between confidence regions and hypothesis tests which are often seen as dual of each other. The confidence region for a parameter μ is a region $\mathcal{M}_\alpha$ such that

$$P(\mathcal{M}_\alpha \ni \mu) = 1 - \alpha.$$

The usual image useful in understanding the reasoning behind the notion of confidence regions (and very nicely illustrated in the Cartoon Guide to Statistics [31]) is the archer and her target. If we know the precision with which the archer hits the target in the sense of the distribution of her arrows in the large circle. We can use it if we are standing behind the target to go back from a single arrow head seen at the back (where the target is invisible and all we see is a square bale of hay) to estimating where we think the centre was.

In particular, if we are lucky enough to have a sampling distribution with a lot of symmetry, we can look at the centre of the sampling distribution and find a good estimate of the parameter and hypothesis testing through the dual confidence region statement is easy.

For the classical hypothesis testing setup to work at all, there are many procedural rules that have to be followed. The main one concerns the order in which the steps are undertaken:

- State the null hypothesis.
- State the alternative.
- Decide on a test statistic and a significance level (Type I error).
- Compute the test statistic for the data at hand.
- Compute the probability of such a value of the test statistic under the null hypothesis (either analytically or through a bootstrap or permutation test simulation experiment).
- Compare this probability (or p-value, as it is called) to the type I error that was pre-specified, if the p-value is smaller than the preassigned type I error, reject the null hypothesis.

In looking at many published instances, it is surprising how often one or more of these steps are violated, in particular it is important to say whether the trees involved in the testing statements are specified prior to consulting the data or not. Data snooping completely invalidates the conclusions of tests that do not account for it (see [30] for a clear statement in this context).

There are ways of incorporating prior information in statistical analyses, these are known as Bayesian methods.

4.2.4 *Bayesian hypothesis testing*

I will not go into the details of Bayesian estimation as the reader can consult Yang, Chapter 3, this volume, who has an exhaustive treatment of Bayesian estimation for phylogenetics in a parametric context. Bayesian statisticians have a completely different approach to hypothesis testing. Parameters are no longer fixed, but are themselves given distributions. Before consulting the data, the parameter is said to have a prior distribution, from which we can actually write statements such as $P(H_0)$ or $P(\tau \in \mathcal{M})$, which would be meaningless in the classical context. After consulting the data D, the distributions becomes restricted to the conditional $P(H_0 \mid D)$ or $P(\tau \in \mathcal{M} \mid D)$.

The most commonly used Bayesian procedure for hypothesis testing is to specify a prior for the null hypothesis, H_0, say for instance with no bias either way, one conventionally chooses $P(H_0) = 0.5$ [50].

Bayesian testing is based on the ratio (or posterior odds)

$$\frac{P(H_0 \mid D)}{P(\overline{H}_0 \mid D)} = \frac{P(D \mid H_0)}{P(D \mid \overline{H}_0)} \times \frac{P(H_0)}{P(\overline{H}_0)}$$

to decide whether the hypothesis H_0 should be rejected, the first ratio on the right is called the Bayes factor; it shows how the prior odds $P(H_0)/P(\overline{H}_0)$ are changed to the posterior odds, if the Bayes factor is small, the null hypothesis is rejected. It is also possible to build sets B with given posterior probability levels: $P(\tau \in B \mid D) = 0.99$, these are called Bayesian **credibility sets**. A clear elementary discussion of Bayesian hypothesis testing is in Chapter 4 of [50].

An example of using the Bayesian paradigm for comparing varying testing procedures in the phylogenetic context can be found in [3]. The author proposes two tests. One compares models two by two using Bayes factors $P(D \mid \mathcal{T}_i)/P(D \mid \mathcal{T}_j)$ and suggests that if the Bayes factor is larger than 100, the evidence is in favour of $\mathcal{T}_i$. However, in real testing situations the evidence is often much less clear cut. In a beautiful example of Bayesian testing applied to the "out-of-Africa" hypothesis, Huelsenbeck and Imennov [46] show cases where the Bayes factor equal to 4.

Another test also proposed by Aris-Brosou [3] uses an average

$$\int_{\mathcal{T},\Omega} p(D \mid \mathcal{T}, \theta) \frac{dP(\mathcal{T}, \theta)}{p(D \mid \mathcal{T}_i)}$$

for which there is not an exact statement of existence as yet, as integration over treespace is undefined. However by restricting himself to a finite number of trees to compare with, this average can be defined using counting measure. Of course the main advantage in the Bayesian approach is the possibility of integrating out all the nuisance parameters, either analytically or by MCMC simulation (see Chapter 3, this volume for details).The software [49] provides a way of generating sets of trees under two differing models and thus some tests can use distances between the distributions of trees under competing hypotheses and the posterior distribution given the data.

4.2.5 *Questions posed as function of the tree parameter*

In all statistical problems, questions are posed in terms of unknown parameters for which one wants to make valid inferences. In the current presentation, our parameter of choice is a semi-labelled binary tree. Sometimes the parameter itself appears in the definition of the null hypothesis,

H_0: The true phylogenetic tree topology τ belongs to a set of trees $\mathcal{M}$.

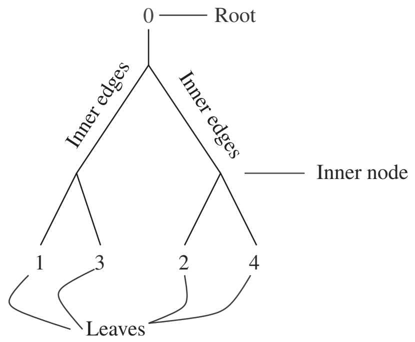

FIG. 4.1. The tree parameter is rooted with labelled leaves and inner branches

For instance the set of trees containing a given clade, or a specific set of trees $\mathcal{M} = \{\tau_1, \tau_2, \ldots, \tau_k\}$ as in Ref. [4].

The parameter space is not a classical Euclidean space, thus introducing the need for many non-standard techniques. The discrete parameter defined as the branching order of the binary rooted tree with n leaves, τ, can take on one of $(2n-3)!!$ values [73] (where $(2n-3)!! = (2n-3)\times(2n-5)\times(2n-7)\times\cdots\times 3\times 1$). $\mathcal{T}_n$ is the branching pattern with the $n-2$ inner branch lengths often considered as nuisance parameters $\theta_1, \theta_2, \ldots, \theta_{n-2}$, left unspecified by H_0 (the pendant edges are sometimes fixed by a constraining normalization of tree so that all the leaves are contemporary). Even for simple hypotheses, the **power function** of the test varies with all the parameters, natural and nuisance. This is resolved by using the standard procedure of setting the nuisance parameters, for example, the edge lengths at their maximum likelihood estimates (MLEs).

We consider rooted trees as in Fig. 4.1 because in most phylogenetic studies, biologists are careful to provide outgroups that root the tree with high certainty, this brings down the complexity of the problem by a factor of n, which is well worth while in practical problems.

The first step is often to estimate the parameter τ by $\hat{\tau}$ computed from the data. In the case of parsimony estimation τ represents a branching order, without edge lengths, however, we can always suppose that in this case the edge lengths are the number of mutations between nodes, the general parameter we will be considering will have edge lengths.

In what follows we will consider our parameter space to be partitioned into regions, each region dedicated to one particular branching order $\hat{\tau}$, estimation can thus be pictured as projecting the data set from the data space into a point $\hat{\tau}$ in the parameter space.

The geometrical construction in Ref. [9] makes this picture more precise. The regions become cubes in dimension $(n-2)$ and the boundary regions are lower dimensional. The first thing to decide when making such a topological construction, is what is the definition of a neighbourhood? Our construction is based on a notion of proximity defined by biologists as nearest neighbour interchange (NNI) moves [54, 81] (also called **Rotation Moves** [78] by combinatorialists), other notions of proximity are also meaningful, in the context of host–parasite comparisons [48] one should use other possible elementary *moves* between neighbouring trees.

This construction enables us to define distances between trees, for both the branching order and the edge enriched trees. With the existence of a distance we are able to define neighbourhoods as balls with a given radius. We will use this distance in much of what follows, but nothing about this distance is unique and many other authors have proposed distances between trees [69].

The boundaries between regions represent an area of uncertainty about the exact branching order, represented by the middle tree in Fig. 4.2. In biological terminology this is called an "unresolved" tree. Biologists call "polytomies" nodes of the tree with more than two branches. These appear as lower dimensional "cube-boundaries" between the regions.

For example, the boundary for trees with three leaves is just a point (Fig. 4.3), while the boundaries between two quadrants in treespace for $n = 4$ are segments (Fig. 4.4).

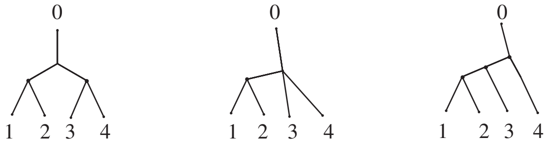

FIG. 4.2. Nearest neighbour interchange (NNI) move, an inner branch becomes zero, then another grows out

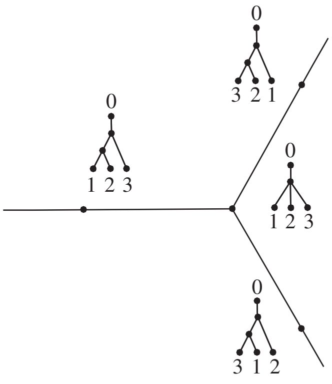

FIG. 4.3. The space of edge enriched trees with three leaves is the union of three half lines meeting at the star tree in the centre, if we limit ourselves to trees with bounded inner edges, the space is the union of three segments of length 1

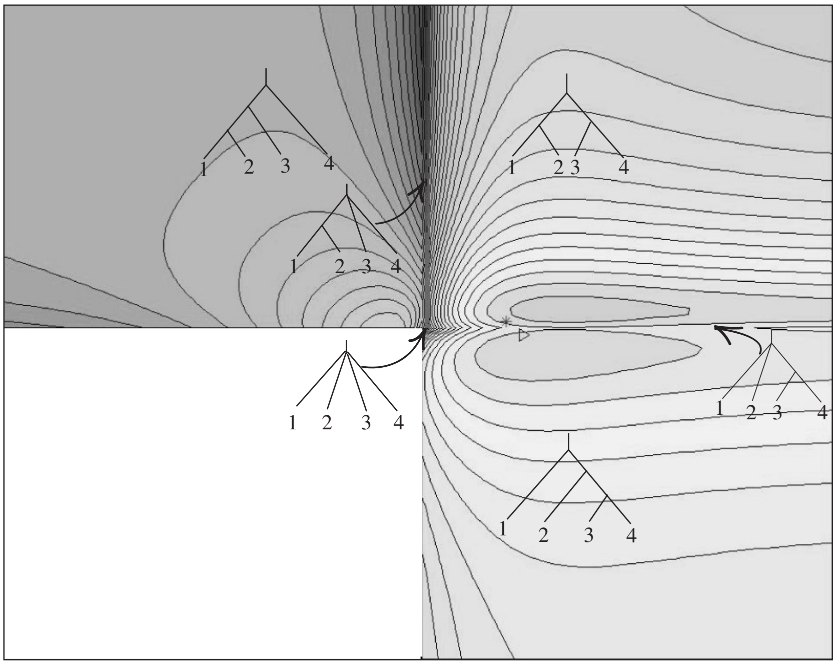

FIG. 4.4. A small part of the likelihood surface mapped onto three neighbouring quadrants of treespace, each quadrant represents one branching order among 15 possible ones for 4 leaves, the true tree that was used to simulate the data is represented as a star close to the horizontal boundary

4.2.6 *Topology of treespace*

Many intuitive pictures of treespace have to be revised to incorporate some of its non-standard properties. Many authors describe the landscape of trees as a real line or plane [14], with the likelihood function as an alternative pattern of mountains and valleys, thus if the sea level rises, islands appear [59].

Figure 4.4 is a representation of the likelihood of a tree with four leaves over only 3 of the 15 possible quadrants for data that was generated according to a true tree with one edge very small compared to the other, we see how the phenomenon of "islands" can occur, we also see how hard it would be to make such a representation for trees with many leaves.

This lacks one essential specificity of treespace: it is not embeddable in such a Euclidean representation because it wraps around itself. Billera *et al.* [9] describe this by defining the *link of the origin* in the following way: all 15 regions corresponding to the 15 possible trees for $n = 4$ share the same origin, we give coordinates to each region according to the edge lengths of their two inner branches, this make each region a square if the tree is constrained to have finite edge lengths. If we take the diagonal line segment $x + y = 1$ in each quadrant, we obtain a graph with an edge for each quadrant and a trivalent vertex for each boundary ray; this graph is called the *link of the origin*. In the case of 4 leaves, we obtain a well-known graph called the Peterson graph, and in higher dimensions, extensions to what we could call Peterson simplices. One of the redeeming properties of treespace as we have described it is that if a group of trees share several edges we can ignore those dimensions and only look at the subspace composed of the trees without these common edges, thus decreasing the dimension of the relevant comparison space.

The wraparound has important consequences for the MCMC methods based on NNI moves, since a wraparound will ensure a speedup in convergence as compared to what would happen a Euclidean space.

The main property of treespace as proved in Ref. [9] is that it is a CAT(0) space, succintly this can be rephrased in the more intuitive fact that *triangles are thin in treespace.* Mathematical details may be found in Ref. [9]: the most important consequences are being a CAT(0) space ensures the existence of convex hulls and distances in treespace [32].

To picture how distances are computed in treespace, Fig. 4.5 shows paths between A and B and between C and D, the latter passes through the star tree and is a cone path that can always be constructed by making all edges zero and then growing the new edges, the distance between two points in tree space is

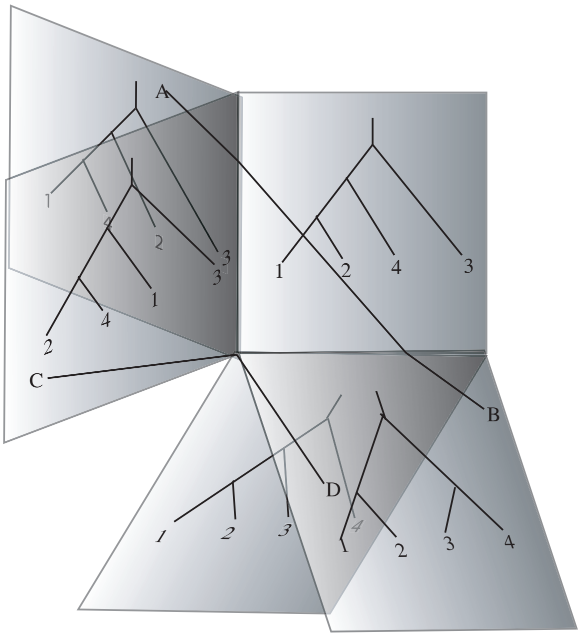

FIG. 4.5. Five of the fifteen possible quadrants corresponding to trees with four leaves and two geodesic paths in treespace, in fact each quadrant contains the star tree and has two other neighbouring quadrants

computed as the shortest path between the points that stays in treespace, thus the geodesic path between A and B does not pass through the star tree. This computation can be intractable, but in real cases, the problem splits down and the distance can be computed in reasonable time [43].

4.2.7 *The data*

The data from which the tree is often estimated are usually matrices of aligned characters for a set of n species.

The data can be:

- Binary, often coming from morphological characters

```
Lemur_cat  00000000000001010100000
Tarsius_s  10000010000000010000000
Saimiri_s  10000010000001010000000
Macaca_sy  00000000000000010000000
Macaca_fa  10000010000000010000000
```

- Aligned:

```
  6  40
Lemur_cat AAGCTTCATA GGAGCAACCA TTCTAATAAT CGCACATGGC
Tarsius_s AAGTTTCATT GGAGCCACCA CTCTTATAAT TGCCCATGGC
Saimiri_s AAGCTTCACC GGCGCAATGA TCCTAATAAT CGCTCACGGG
Macaca_sy AAGCTTCTCC GGTGCAACTA TCCTTATAGT TGCCCATGGA
Macaca_fa AAGCTTCTCC GGCGCAACCA CCCTTATAAT CGCCCACGGG
Macaca_mu AAGCTTTTCT GGCGCAACCA TCCTCATGAT TGCTCACGGA
```

- Gene order (see the Chapters 9 to 13, this volume for some examples).

An important property of the data is that they come with their own metrics. There is a meaningful notion of proximity for two data sets, whether the data are permutations, Amino Acid or DNA sequences. One of the points we want to emphasize in this chapter is that we often have less data than actually needed given the multiplicity of choices we have to make when making decisions involving trees. Most statistical tests in use suppose that the columns of the data (characters) are independent. In fact we know that this is not true, and in highly conserved regions there are strong dependencies between the characters. There is thus much less information in the data than meets the eye. The data may contain 1000 characters, but be equivalent only to 50 independent ones.

4.2.8 *Statistical paradigms*

The algorithms followed in the classical *frequentist* context are:

- Estimate the parameter (either in a parametric (ML) way, semiparametric (Distance-based methods), or non-parametric way (Parsimony)).
- Find the sampling distribution of the estimator under the null.

On the other hand Bayesians follow the following procedure

- Specify a Prior Distribution for the parameter.
- Update the Prior using the Data.
- Compute the Posterior Distribution.

Both use the result of the last steps of their procedures to implement the Hypothesis tests. Frequentists use the estimate and the sampling distribution of the tree parameter to do tests, whether parametric or non-parametric. This is the distribution of the estimates $\hat{\tau}$ when the data are drawn at random from their parent population.

In the case of complex parameters such as trees, no analytical results exist about these sampling distributions, so that the Bootstrap [20, 23] is often employed to provide reasonable approximations to such unknown sampling distributions.

Bayesians use the posterior distribution to compute estimates such as the mode of the posterior (MAP) estimator or the expected value of the posterior and to compute Bayesian credibility regions with given level. More important is the fact that usually Bayesians assign a prior probability to the null hypothesis, such as $P(H_0) = 1/2$ and using this prior and the data can compute $P(H_0 \mid \text{Data})$. This computation is impossible in the frequentist context, only computations based on the sampling distribution are allowed.

4.2.9 *Distributions on treespace*

As we see, in both paradigms the key element is the construction of either the sampling distribution or the posterior distribution, both distributions in treespace. We thus need to understand distributions on treespace. If we had a probability density f over treespace, we could write statements such as eqn (3) in Aris-Brosou [4] that integrates the likelihood $\ell(\theta, \mathcal{T} \mid D)$ over a subset of trees $\mathfrak{T}$:

$$h_{0,f} = \int_{\mathfrak{T}} \ell(\theta, \mathcal{T} \mid D) df(\mathcal{T}).$$

This allows the replacement of a composite null hypotheses of equality of a set of trees by an integrated simple hypotheses as suggested by Lehmann's [52] adaptation of the Bayesian procedure. The integral is undefined unless we have such a probability distribution on treespace.

The basic example of a distribution on treespace that we would like to summarize is the sampling distribution, that we will now define in more detail. Suppose the data comes from a distribution $\mathcal{F}$, and that we are given many such data sets, as shown in Fig. 4.6. Estimation of the tree from the data provides a projection onto treespace for each of the data sets, thus we obtain many estimates $\hat{\tau}_k$.

We need to know what this true "theoretical" sampling distribution is in order to build confidence statements about the true parameter.

The true **sampling distribution** is usually inaccessible, as we are not given many sets of data from the distribution $\mathcal{F}$ with which to work. Figure 4.7 shows

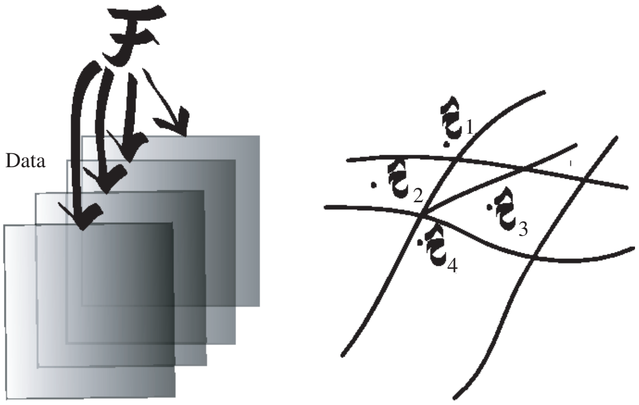

FIG. 4.6. The true sampling distribution lies in a non-standard parameter space

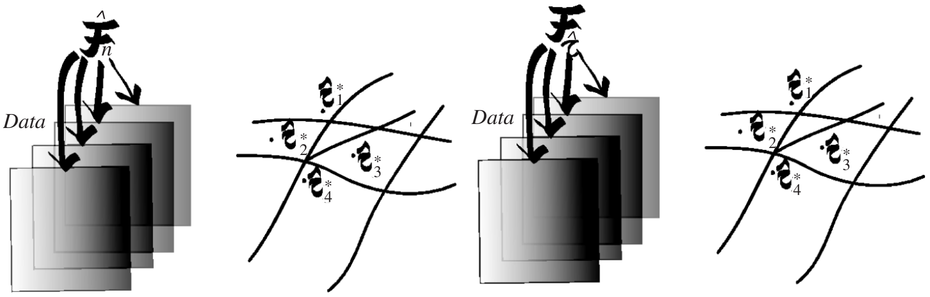

FIG. 4.7. Bootstrap sampling distributions: non-parametric (left), parametric (right)

how the non-parametric bootstrap replaces $\mathcal{F}$ with the **empirical distribution** $\hat{\mathcal{F}}_n$, new data sets are "plausible" perturbations of the original, drawn from the empirical cumulative distribution instead of the unknown $\mathcal{F}$. Data are created by drawing repeatedly from the empirical distribution given from the original data, for each new data set a new tree $\hat{\tau}_k^*$ is estimated, and thus there is a simulated sampling distribution computed by using the multinomial reweighting of the original data [23]. Note that even if we generate a large number of resamples, the bootstrap resampling distribution cannot overcome the fact that it is only an approximation built from one data set. It is actually possible to give the complete bootstrap sampling distribution without using Monte Carlo at all [17], nonetheless the bootstrap remains an approximation as it replaces the unknown distribution $\mathcal{F}$ by the empirical distribution constructed from one sample.

If the data are known to come from a parametric distribution with an unknown parameter such as the edge-weighted tree $\mathcal{T}$, the parametric distribution produces simulated data set by supposing the estimate from the original data $\hat{\mathcal{T}}$ is the true estimate and generating the data from that model as indicated by the right side of Fig. 4.7. This means generating many data sets by simulating sequences from the estimated tree following the Markovian model of evolution.

However, given the large number of possible trees and the small amount of information, both these methods may have problems finding the sampling distribution if it is not simplified. If we consider the simplest possible distribution on trees, we will be using the uniform distribution, however, there are an exponentially growing number of trees. This leads to paradoxes such as the blade of grass type argument [68]: if we consider the probability of obtaining a tree τ_0 we will have conclusions such as $P(\hat{\tau} = \tau_0) = 1/(2n-3)!!$ this becomes exponentially small very quickly, making for paradoxical statements.[1]

Overcoming the discreteness and size of the parameter space. If one wanted to use a sample of size 100 to infer the most likely of 10,000 possible choices, one would need to borrow strength from some underlying structure. Thinking of the choices as boxes that can be ordered in a line with proximity of the boxes being meaningful shows that we can borrow information from "neighbouring" boxes. We will see as we go along that the notion of neighbouring trees is essential to improving our statistical procedures.

We can imagine creating useful features for summarizing the distribution or treespace (either Bayesian posterior or Bootstrapped sampling distributions).

The most common summary in use is the presence or absence of a clade. As compared to the original tree, this would be a vector of length $n-2$. If we just wanted to invent all the clades in the data, the number of possible clades is the number of bipartitions where both sets have at least 2 leaves. The complete feature vector in that case would be a vector of length $2^{n-1}-n-1$. This multidimensional approach can be followed through by doing an analysis of the data as if it were a contingency table and we could keep statements of the kind "clade (1,2) is always present when clade (4,5) is present" thus improving on the basic confidence values currently in use.

Other features might be incorporated into an exponential distribution such as Mallows' model [60] that was originally implemented for ranked data

$$P(\tau_i) = Ke^{-\lambda d(\tau_i,\tau_0)},$$

as described in Ref. [40]. This distribution uses a central tree τ_0 and a distance d in treespace. Mallows model would work well if we had strong belief in a very symmetrical distribution around a central tree. In reality this does not seem to be the case, so a more intricate mixture model would be required. One could imagine having the mixture of two underlying trees which might have biological meaning. Other distributions of interest are extensions of the Yule process (studied by Aldous [1]) or exponential families incorporating information

[1] After choosing a blade of grass in a field, one cannot ask, what were the chances of choosing this blade? With probability one, I was going to choose one [19].

about the estimation method used. The reason for doing this is that Gascuel [29] has shown the influence of the estimation method chosen (parsimony, maximum likelihood, or distance based) on the shape of the estimated tree. We could build different exponential families running through certain important parameters such as "balance", or tree width as studied by evolutionary biologists who use the concept of tree shape (see [37, 63, 65]).

Some methods for comparing trees measure some property of the data with regards to the tree, such as the minimum number of mutations along the tree to produce the given data (the parsimony score) or the probability of the data given a fixed evolutionary model with parameters $\alpha_1, \alpha_2, \ldots, \alpha_k$ and a fixed tree

$$P(D \mid \mathcal{T}_n, \alpha) = L(\mathcal{T}_n).$$

This, considered as a function of $\mathcal{T}_n$ defines the likelihood of $\mathcal{T}_n$. Sometimes this is replaced by the likelihood of a branching pattern τ maximized and the branch lengths $\theta_1, \ldots, \theta_{2n-2}$ are chosen to maximize the likelihood.

The lack of a natural ordering in the parameter space encourages the use of simpler statistical parameters. The presence/absence of a given clade, a confidence level, a distance between trees are all acceptable simplifying parameters as we will see. This multiplicity of riches is something that also occurs in other areas of statistics, for instance when choosing between a multiplicity of graphical models. In that domain, researchers use the notion of "features" characterizing shared aspects of subsets of models.

For one particular observed value, say 1.8921 of a real-valued statistic it is meaningless to ask what would the probability $P(Y = 1.8921)$ be equal to, but we can ask the probability of Y belonging to a neighbourhood around the value 1.8921. The definition of features enables the definition of meaningful neighbourhoods of treespace if the features can be defined by a continuous function from treespace to feature space. This has another advantage, as explained in Ref. [9] the parameter space is not embedded in either the real line $\mathbb{R}$ nor an euclidean space such as $\mathbb{R}^d$, on the other hand we can choose the features to be real valued.

Returning to testing, one of the problems facing a biologist is that natural comparisons are not nested within each other. Efron [21] carries out a geometrical analysis of the problem of comparing two non-nested simple linear models, and the analysis is already quite intricate. When comparing a small number of models, the number of parameters grows, but the degrees of freedom remain manageable. Yang *et al.* [83] already noticed that comparing tree parameters is akin to model comparison. However, in this case the number of available models (the trees) increases exponentially with the number of species and the data will never be sufficient to choose between them. Classical model comparison methods such as the AIC and BIC cannot be applied in their vanilla versions here. We have exponentially many trees to choose from, and in the absence of a "continuum" and an underlying statistic providing a natural ordering of the models, we will be unable to use even a large data set to compare the multiplicity of possibilities. (Think of trying to choose between 1 million pictures when only a thousand samples from them exist.)

There is, however, a solution. If we think of each model as a box, each with an unknown probability, if the sampling distribution throws K balls into the boxes and K is much smaller than the number of boxes, then we cannot conclude. However, if we have a notion of neighbourhood boxes, we can borrow strength from the neighbouring boxes.

Remember in this image, that if the balls correspond to the trees obtained by a Bootstrap resample, we cannot increase indefinitely the number of balls and hope to fill all possible boxes. The non-parametric Bootstrap cannot provide more information than is available in the sample.

The classical statistical location summary in the case of trees would be the **mean** and the **median**, and thus we could use the Bootstrap to estimate bias as in Ref. [8]. The notion of **mean** (centre of the distribution as defined using an integral of the underlying probability distribution) supposes that we already have a probability distribution defined on treespace and know how to integrate. These are currently open problems. Associated to this view of a "centre" of a distribution of trees, we can ask the question: *What distribution is the "majority rule consensus" a centre of?*. This would enable more meaningful statistical inferences using the consensii that biologists so often favour. The median, another useful location statistic, can be defined by either of the various multivariate extensions of the univariate median to the multivariate median (in particular Tukey's multivariate median [80]), which we revisit in the multivariate section below.

Usually the best results in hypothesis testing are obtained by using a statistic that is centred and rescaled like the t-statistic, by dividing it by its sampling variance, here this cannot be defined. By analogy we can suppose that it is beneficial to divide by a similar statistic, for instance $\{E_{P_n} d^2(\hat{\tau}, \tau)\}^{-1/2}$ (where d is a distance defined on tree space and E_{P_n} is the expectation with regards to an underlying distribution P_n) is an ersatz-standard deviation.

4.3 Different types of tests involving phylogenies

There are two main types of statistical testing problems involving phylogenies. First, tests involving the tree parameter itself of the form $P(\tau \in \mathcal{M})$ the second type are tests that treat the phylogenetic tree as a nuisance parameter and will be treated in the second paragraph.

4.3.1 *Testing τ_1 versus τ_2*

The Neyman Pearson theorem ensures that the case of a parametric evolutionary Markovian model the likelihood ratio test as introduced as the Kishino–Hasegawa [34] test will be the most powerful for comparing two prespecified trees. A very clear discussion of the case where one combinatorial tree τ_1 is compared to an alternative τ_2 is given in Ref. [30]. In particular the authors explain how important the assumption that the trees were specified prior to seeing the data. The problem of both estimating and testing a tree with the same data is a more complicated problem and needs adjustments for multiple comparisons as carried out by Shimodaira and Hasegawa [76]. It is definitely the case that the use of the same data to estimate and test a tree is an incorrect procedure.

The use of the non-parametric bootstrap when comparing trees where a satisfactory evolutionary model is known (and may have been used in the estimation of the trees τ_1 and τ_2 to be compared) is not a coherent strategy as the most powerful procedure is to keep the parametric model and use this to generate the resampled data using the parametric bootstrap as implemented by `seqgen` [70] for instance.

4.3.2 *Conditional tests*

Another class of hypothesis tests are those included in what is commonly known as the *Comparative Method* [33, 62]. In this setting, the phylogenetic tree is a nuisance parameter and the interest is in the distribution of variables conditioned on the tree being given. For instance if we wanted to study a morphological trait but substract the variability that can be explained by the phylogenetic relationship between the species, we may (following Ref. [26]), condition on the tree and make a Brownian motion model of the variation of a variable on the tree. More recently, ([44, 57]) propose another parametric model, akin to an ordinary linear mixed model. The variability is decomposed into heritable and residual parts, quantifying the residuals conditions out of the phylogenetic information.

Some recent work enables incorporation of incomplete phylogenetic information [45] providing a way of conducting such tests in a parametric setup where the phylogeny is not known. It would also be interesting to have a Bayesian equivalent of this procedure that could enable the incorporation of some information about the tree we want to condition on, without knowing it exactly.

4.3.3 *Modern Bayesian hypothesis testing*

The Bayesian outlook in hypothesis testing is as yet underdeveloped in the phylogenetic literature but the availability of posterior distributions through Monte Carlo Markov chain (MCMC) algorithms makes this type of testing possible in a rigid parametric context [55, 64, 84]. Useful software have been made available in Refs. [49, 51]. Biologists wishing to use these methods have to take into account the main problem with MCMC (see the review in Ref. [47]):

1. We don't know how long the algorithms have to run to reach stationarity, the only precise theorems [2, 18, 74] have studied very simple symmetric methods, without any Metropolis weighting.
2. Current procedures are based on a restrictive Markovian model of evolution; no study of the robustness of these methods to departure from the Markovian assumptions is available.

One large open question in this area is how to develop non-parametric or semiparametric priors for Bayesian computations in cases where the Markovian model is not adequate. One possibility is to use both the information on the tree shape that is provided both by the estimation method and the phylogenetic noise level [36, 38].

4.3.4 *Bootstrap tests*

I have explained in detail elsewhere [40] some of the caveats to the interpretation of bootstrap support estimates as actual confidence values in the sense of hypothesis testing. If we wanted to test only one clade in the tree, we could consider the existence of this clade as a Bernoulli 0/1 variable and try to estimate it through the plug in principle by using the Bootstrap [25], however, if the model used for estimating the tree is the Markovian Model, we should use the parametric bootstrap, generating new data through simulation from the estimated tree [70]. Using the multinomial non-parametric bootstrap would be incoherent. This procedure allows the construction of one confidence value that can be interpreted on its own. However, two common extensions to this are invalid. If we want to have confidence values on all clades at once, we will be entering the realm of multiple testing: we are using the same data to make confidence statements about different aspects of the data, and statistical theory [53] is very clear about the inaccuracies involved in reporting all the numbers at once on the tree.

We cannot reconstruct the complete bootstrap resampling distribution from the numbers on the clades, this is because these numbers taken together do not form a sufficient statistic for the distribution on treespace (this is discussed in detail in Ref. [41]).

Finally, we cannot compare bootstrap confidence values from one tree to another. This is due to the fact that the number of alternative trees in the neighbourhood of a given tree with a pre-specified size is not always the same. Zharkikh and Li [85] already underlined the importance of taking into account that a given tree may have k alternatives and through simulation experiments, asked the relevant question: *How many neighbours for a given tree?* In fact, through combinatorics developed in Ref. [9]'s continuum of trees (see Section 4.1), we know the number of neighbours of each tree in a precise sense. In this geometric construction each tree topology τ_n with n leaves is represented by a cube of dimension $n - 2$ each dimension representing the inner edge lengths which are supposed to be bounded by one. Figure 4.8 shows the neighbourhoods of two such trees with four leaves. Each quadrant will have as its origin the star tree, which is not a true binary tree since all its inner edges have lengths zero. The point on the left represents a tree with both inner edges close to 1, and only has as neighbours, trees with the same branching order. The point on the right has one of its edges much closer to zero, so has two other different branching orders (combinatorial trees) in its neighbourhood.

For a tree with only two inner edges, there is the only one way of having two edges small: to be close to the origin-star tree and thus the tree is in a neighbourhood of 15 others. This same notion of neighbourhood containing 15 different branching orders applies to all trees on as many leaves as necessary but who have

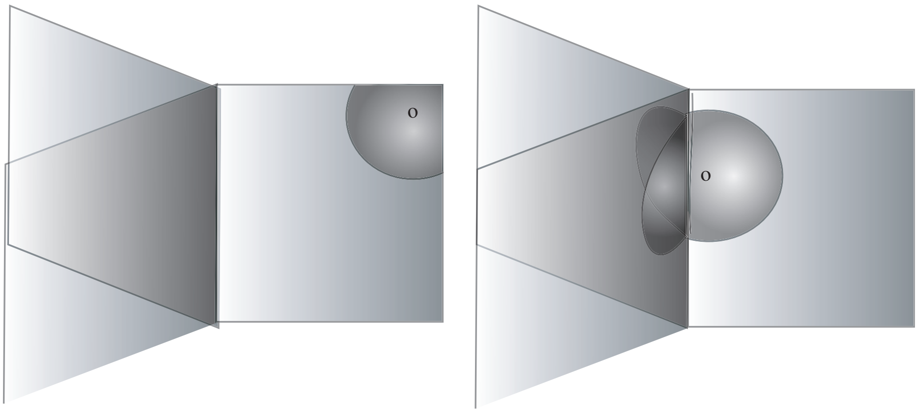

FIG. 4.8. A one tree neighbourhood and a three tree neighbourhood

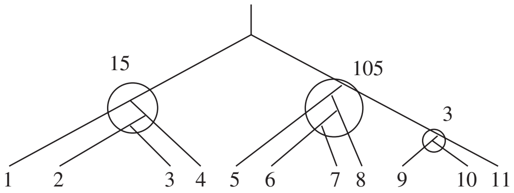

FIG. 4.9. Finding all the trees in a neighbourhood of radius r, each circle shows a set of contiguous edges smaller than r, from left to right we see subtrees with 2, 3, and 1 inner edge respectively

two contiguous "small edges" and all the other inner edges significantly bigger than 0.

This picture of treespace frees us from having to use simulations to find out how many different trees are in a neighbourhood of a given radius r around a given tree. All we have to do is check the sets of contiguous small edges in the tree (say, smaller than r), for example, if there is only one set of size k, then the neighbourhood will contain $(2k-3)!!$ different branching orders (combinatorial trees).

The circles represented in Fig. 4.9 show how all edges smaller than this radius r define the contiguous edge sets. On the left there are two small contiguous edges, in the middle there are three small contiguous edges and on the right there is only one, underneath each disjoint contiguous set, we have counted the number of trees in the neighbourhood of this contiguous set. Here we have three contiguous components, thus a product of three factors for the overall number of neighbours.

In this case the number of trees within a radius r will be the product of the tree numbers $15 * 105 * 3 = 4725$. In general: If there are m sets of contiguous

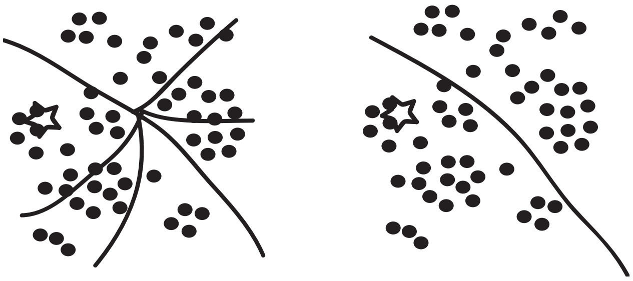

FIG. 4.10. Bootstrap sampling distributions with different neighbourhoods

edges of sizes $(n_1, n_2, \ldots, n_m)$ there will be

$$(2n_1 - 3)!! \times (2n_2 - 3)!! \times (2n_3 - 3)!! \cdots \times (2n_m - 3)!!$$

trees in the neighbourhood. A tree near the star tree at the origin will have an exponential number of neighbours. This explosion of the volume of a neighbourhood at the origin provides for interesting mathematical problems that have to be solved before any considerations about consistency (when the number of leaves increases) can be addressed.

Figure 4.10 aims to illustrate the sense in which just simulating points from a given tree (star) and counting the number of simulated points in the region of interest may not directly inform one on the distance between the true tree (star) and the boundary. The boundary represents the cutoff from one branching order to another, thus separating treespace into regions, each region represents a different branching order. If the distribution were uniform in this region, we would be actually trying to estimate the distance to the boundary by counting the stars in the same region, it is clear that the two different configurations do not give the same answer, whereas the distances are the same. In general we are trying to estimate a weighted distance to the boundary where the weights are provided by the local density.

These differing number of neighbours for different trees show that the bootstrap values cannot be compared from one tree to another. Again, we encounter the problem that "p-values" depend on the context and cannot be compared across studies. This was implicitly understood by Hendy and Penny in their NN Bootstrap procedure (personal communication).

In any classical statistical setup, p-values suffer from a lack of "qualifying weights", in the sense that this one number summary, although on a common scale does not come with any information on the actual amount of information that was used to obtain it. Of course this is a common criticism of p-values by Bayesians (for intricate discussions of this important point see [5, 7, 75], for a textbook introduction see [50]). This has to be taken into account here, as the amount of information available in the data is actually insufficient to conclude in a refined way [66]. For once, there are theorems providing bounds to the amount of precision (the size of the tree) that can be inferred from a given data set (see Chapter 14, this volume). Thus, we should be careful not to provide the equivalent of 15 significant digits for a mean computed with 10 numbers, spread around 100 with a standard deviation of 10 (in this case the standard error would be around 3, so even one significant digit is hubris).

4.4 Non-parametric multivariate hypothesis testing

There is less literature on testing in the non-parametric context; Sitnikova *et al.* [77] who provide interior branch tests and some authors who have permutation test for Bremer [10] support for parsimony trees. In order to be able to design non-parametric tests, we have to leave the realm of the reliance on a molecular clock or even a Markovian model for evolution and explore non-parametric or semiparametric distributions on treespace. To do this we will use the analogies provided from non-parametric multivariate statistics on Euclidean spaces.

4.4.1 *Multivariate confidence regions*

There is an inherent duality in statistics between hypothesis tests and confidence regions for the relevant parameter

$$P(\mathcal{M}_\alpha \ni \tau) = 1 - \alpha.$$

The complement to $\mathcal{M}_\alpha$ provides the rejection region of the test. It is important to note that in this probabilistic statement, the frequentist interpretation is that the region $\mathcal{M}_\alpha$ is random, built from the random variables observed as data and that the parameter has a fixed unknown value τ. For a fixed region B and a parameter τ, the statement $P(\tau \in B)$ is meaningless for a frequentist. Bayesians have a natural way of building such regions, often called **credibility regions** as they have randomness built into the parameters through the posterior distribution, so finding the region that covers $1 - \alpha$ of the posterior probability is quite amenable once the posterior has been either calculated or simulated. However, all current methods for such calculations are based on the parametric Markovian evolutionary model. If we are unsure of the validity of the Markovian model (or absence of a molecular clock, for an example with an unbeatable title see [82]), we can use symmetry arguments leading to semiparametric or non-parametric approaches.

We have found that there are several important questions to address when studying the properties of tests based on confidence regions [22]. One concerns the curvature of the boundary surrounding a region of treespace; the other the number of different regions in contact with a particular frontier. The latter is answered by the mathematical construction of Ref. [9]. However, although the geometric analysis provided in Ref. [9] does show that the natural geodesic distances and the edges of convex hulls in treespace are negatively curved, exact bounds on the amount of curvature are not yet available.

In order to provide both classical non-parametric and Bayesian non-parametric confidence regions, we will use Tukey's approach of [80] involving the

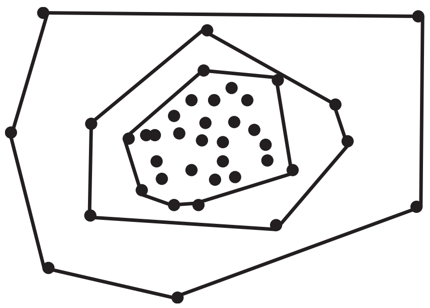

FIG. 4.11. Successive convex hulls built on a scatterplot

construction of regions based on convex hulls. He suggested peelin convex hulls to construct successive “deeper” confidence regions as illustrated in Fig. 4.11. Liu and Singh [56] have developed this for ordering circular data for instance. Here we can use this as a non-parametric method for estimating the “centre” of a distribution in treespace, as well finding central regions holding say 90% of the points.

Example 1 Confidence regions constructed by bootstrapping.

Instead of summarizing the bootstrap sampling distribution by just presence or absence of clades we can explore whether 90% of bootstrap trees are in a specific region in treespace. We can also ask whether the sampling distribution is centred around the star tree, which would indicate that the data does not have strong treelike characteristics.

Such a procedure would be as follows:

- Estimate the tree from the original data call this, t_0.
- Generate K bootstrap trees.
- Compute the 90% convex envelope by peeling the successive hulls, until we have a convex envelope containing 90% of the bootstrap trees call this $C_{0.10}$.
- Look at whether $C_{0.10}$ contains the star tree.
- If it does not, the data are in fact treelike. □

Example 2 Are two data sets congruent, suggesting that they come from the same evolutionary process?

This is an important question often asked before combining datasets[11].

This can be seen as a multidimensional two sample test problem. We want to see if the two bootstrap sampling distributions overlap significantly (A and B). Here we use an extension of the Friedman–Rafsky (FR) [28] test. This method is inspired by the Wald–Wolfowitz test, and solves the problem that there is no natural multidimensional “ordering.” First the bootstrap trees from bootstrapping both data sets are organized into a minimal spanning tree following the classical

Minimal Spanning Tree Algorithm (a greedy algorithm is easy to implement).

- Pool the two bootstrap samples of points in treespace together.
- Compute the distances between all the trees, as defined in Ref. [9].
- Make a minimal spanning ignoring which data set they came from (labels A and B).
- Colour the points according to the data sets they came from.
- Count the number of "pure" edges, that is the number of edges of the minimal spanning tree whose vertices come from the same sample, call this the test statistic S_0, if S_0 is very large, we will reject the null hypothesis that the two data sets come from the same process.
  (An equivalent statistic is provided by taking out all the edges that have mixed colours and counting how many "separate" trees remain.)
- Compute the permutation distribution of S^* by reassigning the labels to the points at random and recomputing the test statistic, say B times.
- Compute the p-value as the ratio

$$\frac{\#\{S_k^* > S_0\}}{B}.$$

This extends to case of more than two data sets by just looking at the distribution of the pure edges as the test statistic. □

Example 3 Using distances between trees to compute the bootstrap sampling distribution.

By computing the bootstrap distribution we can give an approximation to the distribution of $(d(\hat{\mathcal{T}}, \mathcal{T}))$ by $d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}})$. This is a statement by analogy to many other theorems about the bootstrap, nothing has been proved in this context. However, this analogy is very useful as it also suggests that the better test statistic in this case is: $d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}})\{\text{var}(d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}}))\}^{-1/2}$ which should have a near pivotal distribution that provides a good approximation to the unknown distribution of $d(\hat{\mathcal{T}}, \mathcal{T})/\{\text{var}(d(\hat{\mathcal{T}}, \mathcal{T}))\}^{-1/2}$ equivalent of a "studentized" statistic [23]. As can be seen in Fig. 4.12, this distribution is not necessarily Normal, or even symmetric.

Such a sampling distribution can be used to see if a given tree $\mathcal{T}_0$ could be the tree parameter responsible for this data. If the test statistic

$$\frac{d(\hat{\mathcal{T}}, \hat{\mathcal{T}}_0)}{\sqrt{\text{var}(d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}}))}}$$

is within the 95th percentile confidence interval around $\hat{T}$ we cannot reject that it could be the true $\mathcal{T}$ parameter for this data. □

Example 4 Embedding the data into $\mathbb{R}^d$.

Finally a whole other class of multivariate tests are available through an approximate embedding of treespace into $\mathbb{R}^d$. Assume a finite set of trees: it could be

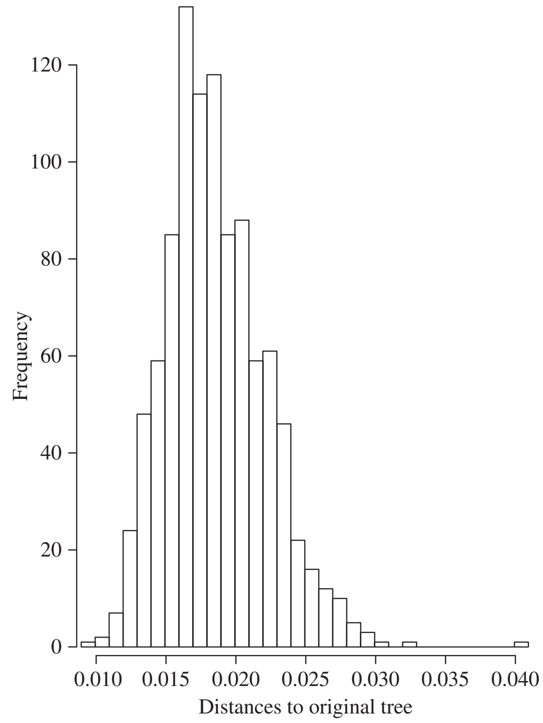

FIG. 4.12. Bootstrap sampling distribution of the distances between the original tree and the bootstrapped trees in the analysis of the Plasmodium F. data analysed in Ref. [22], the distances were computed according to Ref. [9]

a set of trees from bootstrap resamples, it could be a pick from a Bayesian posterior distribution, or sets of trees from different types of data on the same species. Consider the matrix of distances between trees and use a multidimensional scaling algorithm (either metric or non-metric) to find the best approximate embedding of the trees in $\mathbb{R}^d$ in the sense of distance reconstruction. Then we can use all the usual multivariate statistical techniques to analyse the relationships between the trees. The likely candidates are

- discriminant analysis that enables finding combinations of the coordinates that reconstruct prior groupings of the trees (trees made from different data sources, molecular, behavioural, phenotypic for instance)
- principal components that provide a few principal directions of variation
- clustering that would point out if the trees can be seen as a mixture of a few tightly clustered groups, thus pointing to a multiplicity in the underlying evolutionary structure, in this case a mixture of trees would be appropriate (see Chapter 7, this volume). □

4.5 Conclusions: there are many open problems

Much work is yet to be done to clarify the meaning of the procedures and tests already in practice, as well as to provide sensible non-parametric extensions to the testing procedures already available.

Here are some interesting open problems:

- Find a test for measuring how close the data are to being treelike, without postulation of a parametric model, some progress on this has been made by comparing the distances on the data to the closest distance fulfilling the four point condition (see Chapter 7, this volume).
- Find a test for finding out whether the data are a mixture of two trees? This can be done with networks as in Chapter 7, this volume, or it can be done by looking at the posterior distribution (see Yang, Chapter 3, this volume) and finding if there is a evidence of bimodality.
- Find satisfactory probability distributions on treespace that enable simple definitions of non-parametric sampling and Bayesian posterior distributions.
- Find the optimal ways of aggregating trees as either expectations for various measures or modes of these distributions.
- Find a notion of differential in treespace to study the influence functions necessary for robustness calculations.
- Quantify how the departure from independence in most biological data influences the validity of using Bootstrap procedures that assume independence.
- Quantify the amount of information in a given data set and find the equivalent number of degrees of freedom needed to fit a tree under constraints.
- Generalize the decomposition into phylogenetic information and non-heritable residuals to a non-parametric setting.

Acknowledgements

This research was funded in part by a grant from the NSF grant DMS-0241246, I also thank the CNRS for travel support and Olivier Gascuel for organizing the meeting at IHP in Paris and carefully reading my first draft.

I would like to thank Persi Diaconis for discussions of many aspects of this work, Elizabeth Purdom and two referees for reading an early version, Henry Towsner and Aaron Staple for computational assistance, Erich Lehmann and Jo Romano for sending me a chapter of their forthcoming book and Michael Perlman for sending me his manuscript on likelihood ratio tests.

References

[1] Aldous, D.A. (1996). Probability distributions on cladograms. In *Random Discrete Structures* (ed. D.A. Aldous and R. Pemantle), pp. 1–18. Springer-Verlag, Berlin.

[2] Aldous, D.A. (2000). Mixing time for a Markov chain on cladograms. *Combinatorics, Probability and Computing*, **9**, 191–204.

[3] Aris-Brosou, S. (2003*a*). How Bayes tests of molecular phylogenies compare with frequentist approaches. *Bioinformatics*, **19**(5), 618–624.

[4] Aris-Brosou, S. (2003*b*). Least and most powerful phylogenetic tests to elucidate the origin of the seed plants in presence of conflicting signals under misspecified models? *Systematic Biology*, **52**(6), 781–793.

[5] Bayarri, M.J. and Berger, J.O. (2000). P values for composite null models. *Journal of the American Statistical Association*, **95**(452), 1127–1142.

[6] Berger, J.O. and Guglielmi, A. (2001). Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, **96**(453), 174–184.

[7] Berger, J.O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association*, **82**, 112–122.

[8] Berry, V. and Gascuel, O. (1996). Interpretation of bootstrap trees: Threshold of clade selection and induced gain. *Molecular Biology and Evolution*, **13**, 999–1011.

[9] Billera, L., Holmes, S., and Vogtmann, K. (2001). The geometry of tree space. *Advances in Applied Mathematics*, **28**, 771–801.

[10] Bremer, K. (1994). Branch support and tree stability. *Cladistics*, **10**, 295–304.

[11] Buckley, T.R., Arensburger, P., Simon, C., and Chambers, G.K. (2002). Combined data, Bayesian phylogenetics, and the origin of the New Zealand Cicada genera. *Systematic Biology*, **51**, 4–15.

[12] Chang, J. (1996*a*). Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences*, **137**, 51–73.

[13] Chang, J. (1996*b*). Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences*, **134**, 189–215.

[14] Charleston, M.A. (1996). Landscape of trees. `http://taxonomy.zoology.gla.ac.uk/mac/landscape/trees.html`.

[15] Diaconis, P. (1989). A generalization of spectral analysis with application to ranked data. *The Annals of Statistics*, **17**, 949–979.

[16] Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics*, **14**, 1–26.

[17] Diaconis, P. and Holmes, S. (1994). Gray codes and randomization procedures. *Statistics and Computing*, **4**, 287–302.

[18] Diaconis, P. and Holmes, S. (2002). Random walks on trees and matchings. *Electronic Journal of Probability*, **7**, 1–18.

[19] Diaconis, P. and Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association*, **84**, 853–861.

[20] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.

[21] Efron, B. (1984). Comparing non-nested linear models. *Journal of the American Statistical Association*, **79**, 791–803.

[22] Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of National Academy of Sciences USA*, **93**, 13429–13434.

[23] Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

[24] Efron, B. and Tibshirani, R. (1998). The problem of regions. *Annals of Statistics*, **26**(5), 1687–1718.

[25] Felsenstein, J. (1983). Statistical inference of phylogenies (with discussion). *Journal Royal Statistical Society, Series A*, **146**, 246–272.

[26] Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, **125**, 1–15.

[27] Fligner, M.A. and Verducci, J.S. (ed.) (1992). *Probability Models and Statistical Analyses for Ranking Data.* Springer-Verlag, Berlin.

[28] Friedman, J.H. and Rafsky, L.C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, **7**, 697–717.

[29] Gascuel, O. (2000). Evidence for a relationship between algorithmic scheme and shape of inferred trees. In *Data Analysis, Scientific Modeling and Practical Applications* (ed. W. Gaul, O. Opitz, and M. Schader), pp. 157–168. Springer-Verlag, Berlin.

[30] Goldman, N., Anderson, J.P., and Rodrigo, A.G. (2000). Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*, **49**, 652–670.

[31] Gonick, L. and Smith, W. (1993). *The Cartoon Guide to Statistics.* Harper-Row Inc., New York.

[32] Gromov, M. (1987). Hyperbolic groups. In *Essays in Group Theory* (ed. S.M. Gersten), pp. 75–263. Springer, New York.

[33] Harvey, P.H. and Pagel, M.D. (1991). *The Comparative Method in Evolutionary Biology.* Oxford University Press.

[34] Hasegawa, M. and Kishino, H. (1989). Confidence limits on the maximum likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. *Evolution*, **43**, 672–677.

[35] Hayasaka, K., Gojobori, T., and Horai, S. (1988). Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution*, **5/6**, 626–644.

[36] Heard, S.B. and Mooers, A.O. (1996). Imperfect information and the balance of cladograms and phenograms. *Systematic Biology*, **5**, 115–118.

[37] Heard, S.B. and Mooers, A.O. (2002). The signatures of random and selective mass extinctions in phylogenetic tree balance. *Systematic Biology*, **51**, 889–897.

[38] Hillis, D.M. (1996). Inferring complex phylogenies. *Nature*, **383**, 130.

[39] Holmes, S. (1999). Phylogenies: An overview. In *Statistics and Genetics* (ed. E. Halloran and S. Geisser), Springer-Verlag, New York.

[40] Holmes, S. (2003*a*). Bootstrapping phylogenetic trees: Theory and methods. *Statistical Science*, **18**, 241–255.

[41] Holmes, S. (2003*b*). Statistics for phylogenetic trees. *Theoretical Population Biology*, **63**, 17–32.

[42] Holmes, S. and Diaconis, P. (1999). Computing with trees. In *31st Symposium on the Interface: Models, Predictions, and Computing (Interface'99).*

[43] Holmes, S., Staple, A., and Vogtmann, K. (2004). Algorithm for computing distances between trees and its applications. Research Report, Department of Statistics, Stanford, CA 94305.

[44] Housworth, E., Martins, E., and Lynch, M. (2004). Phylogenetic mixed models. *American Naturalist*, **163**, 84–96.

[45] Housworth, E.A. and Martins, E.P. (2001). Conducting phylogenetic analyses when the phylogeny is partially known: Random sampling of constrained phylogenies. *Systematic Biology*, **50**, 628–639.

[46] Huelsenbeck, J.P. and Imennov, N.S. (2002). Geographic origin of human mitochondrial DNA: Accommodating phylogenetic uncertainty and model comparison. *Systematic Biology*, **51**, 155–165.

[47] Huelsenbeck, J.P., Larget, B., Miller, R.E., and Ronquist, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, **51**, 673–688.

[48] Huelsenbeck, J.P., Rannala, B., and Yang, Z. (1997). Statistical tests of host–parasite cospeciation. *Evolution*, **51**, 410–419.

[49] Huelsenbeck, J.P. and Ronquist, F. (2001). MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

[50] Jaynes, E.T. (2003). *Probability Theory: The Logic of Science* (ed. G.L. Bretthorst). Cambridge University Press, Cambridge.

[51] Larget, B. and Simon, D. (2001). Bayesian analysis in molecular biology and evolution. `www.mathcs.duq.edu/larget/bambe.html`.

[52] Lehmann, E.L. (1997). *Testing Statistical Hypotheses.* Springer-Verlag, New York.

[53] Lehmann, E.L. and Romano, J. (2004). *Testing Statistical Hypotheses* (3rd edn). Springer-Verlag, New York.

[54] Li, M., Tromp, J., and Zhang, L. (1996). Some notes on the nearest neighbour interchange distance. *Journal of Theoretical Biology*, **182**, 463–467.

[55] Li, S., Pearl, D.K., and Doss, H. (2000). Phylogenetic tree construction using MCMC. *Journal of the American Statistical Association*, **95**, 493–503.

[56] Liu, R.Y. and Singh, K. (1992). Ordering directional data: Concepts of data depth on circles and spheres. *The Annals of Statistics*, **20**, 1468–1484.

[57] Lynch, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution*, **45**, 1065–1080.

[58] MacFadden, B.J. and Hulbert Jr, R. (1988). Explosive speciation at the base of the adaptive radiation of miocene grazing horses. *Nature*, **336**, 466–468.

[59] Maddison, D.R. (1991). The discovery and importance of multiple islands of most parsimonious trees. *Systematic Zoology*, **40**, 315–328.

[60] Mallows, C.L. (1957). Non-null ranking models. I. *Biometrika*, **44**, 114–130.

[61] Marden, J.I. (1995). *Analyzing and Modeling Rank Data.* Chapman & Hall, London.

[62] Martins, E.P. and Hansen, T.F. (1997). Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*, **149**, 646–667.

[63] Martins, E.P. and Housworth, E.A. (2002). Phylogeny shape and the phylogenetic comparative method. *Systematic Biology*, **51**, 1–8.

[64] Mau, B., Newton, M.A., and Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, **55**, 1–12.

[65] Mooers, A.O. and Heard, S.B. (1997). Inferring evolutionary process from the phylogenetic tree shape. *Quarterly Review of Biology*, **72**, 31–54.

[66] Nei, M., Kumar, S., and Takahashi, K. (1998). The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. *Proceedings of the National Academy of Sciences USA*, **95**, 12390–12397.

[67] Newton, M.A. (1996). Bootstrapping phylogenies: Large deviations and dispersion effects. *Biometrika*, **83**, 315–328.

[68] Penny, D., Foulds, L.R., and Hendy, M.D. (1982). Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. *Nature*, **297**, 197–200.

[69] Penny, D. and Hendy, M.D. (1985). The use of tree comparison metrics. *Systematic Zoology*, **34**, 75–82.

[70] Rambaut, A. and Grassly, N.C. (1997). Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, **13**, 235–238.

[71] Rice, J. (1992). *Mathematical Statistics and Data Analysis.* Duxbury Press, Wadsworth, Belmont, CA.

[72] Sanderson, M.J. and Wojciechowski, M.F. (2000). Improved bootstrap confidence limits in large-scale phylogenies with an example from neo-astragalus (leguminosae). *Systematic Biology*, **49**, 671–685.

[73] Schröder, E. (1870). Vier combinatorische probleme. *Zeitschrift fur Mathematik und Physik*, **15**, 361–376.

[74] Schweinsberg, J. (2001). An O(n^2) bound for the relaxation time of a Markov chain on cladograms. *Random Structures and Algorithms*, **20**, 59–70.

[75] Sellke, T., Bayarri, M.J., and Berger, J.O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, **55**(1), 62–71.

[76] Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, **16**, 1114–1116.

[77] Sitnikova, T., Rzhetsky, A., and Nei, M. (1995). Interior-branch and bootstrap tests of phylogenetic trees. *Molecular Biology and Evolution*, **12**, 319–333.

[78] Sleator, D.D., Tarjan, R.E., and Thurston, W.P. (1992). Short encodings of evolving structures. *SIAM Journal of Discrete Mathematics*, **5**(3), 428–450.

[79] Thompson, E.A. (1975). *Human Evolutionary Trees.* Cambridge University Press.

[80] Tukey, J.W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, Volume 2 (ed. R.D. James), pp. 523–531. Canadian Mathematical Congress, Montreal, Vancouver.

[81] Waterman, M.S. and Smith, T.F. (1978). On the similarity of dendograms. *Journal of Theoretical Biology*, **73**, 789–800.

[82] Wiegmann, B.M., Yeates, D.K., Thorne, J.L., and Kishino, H. (2003). Time flies, a new molecular time-scale for brachyceran fly evolution without a clock? *Systematic Biology*, **52**(6), 745–756.

[83] Yang, Z., Goldman, N., and Friday, A.E. (1995). Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Systematic Biology*, **44**, 384–399.

[84] Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Molecular Biology and Evolution*, **14**, 717–724.

[85] Zharkikh, A. and Li, W.H. (1995). Estimation of confidence in phylogeny: The complete and partial bootstrap technique. *Molecular Phylogenetic Evololution*, **4**, 44–63.