



Permutation Tests for Analysing Cospeciation in Multiple Phylogenies

Lazarus Katana Mramba

Submitted in accordance with the requirements for the degree of

Master of Science in Statistics

The University of Leeds, School of Mathematics

September 2010

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

Abstract

The purpose of this study is to develop permutation test statistics that can be used to analyse cospeciation in three phylogenies. The null hypothesis, H_0 , is that the three phylogenies are not related, indicating that the host species evolved independently from their parasite species. The alternative hypothesis, H_1 , is that there is a close relationship among the three phylogenies which could indicate cospeciation.

Three test statistics have been developed. The first one is the Pearson's partial correlation coefficient, r_* , where $*$ is either $xy.z$, $xz.y$ or $yz.x$. A triangular association matrix is used when computing the observed, r_*^{obs} , and the permuted partial correlation coefficients, r_*^p . To test the significance of r_*^{obs} , three partial p values: P_x , P_y and P_z (for X , Y and Z held constant respectively) are calculated by summing $r_*^p \geq r_*^{obs}$ and dividing by the total number of permutations, N . To conclude on the overall significance of the results, a geometric p value, P_{gm} , is calculated.

The second test statistic is the eigenvalue, λ_r , computed from the correlation matrix based on a principal component analysis (PCA) method. With this statistic, the eigenvalues λ_r are derived from the correlation coefficients of matrix D , where matrix D takes pairs of triangles of relations from the patristic distances to form its rows and the columns are the three trees X , Y and Z . Only the first λ_r corresponding to PC1 is considered since it has the highest variance, is the largest and explains most of the proportion of variation in the data. Similarly, λ_r^{obs} and λ_r^p are calculated as above. To test the significance of the observed eigenvalues, a p value, P_{λ_r} , is computed by summing the number of $\lambda_r^p \geq \lambda_r^{obs}$ and dividing by N .

The third test statistic is the eigenvalue, λ_c , computed using the covariance matrix. The procedure is exactly the same as the one discussed above and λ_c^{obs} and λ_c^p are calculated. A p value (P_{λ_c}) is also computed to test the significance of the results.

Computation of type I error suggests that using PCA correlation structure produces uniformly distributed p values as well as for the covariance structure but not when partial correlation statistic is used. These results are irrespective of the size of the phylogenies. The power to reject H_0 drops very fast as more association triangles are added or substituted when the partial statistic is used at 0.01 significance level for small phylogenies. When large phylogenies are used, the power to reject H_0 is found to be consistently high in all the three statistics.

In conclusion, the permutation method on PCA eigenvalues produces reliable test statistics that can be used to test for cospeciation of multiple phylogenies. The partial test statistics can be used but may give biased results for small phylogenies.

Contents

1	Introduction	1
1.1	Background	1
1.2	Host-Parasite relationships	2
1.3	Background on phylogenetic trees.	5
1.3.1	Properties of phylogenetic trees	5
2	Statistical Background	9
2.1	Correlation Coefficients	9
2.2	Principal Component Analysis	12
2.2.1	Eigenvalues and Eigenvectors	14
2.2.2	PCA using a correlation matrix	16
3	Methods	19
3.1	Test statistics	19
3.1.1	Test statistic for partial correlation	25
3.1.2	Test statistic for PCA	26
3.1.3	Permutations under the Null hypothesis	28
3.1.4	Type I error	30
3.2	Power Simulations	36
3.2.1	First approach: Adding random triangles	37
3.2.2	Second approach: Replacing triangles	40
4	Results	43
4.1	Results under the null hypothesis	43
4.2	Results under the alternative hypothesis	47
4.2.1	Results under a perfect H_1 condition	47
4.2.2	Results for adding random triangles	51
4.2.3	Results for replacing triangles	52
5	Conclusion	55
5.1	Discussion	55
5.2	Further work	57
A	R programs	59
A.1	R-functions	59

A.1.1	The <i>nperm</i> function	59
A.1.2	The <i>simdata</i> function	66
A.1.3	The <i>addtriangles</i> function	68
A.1.4	The <i>replacetriangles</i> function	69
B	Supplementary Tables	71
B.1	Extra tables	71
	Bibliography	72

List of Figures

1.1	Foodweb	3
1.2	Titrophic relationships	3
1.3	Gophers-lice association	4
1.4	A rooted phylogenetic tree.	6
1.5	Unrooted tree	7
1.6	Multidimensional scaling plot	8
3.1	Matrix D for random trees	22
3.2	Plotted random trees	22
3.3	Exploring matrix D under H_0	23
3.4	Matrix D for trees under H_1	24
3.5	Matrix D plots under H_1	24
3.6	Three random phylogenetic trees with 10 tips	29
3.7	An association matrix under H_0	29
3.8	Type I error plots for 10 p values	31
3.9	Type I error plots for 20 p values	32
3.10	Type I error plots for 50 p values on trees with 10 tips	34
3.11	Type I error plots for 10 tips trees	34
3.12	Type I error for 15 tips trees	35
3.13	An association matrix under H_1	36
3.14	Tree topologies under H_1	37
3.15	Trees after adding random triangles	39
3.16	Trees after replacement with random triangles	41
4.1	Pairwise correlation coefficients under H_0	44
4.2	Observed Results under H_0	45
4.3	Density plots under the null hypothesis	46
4.4	Type I error plots for trees with 10 tips	47
4.5	Type I error distributions for 15 tips	48
4.6	Pairwise correlation plots under H_1	49
4.7	Density plots under alternative hypothesis	49
4.8	PCA plots under H_1	50
4.9	Power curves for trees with 10 tips	51
4.10	Power curves for trees with 20 tips	52
4.11	Power curves for trees with 10 tips (substituting triangles)	53

4.12 Power curves for trees with 20 tips (substituting triangles)	54
---	----

List of Tables

2.1	PCA data structure	12
2.2	Principal component loadings.	16
2.3	Principal components from covariance loadings.	16
2.4	Principal components computed from a correlation matrix.	16
3.1	A general table for matrix D	21
3.2	10 p for trees with 10 tips	31
3.3	20 P values for trees with 10 tips	32
3.4	50 P values for trees with 10 tips	33
3.5	Type I and type II errors	36
4.1	Principal components under H_0	44
4.2	Principal components under perfect conditions of H_1	48
B.1	The rotation of the principal components.	71
B.2	Association matrix used under H_0	71
B.3	matrix D	72
B.4	100 p values for trees with 15 tips	73

Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or institution of learning.

In the attached submission I have not presented anyone else's work as my own. Where I have taken advantage of the work of others, I have given full acknowledgment. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the School's policy on mitigation and procedures for the submission of statements and evidence of mitigation.

Signed: _____

Symbols & abbreviations

MDS	Multidimensional scaling
PCA	Principal component analysis
PC	Principal component
r_{xy}	Pearson's pairwise correlation coefficient between x and y
r_{yz}	Pearson's pairwise correlation coefficient between y and z
r_{xz}	Pearson's pairwise correlation coefficient between x and z
$r_{xy.z}^{obs}$	observed partial correlation coefficient between x and y after controlling for z
$r_{xz.y}^{obs}$	observed partial correlation coefficient between x and z after controlling for y
$r_{yz.x}^{obs}$	observed partial correlation coefficient between y and z after controlling for x
$r_{xz.y}^p$	permuted partial correlation coefficient between x and z after controlling for y
$r_{xy.z}^p$	permuted partial correlation coefficient between x and y after controlling for z
$r_{yz.x}^p$	permuted partial correlation coefficient between y and z after controlling for x
P_x	partial p value of y and z after controlling for x
P_z	partial p value of x and y after controlling for z
P_y	partial p value of x and z after controlling for y
P_{gm}	geometric p value of P_z , P_x and P_y
λ_r^{obs}	observed eigenvalues computed from the correlation matrix
λ_c^{obs}	observed eigenvalues computed from the covariance matrix
λ_r^p	permuted eigenvalues computed from the correlation matrix
λ_c^p	permuted eigenvalues computed from the covariance matrix
P_{λ_r}	p values of eigenvalues computed from the correlation matrix
P_{λ_c}	p values of eigenvalues computed from the covariance matrix

Acknowledgments

It is my pleasure to thank everyone who has directly or indirectly made this dissertation a success.

First and foremost, I would like to thank the Almighty God for the gift of life and grace that has brought me this far.

I am indebted to my core supervisor, Professor Walter R. Gilks for his superb statistical ideas and for sharing his vast knowledge of statistics. Special thanks to Kerstin Hommola, a Ph.D. statistics student who has consistently provided support and guidance through out this project.

I owe my deepest gratitude to Dr. Stuart Barber, the MSc statistics coordinator, who has actively and willingly acted as both my project supervisor and unceasingly provided brilliant statistical guidance and Latex support in the whole MSc course and particularly in this project.

It is an honour for me to have had Professor Charles C. Taylor who is a senior statistician in the department of statistics and serving as the head of the school of mathematics as my personal tutor. He has been a cheerful career guider and a source of encouragement and motivation.

This course would not have been successful without funding from Kenya Medical Research Institute-Wellcome Trust, Kilifi, Kenya and without the support of Dr. Gregory Fegan, the current senior statistician of KEMRI-Kilifi.

Last but not least, I would like to thank all my lecturers who have played a major role in shaping my academic life and future aspirations and to my dearest and beloved wife Dorothy and son Gift for their unfailing love and moral support.

Chapter 1

Introduction

1.1 Background

Cospeciation is defined by Page (2003) as the joint speciation of lineages that are known to have an ecological association allowing for either of them to have speciated slightly before or after the other. A typical example is that of a host-parasite association. *Coevolution* may be defined as the joint evolution of any two or more associated organisms and *coadaptation* is a synonym of coevolution meaning the reciprocal adaptation in a group of associated organisms.

When two phylogenetic trees have similar topologies, they are said to be *congruent* and therefore *incongruent* if they have discordance between their topologies (Page, 2003). This idea of congruency has been implicated with cospeciation (Brooks and McLennan, 1991) and that congruent phylogenies signal cospeciation whereas incongruence imply host switching. Also, there is a general consensus derived from Fahrenholz's rule (Fahrenholz, 1913), that parasites diversify together with their hosts. The rule states that parasites' phylogenies reflect hosts' phylogenies. However, complete divergence occurs only when cospeciation is regarded as the only exclusive process (Page, 2003). This is not likely to be the case as there are other biological and environmental processes that occur between host-parasite associations. Thus the incongruence of host-parasite phylogenies would be due to other events besides cospeciation such as parasites switching host lineages, duplication, where the parasites independently speciate from their hosts, extinction, parasites failure to speciate with their hosts, or failing to colonize the descendants of a whole speciating host-lineage (Paterson et al., 1999; Paterson and Gray, 1997; Page, 1990b, 1996b).

The use of phylogenetic trees and DNA or protein sequences in biology is fundamental to the understanding of the evolutionary relationships and is one of the primary driving tools for

describing these associations. New developments in the study of host-parasite phylogenies have given insights into the complexity and necessity of reliable statistical methods that can be used to infer the history of an association between them (Page, 2003). However, reliable statistical test appropriate for assessing cospeciation of more than two parasite-host phylogenies in order to quantify various biological phenomena remain a statistical paradox (Choi and Gomez, 2009).

In parasitology, parasites have long been used to infer their host phylogeny (Klassen, 1992). If these parasites cospeciated with their hosts then the parasite phylogeny reflects the host phylogeny assuming that the rate at which the parasite evolves is lower than that of the host. Thus the parasite will retain some traits that the host lost. In contrary to this ideology, evidence from studies in molecular biology has shown that parasites evolve at a higher rate than their hosts (Hafner et al., 1994; Moran et al., 1995).

Page (2003) points out that a basic test of cospeciation is one that gives a significant similarity between the topologies of host and parasite phylogenies that is not due to chance alone. Questions about the timing of speciation emerge if host-parasite phylogenies are found to be identical.

Trees are used to analyse host-parasite cospeciation because the method is widely applicable to different types of datasets such as data from molecular and morphological designs. Trees can be represented in other formats that is not necessarily phylogeny. For instance, Becerra (1997) used chemical similarity to compare the phylogeny of host plants with their parasite (blepharida beetles) phylogeny because the insects' evolutionary history is more identical to the host chemistry than the host phylogeny.

1.2 Host-Parasite relationships

All organisms in any ecosystem are linked biochemically (Ahmad et al., 2004). A simple relationship is composed of two or three trophic (eating) levels with feeding relationships. Hosts and parasites form complex food webs that span over several trophic levels. Figure 1.1 shows an example. A subset of the complex food web is shown in figure 1.2 with triangular associations.

An association is said to be ditrophic when the species of one host are associated with that of one parasite. One common example is the gopher-lice association. A simplified relationship

¹<http://picsdigger.com/image/f29fec67/>

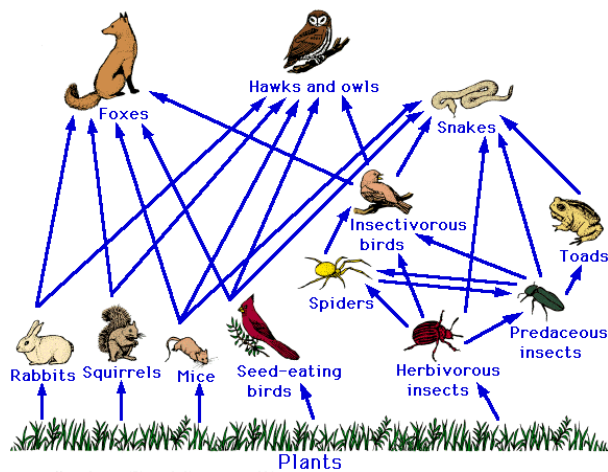


Figure 1.1: Foodweb ¹.

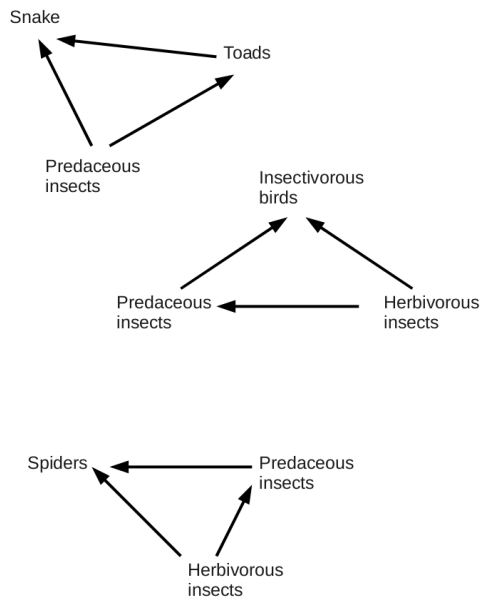


Figure 1.2: Trophic relationships derived from the foodweb. The species at the pointer feeds on the other. For example: the snakes feed on both the predaceous insects and the toads as the toads feed on the predaceous insects as well.

is given in figure 1.3. The gophers represent the host relating with their parasite which is lice in this case. From this figure, it can be seen that the relationship can be complex in that a single species from the host may be infected by multiple parasite species or multiple host

species infected by the same parasite species in addition to the one to one associations.

There are basically three trophic levels in an ecosystem. The first trophic level is composed of primary producers of energy, which are mainly plants. The second trophic level is made up of those that feed directly on the plants, called herbivores or primary consumers and the third level consists of predators. These three levels collectively form a trophic system. Examples of trophic relations include plants-herbivores-insects (pollinators), deceit-pollinated plants limited by the pollinators for seed set and the herbivores relying on the plants for fitness, and a parasitoids-plants-herbivores relationship (Micha et al., 2000), where varieties of plants and herbivores confront the parasitoids.

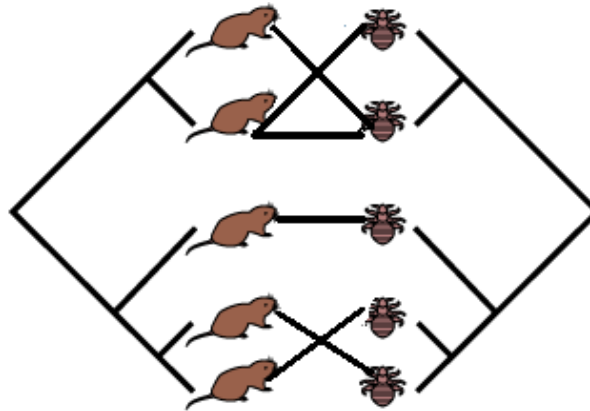


Figure 1.3: Host-Parasite Ditrophic associations ².

Huelsenbeck et al. (2000, 2001) employ maximum likelihood estimation in a Bayesian framework to analyse the ditrophic host-parasite relations whereas Hommola et al. (2009) use a permutation test. We seek to extend the permutation method to more than two phylogenetic associations.

Permutation methods are useful in that they do not require some parametric assumptions and are therefore reliable since they have no underlying distribution assumptions. One approach suggested by Lapointe and Legendre (1992) is to analyse the statistical significance of the matrix correlation coefficient by comparing independent phylogenetic trees and creating tables of critical values of the Pearson's cross product matrix correlation coefficient. However, our method does not require tables of critical values since the critical values are directly computed from the permuted statistics and compared with the observed statistics. Details are given in section 3.

²<http://evolution.berkeley.edu/evosite/evol01/VClhCospeciation.shtml>

1.3 Background on phylogenetic trees.

1.3.1 Properties of phylogenetic trees

Phylogenetic trees are diagrammatic representations of the evolutionary relationships (Ewens and Grant, 2001) that occur between taxonomic groups. Trees can be *rooted* (Figure 1.4) or *unrooted* (Figure 1.5). A biological phylogenetic tree must have a root, also called the ancestor of all the leaves. The leaves of a tree represent the species that are most current and are the terminal nodes of a rooted tree. Trees are said to be rooted if their branch lengths represent evolutionary time, with all species in the tree sharing a common ancestor. For cases of unrooted trees, there is no direction to indicate how the evolutionary time might have flowed. The trees are binary implying that an edge that branches will split to two daughter edges. The edges explain the evolutionary divergence that is associated to it and define the measure of distances between the species. Because of the fact that the current species share a common ancestor, they are likely to have similar traits in their DNA or protein sequences inherited from their common ancestor which has changed over evolutionary time through mutational processes such as substitution, deletion and insertion.

Phylogenetic trees convey two types of information:

1. The topology defines the branching order of the trees and the way the species are distributed among the leaves.
2. The branch lengths represent phylogenetic time, measured by the average amount of mutational change.

An unrooted tree of n species has $n - 2$ inner nodes, and $n + (n - 2) - 1 = 2n - 3$ branches (Durbin et al., 2009). In general, for n species, there are, in total, $(2n - 3)!!$ different rooted and $(2n - 5)!!$ different unrooted tree topologies, where $!!$ denotes double factorial. Leaves are labelled from 1 to n . The branch nodes are assigned the numbers $(n+1)$ to $(2n-1)$ reserving the $2n - 1$ for the root node. The number of possible trees from n leaves grow rapidly with n .

The study of cospeciation is limited by the fact that it relies on the accuracy of trees. The assumption is that the host-parasite trees are accurate measures of the evolutionary time. The distance between two or more species from their recent common ancestor may be defined in terms of years if this is known. However, the evolutionary time is not accurately known hence surrogate distances are used instead. However, since all the computations are based on these phylogenetic trees, it is assumed here that the distances are known.

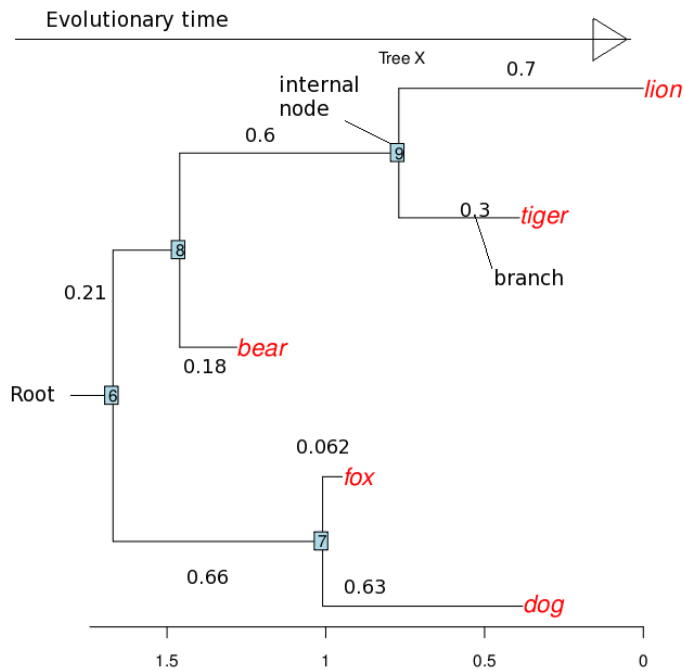


Figure 1.4: A rooted phylogenetic tree.

Definition

Suppose x_i , x_j and x_l are nodes in a tree X . Let $d(x_i, x_j)$ define the distance between nodes x_i and x_j . This distance is said to be *metric* (Mardia et al., 1979; Gentleman et al., 2005) if the following conditions for are satisfied:

- (i) $d(x_i, x_j) \geq 0, \forall i \neq j$
- (ii) $d(x_i, x_j) = 0$ iff $i = j$,
- (iii) $d(x_i, x_j) = d(x_j, x_i)$,
- (iv) $d(x_i, x_j) \leq d(x_i, x_l) + d(x_l, x_j) \forall x_i, x_j, x_l$ in X .

Patristic distances are additive phylogenetic distances obtained by summing up the branch lengths on a path between two nodes of a tree. The patristic distance describes the genetic changes in a tree (Fourment and Gibbs, 2006). Patristic distances can either be ultrametric or non-ultrametric. Non-ultrametric distances satisfy the above four conditions whereas ultrametric distances must meet the condition below in addition to the four stated above:

- (v) $d(x_i, x_l) \leq \max\{d(x_i, x_j), d(x_j, x_l)\}$

All the trees generated in this document are non-ultrametric.

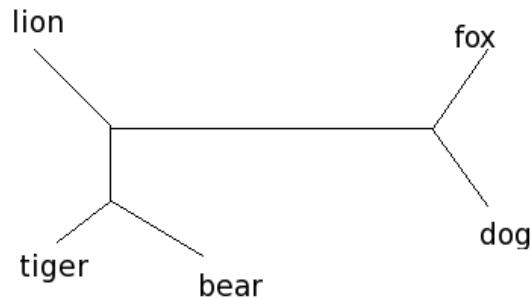


Figure 1.5: An unrooted tree.

Example 1.3.1. Calculating patristic distances

The phylogenetic tree in figure 1.4 is constructed using *R*, (R Development Core Team, 2010) and by loading library *ape* developed by Paradis (2006). The function *rtree* generates non-ultrametric phylogenetic trees by randomly splitting their edges. The standard format is called *newick* and is of class “phylo”. It represents the branch lengths. A practical example for the branch lengths for figure 1.4 is

```
"((dog:0.63,fox:0.062):0.66,(bear:0.18
(tiger:0.38,lion:0.77):0.69):0.21);"
```

The patristic distance between the dog and the fox is given by $0.63+0.062 = 0.692$. Similarly, the patristic distance between the dog and the bear is $0.63+0.66+0.18+0.21=1.68$. All pairwise patristic distances can be calculated to give the matrix of the patristic distances given below.

	<i>dog</i>	<i>fox</i>	<i>bear</i>	<i>tiger</i>	<i>lion</i>
<i>dog</i>	0.00	0.69	1.68	2.57	2.96
<i>fox</i>	0.69	0.00	1.11	2.00	2.39
<i>bear</i>	1.68	1.11	0.00	1.25	1.64
<i>tiger</i>	2.57	2.00	1.25	0.00	1.15
<i>lion</i>	2.96	2.39	1.64	1.15	0.00

Points representing the patristic distance matrix can be plotted on a Cartesian plane using the Multidimensional Scaling (MDS) methods to graphically visualize which species are close. MDS can be classical or non-classical, with classical allowing metric input and metric output and being easy to interpret whereas non-classical uses ranks of the distances corresponding to

the ranks of their dissimilarities. More details are not covered.

Figure 1.6 is a 2-dimensional classical MDS plot of the patristic distance matrix. It can be said that dogs are more similar to foxes than to lions and lions are more similar to tigers than to bears, may be due to evolutionary mutational processes.

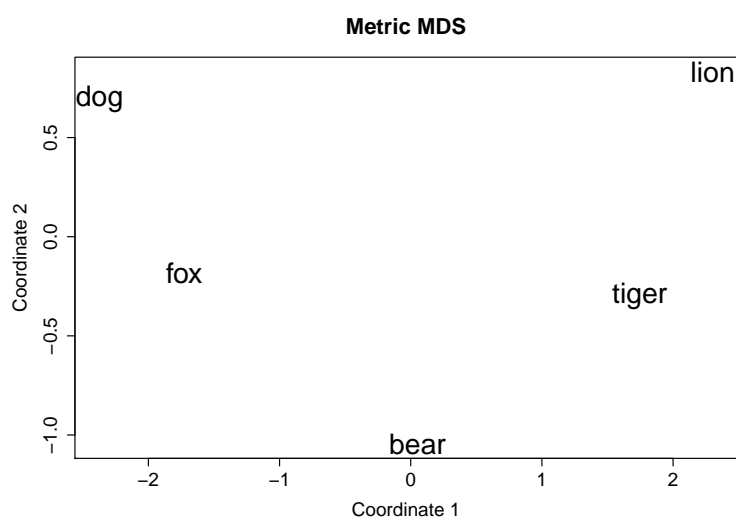


Figure 1.6: A 2D plot of the patristic distance matrix using classical MDS technique.

Chapter 2

Statistical Background

2.1 Correlation Coefficients

A population correlation coefficient, ρ , is a standard measure of linear relationships between sets of random variables, with ρ taking values $-1 \leq \rho \leq +1$. We might be interested in testing the $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$. This method is scale invariant and therefore does not change when the scale of the variables is changed. It is due to this property that correlation measures are preferred to covariance measures. However, the sample correlation coefficient, r , is a biased estimator for ρ and both ρ and r are not appropriate for measuring non-linear relationships.

Let X , Y and Z be three quantitative random variables representing the three trees. The relationship between, say X and Y is characterised by the population covariance σ_{xy} and the population correlation coefficient ρ . These parameters are used to indicate lack of independence and also to quantify the strength of a linear relationship (Jobson, 1991) between the two variables.

A random sample of X and Y of size n , (x_i, y_i) , $i = 1, 2, \dots, n$, can be used to obtain unbiased estimators μ_x, μ_y, σ_x^2 and σ_y^2 by computing \bar{x}, \bar{y}, s_x^2 and s_y^2 respectively. The sample covariance s_{xy} is the unbiased estimator of the covariance parameter σ_{xy} given by

$$s_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / (n - 1). \quad (2.1)$$

Pearson's pairwise correlation coefficient

The Pearson's pairwise product moment correlation coefficient, r , measures the degree of linearity between two variables and is derived from ρ .

$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}, \quad (2.2)$$
$$\text{thus, } r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

When the association is not linear but is monotonic, the Spearman's rank correlation coefficient, r_s , becomes useful, which is a Pearson's r calculated on the ranks of the variables. All correlation coefficients calculated in this study are based on the Pearson's method.

Partial correlation coefficient

In calculating sample partial correlation coefficient, a third variable has to be controlled for. This makes it possible to distinguish direct correlation between X and Y , and their correlation due to a shared common cause, Z (Crawley, 2007). The null hypotheses in this setting are:

$$H_0 : \rho_{xy.z} = 0,$$

$$H_0 : \rho_{yz.x} = 0,$$

$$H_0 : \rho_{xz.y} = 0.$$

where $\rho_{xy.z}$ is the population partial correlation of variables X and Y , after controlling for Z . Similarly, $\rho_{yz.x}$ is the population partial correlation of variables Y and Z , after controlling for X and $\rho_{xz.y}$ is the population partial correlation of variables X and Z , after controlling for Y .

Partial correlation coefficients are derived from the Pearson's correlation coefficient. When there are three random variables, it is necessary to compute all the three pairwise and partial correlation coefficients because they will be used in computing the three partial p values. The sample partial correlation coefficient between variables X and Y after controlling for Z is given by

$$r_{xy.z} = \frac{r_{xy} - r_{yz}r_{xz}}{\sqrt{(1 - r_{yz}^2)(1 - r_{xz}^2)}}. \quad (2.3)$$

The sample partial correlation coefficient between Y and Z after controlling for X is given

by

$$r_{yz.x} = \frac{r_{yz} - r_{xy}r_{xz}}{\sqrt{(1 - r_{xy}^2)(1 - r_{xz}^2)}}. \quad (2.4)$$

Similarly, the sample partial correlation coefficient between X and Z after controlling for Y is given by

$$r_{xz.y} = \frac{r_{xz} - r_{xy}r_{yz}}{\sqrt{(1 - r_{xy}^2)(1 - r_{yz}^2)}}. \quad (2.5)$$

Interpretation

The value $r_{yz.x} = 0.9$ implies that the variation in Z not explained by X has a strong correlation with the part of Y that is not related to X .

Example 2.1.1. Partial correlation coefficients

Consider the correlation matrix, \mathbf{R} , given by

$$\begin{array}{c} X \quad Y \quad Z \\ X \begin{pmatrix} 1 & r_{xy} & r_{xz} \\ r_{xy} & 1 & r_{yz} \\ r_{xz} & r_{yz} & 1 \end{pmatrix} \\ Y \\ Z \end{array}$$

The proportion of variation in variable Y that is explained by both variables X and Z , denoted by $r_{y.xz}^2$ is given by

$$\begin{aligned} r_{y.xz}^2 &= r_{xy}^2 + r_{yz.x}^2(1 - r_{xz}^2) \\ &= r_{xy}^2 + \frac{[r_{yz} - r_{xy}r_{xz}]^2}{1 - r_{xz}^2}. \end{aligned} \quad (2.6)$$

Here, $r_{y.xz}$ is called the coefficient of multiple correlation between the variable Y and variables X and Z . The additional variation in variable Y explained by variable X when variable Z is held constant is given by

$$r_{xy.z}^2(1 - r_{yz}^2) \quad (2.7)$$

and the total variation in variable Y explained by both variables X and Z is given by

$$r_{y.xz}^2 = r_{yz}^2 + r_{xy.z}^2(1 - r_{yz}^2) \quad (2.8)$$

2.2 Principal Component Analysis

This is one of the most widely used multivariate techniques first described by Pearson (1901). The details are also provided by Jolliffe (1986), Jackson (1991) and Timm (2002). The main objective of the principal component analysis (PCA) technique is to transform a set of p variables X_1, X_2, \dots, X_p to a set of p uncorrelated hypothetical constructs commonly known as the principal components Z_1, Z_2, \dots, Z_p arranged in order of their importance (Manly, 2005). These principal components can be used in many ways. They are used to explain the dependencies existing among variables and to detect existing relationships among observations (Timm, 2002).

The principal components need to be uncorrelated because they measure different dimensions of the data. They are ordered in a way that the variance $var(Z_1) \geq var(Z_2) \geq \dots \geq var(Z_p)$. This technique tends to reduce the large number of variables to a smaller number of transformed variables. It works best if the original variables are highly correlated.

Algebraic basis of PCA

Suppose that there are p variables X_1, X_2, \dots, X_p and n observations. Table 2.1 displays the structure of the data. The principal components are a linear combination of these variables.

Case	X_1	X_2	\dots	X_p
1	X_{11}	X_{12}	\dots	X_{1p}
2	X_{21}	X_{22}	\dots	X_{2p}
\vdots	\vdots	\vdots	\dots	\vdots
n	X_{n1}	X_{n2}	\dots	X_{np}

Table 2.1: A dataset structure with variables coded X_1, X_2, \dots, X_p and n observations ¹.

Let the coefficients of the linear combination of the p variables be denoted as $[b_{11}, b_{12}, \dots, b_{1p}] = \mathbf{b}_1$. Similarly, $[b_{21}, b_{22}, \dots, b_{2p}] = \mathbf{b}_2$ and $[b_{31}, b_{32}, \dots, b_{3p}] = \mathbf{b}_3$. Let $X_1, X_2, \dots, X_p = \mathbf{X}$.

The first principal component Z_1 is given by

$$Z_1 = b_{11}X_1 + b_{12}X_2 + \dots + b_{1p}X_p = \mathbf{b}_1\mathbf{X}, \quad (2.9)$$

and varies for each observation subject to the following constraint

$$b_{11}^2 + b_{12}^2 + \dots + b_{1p}^2 = \mathbf{b}_1'\mathbf{b}_1 = 1. \quad (2.10)$$

¹Table taken from Manly (2005, Chapter 6).

This implies that the variance of Z_1 is as large as possible given the constraint on the constants b_{1j} . Without the constraint, $var(Z_1)$ can be increased by increasing any one of the constants.

The second principal component

$$Z_2 = b_{21}X_1 + b_{22}X_2 + \dots + b_{2p}X_p = \mathbf{b}_2\mathbf{X} \quad (2.11)$$

is chosen such that $var(Z_2)$ is large subject to the condition that both Z_1 and Z_2 have zero correlation and that

$$\begin{aligned} b_{21}^2 + b_{22}^2 + \dots + b_{2p}^2 &= \mathbf{b}'_2\mathbf{b}_2 = 1. \\ \text{and } \mathbf{b}'_2\mathbf{b}_1 &= 0 \end{aligned} \quad (2.12)$$

The third principal component

$$Z_3 = b_{31}X_1 + b_{32}X_2 + \dots + b_{3p}X_p = \mathbf{b}_3\mathbf{X}. \quad (2.13)$$

is such that $var(Z_3)$ is large subject to the constraint that

$$\begin{aligned} b_{31}^2 + b_{32}^2 + \dots + b_{3p}^2 &= \mathbf{b}'_3\mathbf{b}_3 = 1 \\ \text{and } \mathbf{b}'_3\mathbf{b}_2 &= 0 \end{aligned} \quad (2.14)$$

and that Z_3 is uncorrelated with both Z_2 and Z_1 .

and similarly, the i th principal component

$$Z_i = b_{i1}X_1 + b_{i2}X_2 + \dots + b_{ip}X_p = \mathbf{b}_i\mathbf{X} \quad (2.15)$$

is such that the $var(Z_i)$ is large subject to the constraint that

$$\begin{aligned} b_{i1}^2 + b_{i2}^2 + \dots + b_{ip}^2 &= \mathbf{b}'_i\mathbf{b}_i = 1 \\ \text{and } \mathbf{b}'_i\mathbf{b}_j &= 0 \end{aligned} \quad (2.16)$$

and that Z_i is uncorrelated with all of the Z_j for $j < i$.

2.2.1 Eigenvalues and Eigenvectors

The principal component analysis technique involves computations of the eigenvalues and eigenvectors. The variances of the components are exactly the eigenvalues of the covariance matrix. If there are p original variables then there will be p eigenvalues some of which may be zeros but cannot be negative for a covariance matrix.

Consider a matrix D of order p . The scalar quantity λ is known as the characteristic root or the root or the eigenvalue of D if $(D - \lambda I_p)$ is singular. By definition, a matrix, say D is said to be *singular* if and only if each eigenvalue of D is 1 or 0. This implies that the determinant of $(D - \lambda I_p)$ must be equal to zero. That is

$$|D - \lambda I_p| = 0. \quad (2.17)$$

Equation 2.17 is sometimes called the characteristic or the eigen equation of the matrix D which is a p degree polynomial in λ with eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_p$. The eigenvalues are ordered such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ and that λ_i corresponds to the i th principal component. If a subset of the eigenvalues are equal such that $\lambda_1 = \lambda_2 = \dots = \lambda_q$, where $q < p$, then, the eigenvalues are said to have multiplicity q . From equation 2.17, the *rank*, $r(D - \lambda_q I_p) < p$ so that the columns of $(D - \lambda_q I_p)$ are linearly dependent. Thus, there exists non-zero vectors v_i such that

$$(D - \lambda_q I_p)v_i = 0 \text{ for } i = 1, 2, \dots, p. \quad (2.18)$$

These vectors v_i which satisfy equation 2.18 are called the eigenvectors or sometimes called the characteristic vectors. The covariance matrix Σ of rank p has the form:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \dots & \Sigma_{1p} \\ \Sigma_{21} & \Sigma_{22} & \dots & \Sigma_{2p} \\ \vdots & \vdots & \dots & \vdots \\ \Sigma_{p1} & \Sigma_{p2} & \dots & \Sigma_{pp} \end{bmatrix}$$

The total variance of the p components satisfy the equation below:

$$\begin{aligned} \lambda_1 + \lambda_2 + \dots + \lambda_p &= \Sigma_{11} + \Sigma_{22} + \dots + \Sigma_{pp} \\ \iff \sum_{i=1}^p \lambda_i &= \text{trace}(\Sigma). \end{aligned} \quad (2.19)$$

Thus the j th component explains a proportion P_j of the total variation on the dataset (Everitt and Dunn, 2001), where

$$P_j = \frac{\lambda_j}{\text{trace}(\Sigma)}, \quad (2.20)$$

and the first n principal components, where $n < p$ will account for a proportion P^* of the total variation where

$$P^* = \frac{\sum_{i=1}^n \lambda_i}{\sum_{i=1}^p \lambda_i} = \frac{\sum_{i=1}^n \lambda_i}{\text{trace}(\Sigma)}. \quad (2.21)$$

The principal components can be obtained from standardized variables Z_i^* (Johnson and Wichern, 1992) so as to have zero mean and a unit variance by

$$\begin{aligned} Z_1^* &= \frac{X_1 - \mu_1}{\sqrt{\sigma_{11}}} \\ Z_2^* &= \frac{X_2 - \mu_2}{\sqrt{\sigma_{22}}} \\ &\vdots \\ Z_p^* &= \frac{X_p - \mu_p}{\sqrt{\sigma_{pp}}} \end{aligned} \quad (2.22)$$

The covariance of variable X_i with component Z_j is given by

$$\text{cov}(X_i, Z_j) = \lambda_j b_{ji}$$

The correlation of variable X_i with component Z_j is

$$r_{x_i, z_j} = \frac{\text{cov}(X_i, Z_j)}{\sigma_{x_i} \sigma_{z_j}} = \frac{\lambda_j b_{ji}}{\sqrt{s_{ii} \lambda_j}} = \frac{\sqrt{\lambda_j} b_{ji}}{\sqrt{s_{ii}}}. \quad (2.23)$$

If the components are extracted from the correlation matrix rather than the covariance matrix, then

$$r_{x_i, z_j} = \sqrt{\lambda_j} b_{ji}. \quad (2.24)$$

The principal components are summarized in table 2.2.

Table 2.3 displays a summary of the loadings from using the covariance matrix.

Variables	Components			
	Z_1^*	Z_2^*	...	Z_p^*
X_1	b_{11}	b_{12}	...	b_{1p}
X_2	b_{21}	b_{22}	...	b_{2p}
\vdots	\vdots	\vdots	\vdots	\vdots
X_p	b_{p1}	b_{p2}	...	b_{pp}
Eigenvalues	λ_1	λ_2	...	λ_p
Eigenvectors	\mathbf{v}_1	\mathbf{v}_2	...	\mathbf{v}_p

Table 2.2: Principal component loadings.

Variables	Components			
	Z_1^*	Z_2^*	...	Z_p^*
X_1	$b_{11}\sqrt{\lambda_1}$	$b_{12}\sqrt{\lambda_2}$...	$b_{1p}\sqrt{\lambda_p}$
X_2	$b_{21}\sqrt{\lambda_1}$	$b_{22}\sqrt{\lambda_2}$...	$b_{2p}\sqrt{\lambda_p}$
\vdots	\vdots	\vdots	\vdots	\vdots
X_p	$b_{p1}\sqrt{\lambda_1}$	$b_{p2}\sqrt{\lambda_2}$...	$b_{pp}\sqrt{\lambda_p}$

Table 2.3: Principal components from covariance loadings.

2.2.2 PCA using a correlation matrix

Instead of using the covariance matrix, the matrix of correlations can be used to compute the principal components. Selecting a set of n principal components, where $n \ll p$, table 2.4 displays a structure of the components using the correlation matrix.

Variables	Components			
	Z_1^*	Z_2^*	...	Z_n^*
X_1	$b_{11}\sqrt{\lambda_1}/\sigma_1$	$b_{12}\sqrt{\lambda_2}/\sigma_1$...	$b_{1n}\sqrt{\lambda_n}/\sigma_1$
X_2	$b_{21}\sqrt{\lambda_1}/\sigma_2$	$b_{22}\sqrt{\lambda_2}/\sigma_2$...	$b_{2n}\sqrt{\lambda_n}/\sigma_2$
\vdots	\vdots	\vdots	\vdots	\vdots
X_p	$b_{p1}\sqrt{\lambda_1}/\sigma_p$	$b_{p2}\sqrt{\lambda_2}/\sigma_p$...	$b_{pn}\sqrt{\lambda_n}/\sigma_p$

Table 2.4: Principal components computed from a correlation matrix.

Example 2.2.1. Using the correlation structure

Suppose that there are two random variables X_1 and X_2 with a correlation matrix Φ given by

$$\Phi = \begin{bmatrix} 1.0 & r \\ r & 1.0 \end{bmatrix}$$

where r is the pairwise correlation coefficient of the variables X_1 and X_2 . Then, the principal

components (PC) of Φ can be calculated by first computing the eigenvalues and the eigenvectors of the matrix Φ . The eigenvalues λ_i are the roots of the equation

$$|\Phi - \lambda I_p| = 0.$$

This leads to a quadratic equation of the form

$$(1 - \lambda)^2 - r^2 = 0,$$

yielding the eigenvalues

$$\lambda_1 = 1 + r \text{ and } \lambda_2 = 1 - r,$$

where $\lambda_1 + \lambda_2 = 2 = \text{trace}(\Phi)$.

To find the eigenvector that corresponds to the eigenvalue λ_1 , the equation

$$\Phi b_1 = \lambda_1 b_1$$

is rescaled leading to the equations

$$\begin{aligned} b_{11} + r b_{12} &= (1 + r) b_{11} \\ r b_{11} + b_{12} &= (1 + r) b_{12}, \end{aligned}$$

both equations being identical and leading to

$$b_{11} = b_{12}.$$

Introducing the normalization constraint such that $b_1^T b_1 = 1$ gives

$$b_{11} = b_{12} = \frac{1}{\sqrt{2}}$$

and the second eigenvector is found in a similar way, giving

$$b_{21} = \frac{1}{\sqrt{2}} \text{ and } b_{22} = \frac{-1}{\sqrt{2}}.$$

Hence, the principal components Z_1 and Z_2 are given by

$$Z_1 = \frac{1}{\sqrt{2}}(X_1 + X_2) \text{ and } Z_2 = \frac{1}{\sqrt{2}}(X_1 - X_2).$$

For $r < 0$, the order of the roots and that of the principal components is reversed whereas if $r = 0$, the roots are both equal to 1 implying that any two orthogonal solutions can represent the two principal components.

Chapter 3

Methods

3.1 Test statistics

The methods introduced in this project develop three types of test statistics under the multivariate statistical techniques and employ permutation method to test the hypotheses stated below. The statistics include the partial correlation coefficients, PCA eigenvalues computed from the correlation matrix and eigenvalues computed from the covariance matrix. The permutation approach is preferred because it makes no assumptions underlying the distribution of the data.

The hypotheses are

H_0 : The host-parasite phylogenies are not related indicating that they have evolved independently;

H_1 : The host-parasite phylogenies are closely related indicating cospeciation.

The dataset consists of phylogenies generated under the null hypothesis H_0 or under the alternative hypothesis H_1 and an association matrix representing the interactions among the phylogenetic trees. A large number of phylogenies can be generated under the permutation method and different statistics calculated.

Notation:

Let the trees be denoted as X, Y, Z . Let x denote a tip from tree X , y from tree Y and z from tree Z respectively.

Let (x, y, z) be a triple such that edges xy, xz and yz all exist in the interaction graphs of XY, XZ and YZ respectively such that (x, y, z) picks out a triangle of interactions.

Let T denote the set of all observed triples (x, y, z) and let n denote the number of elements of T where $n > 1$. Let (x_i, y_i, z_i) denote the i th triple of T .

Suppose i and j are distinct elements of T . Let $d(x_i, x_j)$ denote the patristic distance between tips x_i and x_j of tree X . Similarly denoting $d(y_i, y_j)$ and $d(z_i, z_j)$ for trees Y and Z respectively.

Generate a matrix D with $n(n - 1)/2$ rows and three columns, where each row represents a distinct pair of triples i and j for $j < i$ and where each column represents a tree. Thus the $(i_1)(i_2)/2 + j$ th row of D will contain $d(x_i, x_j), d(y_i, y_j), d(z_i, z_j)$.

Example 3.1.1. Calculating matrix D

Calculation of matrix D can be done from data that has been generated either under H_0 , which represents independent evolution or under H_1 to represent cospeciation. The following trees have been generated under H_0 with their branch lengths given in *newick* format as

```
## Tree X
" ( (3:0.629, 4:0.0618) :0.661, (1:0.177,
(2:0.384, 5:0.77) :0.687) :0.206) ; "
## Tree Y
" ( (5:0.126, 2:0.267) :0.652, ( (3:0.382, 4:0.87) :
0.0134, 1:0.34) :0.386) ; "
## Tree Z
" ( (3:0.724, 1:0.411) :0.108, ( (5:0.783, 2:0.553) :
0.647, 4:0.53) :0.821) ; "
```

A triangular association matrix is assigned randomly and for this example, is

$$\begin{array}{c} X \quad Y \quad Z \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} \begin{pmatrix} 4 & 3 & 1 \\ 1 & 5 & 2 \\ 3 & 3 & 3 \\ 4 & 2 & 4 \\ 4 & 1 & 3 \end{pmatrix} \end{array}.$$

This association matrix has $n = 5$, $T_1 = (4, 3, 1)$, $T_2 = (1, 5, 2)$, $T_3 = (3, 3, 3)$, $T_4 = (4, 2, 4)$ and $T_5 = (4, 1, 3)$. Using this association matrix, the trees can be plotted as shown in figure 3.2. Patristic distances for each tree are calculated by adding the branch lengths between nodes as explained in example 1.3.1. The three patristic distances for X , Y and Z are:

$$X = \begin{matrix} & \begin{matrix} 3 & 4 & 1 & 2 & 5 \end{matrix} \\ \begin{matrix} 3 \\ 4 \\ 1 \\ 2 \\ 5 \end{matrix} & \begin{pmatrix} 0.00 & 0.69 & 1.67 & 2.57 & 2.95 \\ 0.69 & 0.00 & 1.11 & 2.00 & 2.39 \\ 1.67 & 1.11 & 0.00 & 1.25 & 1.63 \\ 2.57 & 2.00 & 1.25 & 0.00 & 1.15 \\ 2.95 & 2.39 & 1.63 & 1.15 & 0.00 \end{pmatrix} \end{matrix}$$

$$Z = \begin{matrix} & \begin{matrix} 3 & 1 & 5 & 2 & 4 \end{matrix} \\ \begin{matrix} 3 \\ 1 \\ 5 \\ 2 \\ 4 \end{matrix} & \begin{pmatrix} 0.00 & 1.13 & 3.08 & 2.85 & 2.18 \\ 1.13 & 0.00 & 2.77 & 2.54 & 1.87 \\ 3.08 & 2.77 & 0.00 & 1.34 & 1.96 \\ 2.85 & 2.54 & 1.34 & 0.00 & 1.73 \\ 2.18 & 1.87 & 1.96 & 1.73 & 0.00 \end{pmatrix} \end{matrix}$$

$$Y = \begin{matrix} & \begin{matrix} 5 & 2 & 3 & 4 & 1 \end{matrix} \\ \begin{matrix} 5 \\ 2 \\ 3 \\ 4 \\ 1 \end{matrix} & \begin{pmatrix} 0.00 & 0.39 & 1.56 & 2.05 & 1.50 \\ 0.39 & 0.00 & 1.70 & 2.19 & 1.65 \\ 1.56 & 1.70 & 0.00 & 1.25 & 0.74 \\ 2.05 & 2.19 & 1.25 & 0.00 & 1.22 \\ 1.50 & 1.65 & 0.74 & 1.22 & 0.00 \end{pmatrix} \end{matrix}$$

Matrix D has $n(n-1)/2$ rows and the number of columns is equivalent to the number of trees. Since $n = 5$, this implies that there are $5(4)/2 = 10$ rows. Each row will contain the triples: $d(x_i, x_j), d(y_i, y_j), d(z_i, z_j)$. A general structure is given in table 3.1 and the calculated matrix D is given in figure 3.1.

	$d(x_i, x_j)$	$d(y_i, y_j)$	$d(z_i, z_j)$
T1,T2	$d(x_1, x_4)$	$d(y_3, y_5)$	$d(z_1, z_2)$
T1,T3	$d(x_3, x_4)$	$d(y_3, y_3)$	$d(z_1, z_3)$
T1,T4	$d(x_4, x_4)$	$d(y_3, y_2)$	$d(z_1, z_4)$
T1,T5	$d(x_4, x_4)$	$d(y_3, y_1)$	$d(z_1, z_3)$
T2,T3	$d(x_1, x_3)$	$d(y_5, y_3)$	$d(z_2, z_3)$
T2,T4	$d(x_1, x_4)$	$d(y_5, y_2)$	$d(z_2, z_4)$
T2,T5	$d(x_1, x_4)$	$d(y_5, y_1)$	$d(z_2, z_3)$
T3,T4	$d(x_3, x_4)$	$d(y_3, y_2)$	$d(z_3, z_4)$
T3,T5	$d(x_3, x_4)$	$d(y_3, y_1)$	$d(z_3, z_3)$
T4,T5	$d(x_4, x_4)$	$d(y_2, y_1)$	$d(z_4, z_3)$

Table 3.1: A general format on how to fill matrix D .

Matrix D can be visualized using 3D plots, boxplots and line graphs among other methods as shown in figure 3.3. The first graph plots the row observations from 1, 2, ..., 10 for each of the columns. Since the lines do not follow a specific pattern and are crossing randomly, then the three trees have no close relationship implying that they might have evolved independently. Test statistics given in section 3.1.1 will be computed to confirm these observations. The same comments are made from the box plots as the box plots have different variances, skewness and varying means of the three columns of the matrix D and from the 3D scatter plot which shows

	X	Y	Z
1	2.57	0.74	1.13
2	1.25	0.00	3.08
3	0.00	1.70	2.85
4	0.00	1.56	3.08
5	1.67	0.74	2.77
6	2.57	1.65	2.54
7	2.57	1.50	2.77
8	1.25	1.70	1.34
9	1.25	1.56	0.00
10	0.00	0.39	1.34

Figure 3.1: The calculated matrix D from this example, where all the trees are random with a random association matrix.

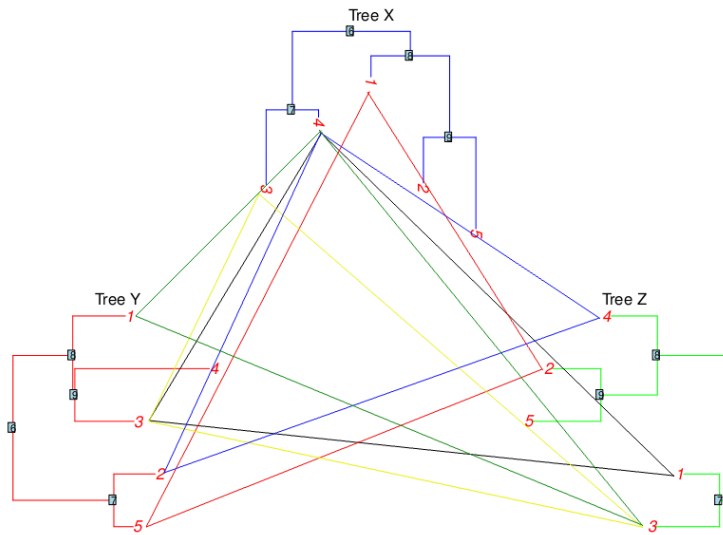


Figure 3.2: An example of random associations among three random phylogenies.

how random the 10 rows of observations are.

Matrix D can be formed from trees that have been generated under H_1 to represent cospeciation. The trees X , Y and Z are set to be very close, by generating them to have the same topologies such as $X = Y = Z$, and to have a triangular association matrix that has associations at corresponding positions of the trees such that $T_1 = (1, 1, 1)$, $T_2 = (2, 2, 2)$, $T_3 = (3, 3, 3)$, $T_4 = (4, 4, 4)$ and $T_5 = (5, 5, 5)$. The new matrix D is given in figure 3.4 and has all the three columns exactly with the same values.

The new plots would be as given in figure 3.5. It is evident from these plots that there is a very strong relationship among the phylogenies, a scenario that would indicate perfect cospeciation. The first plot has all the three lines X , Y , Z appearing as one. The box plots have same

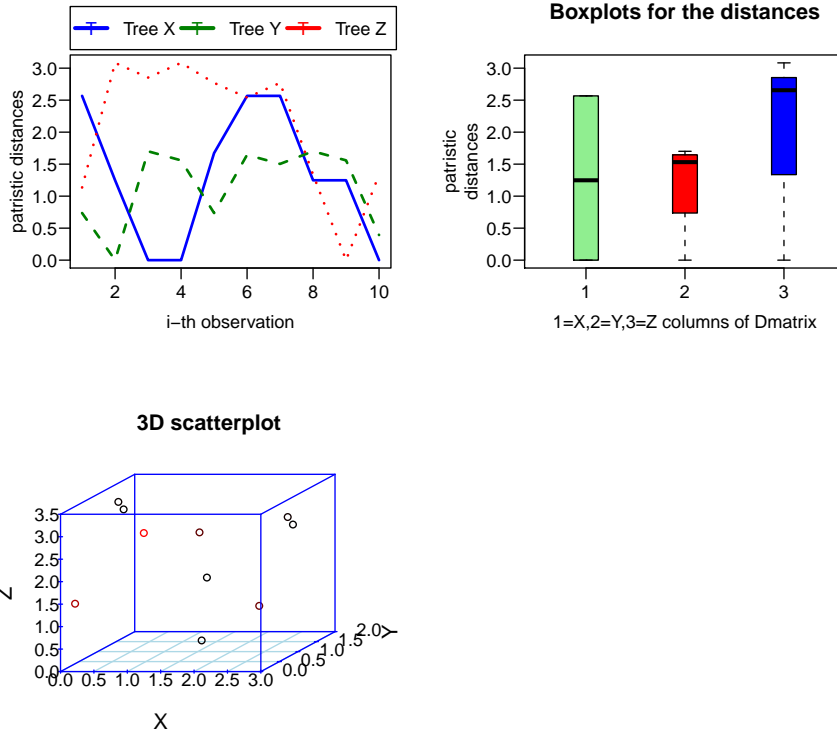


Figure 3.3: Plots to visualize matrix D from trees generated under H_0 to represent independent evolution of the Host-Parasite phylogenies.

variances, same skewness, same mean of the patristic distances and the scatter plot shows that all the 10 observations lie on the line $x = y = z$ which indicates high dependencies.

	X	Y	Z
1	0.69	0.69	0.69
2	1.67	1.67	1.67
3	2.57	2.57	2.57
4	2.95	2.95	2.95
5	1.11	1.11	1.11
6	2.00	2.00	2.00
7	2.39	2.39	2.39
8	1.25	1.25	1.25
9	1.63	1.63	1.63
10	1.15	1.15	1.15

Figure 3.4: Matrix D calculated from trees generated under H_1 , where $X = Y = Z$ and the triangular associations are at corresponding positions of the trees.

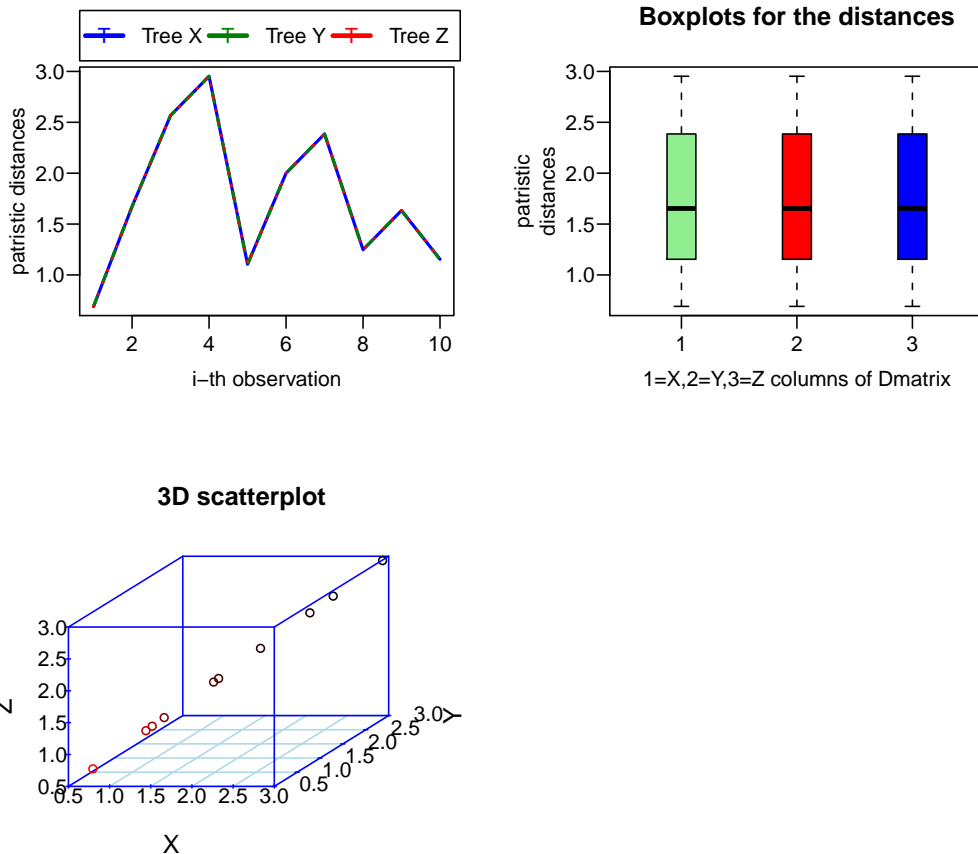


Figure 3.5: Plots to visualize matrix D for trees $X = Y = Z$ and with the associations being at corresponding positions of the trees. This scenario represents cospeciation.

3.1.1 Test statistic for partial correlation

Let the sample partial correlation coefficients from the observed data be denoted by $r_{xy.z}^{obs}$, $r_{yz.x}^{obs}$ and $r_{xz.y}^{obs}$. For each permutation i of the labels of the trees, let the sample partial correlation coefficients obtained after a large number of permutations, say N be denoted by $r_{xy.z_i}^p$, $r_{yz.x_i}^p$ and $r_{xz.y_i}^p$.

To test the significance of the observed partial correlations, the partial p values are computed by summing the number of permuted partial correlation coefficients that are greater than or equal to the observed partial correlation coefficient and divided by the number of permutations (equation 3.1). The geometric p value of the three partial p values is then computed as the root of the product of the three p values given by equation 3.4.

The partial p value after controlling for variable Z is calculated as:

$$P_z = \frac{1}{N} \sum_{i=1}^N I(r_{xy.z_i}^p \geq r_{xy.z}^{obs}), \quad (3.1)$$

where N is the total number of permutations, and $r_{xy.z_i}^p$ is the partial correlation coefficient between variables X and Y while controlling for variable Z for the i th permutation.

The partial p value after controlling for variable X is calculated as:

$$P_x = \frac{1}{N} \sum_{i=1}^N I(r_{yz.x_i}^p \geq r_{yz.x}^{obs}). \quad (3.2)$$

and likewise, the partial p value after controlling for variable Y is given by

$$P_y = \frac{1}{N} \sum_{i=1}^N I(r_{xz.y_i}^p \geq r_{xz.y}^{obs}). \quad (3.3)$$

Here,

$$I(r_{\star}^p \geq r_{\star}^{obs}) = \begin{cases} 1 & \text{if } r_{\star}^p \geq r_{\star}^{obs} \\ 0 & \text{otherwise} \end{cases}$$

where \star stands for either $xy.z_i$, or $yz.x_i$ or $xz.y_i$.

The geometric p value

The geometric p value is considered to be the overall test of significance of the observed partial

correlation coefficients and is given by

$$P_{gm} = \prod_{i=1}^p \{P_z P_x P_y\}^{1/p}. \quad (3.4)$$

For the three variables (or phylogenies), $p = 3$.

Decision

If $P_{gm} \leq \alpha$, where α is the significance level, H_0 is rejected and a conclusion made that there is a close relationship among the three host-parasite systems which could indicate cospeciation. Otherwise, there is no evidence to reject H_0 .

3.1.2 Test statistic for PCA

Computations are done on matrix D , the matrix that takes pairs of the triangular relationships from the patristic distances of the three phylogenetic trees X , Y , and Z . The observed eigenvalues are computed from both the covariance matrix and the correlation matrix before any permutations to the data has been done and are denoted as:

$$\lambda_r^{obs} = \text{eigen}\{\text{correlation}(\text{matrix}(D))\} \quad (3.5)$$

$$\lambda_c^{obs} = \text{eigen}\{\text{covariance}(\text{matrix}(D))\}. \quad (3.6)$$

Each of these produces three eigenvalues. The first, second and third eigenvalues correspond to the first, second and third principal components respectively. For each permutation i , for say a large number of permutations, N , the permuted eigenvalues are denoted as

$$\lambda_{r_i}^p = \text{eigen}\{\text{correlation}(\text{matrix}(D))_i\} \quad (3.7)$$

$$\lambda_{c_i}^p = \text{eigen}\{\text{covariance}(\text{matrix}(D))_i\}. \quad (3.8)$$

Only the first eigenvalues for both the covariance and correlation structures are used for the statistics. This is because they are usually the largest in size, and sometimes the eigenvalues for PC2 and PC3 can be negligibly small, they explain the largest proportion of variation in the data and have the highest variance.

The p values are then calculated as

$$P_{\lambda_r} = \frac{1}{N} \sum_{i=1}^N I(\lambda_{r_i}^p \geq \lambda_r^{obs}) \quad (3.9)$$

$$P_{\lambda_c} = \frac{1}{N} \sum_{i=1}^N I(\lambda_{c_i}^p \geq \lambda_c^{obs}), \quad (3.10)$$

where P_{λ_r} is the p value due to the correlation structure and P_{λ_c} is the p value due to the covariance structure and where

$$I(\lambda_{\star}^p \geq \lambda_{\star}^{obs}) = \begin{cases} 1 & \text{if } \lambda_{\star}^p \geq \lambda_{\star}^{obs} \\ 0 & \text{otherwise} \end{cases}$$

where \star stands for either r_i , or c_i .

Decision is made based on the p values and the set level of significance, α . If $p \leq \alpha$, H_0 is rejected and a close relationship of the three host-parasite system is inferred. Otherwise, there is no sufficient evidence to reject H_0 .

Permutation Algorithm

Step 1: Set the significance level α .

Step 2: For each of the three phylogenetic trees X , Y and Z , calculate their patristic distances.

Step 3: Compute matrix D using both the patristic distances and the association matrix.

Step 4: Using matrix D , calculate the three observed pairwise correlation coefficients: r_{xy}^{obs} , r_{yz}^{obs} and r_{xz}^{obs} and partial correlation coefficients $r_{xy.z}^{obs}$, $r_{yz.x}^{obs}$ and $r_{xz.y}^{obs}$.

Step 5: Compute the observed eigenvalues from both the covariance and the correlation matrices and consider the eigenvalues from the first principal components.

Step 6: Permute the labels of the trees.

step 7: Repeat steps 2 to 6 for a large number of times, say N . For each permutation i , compute the permuted partial correlations coefficients $r_{xy.z_i}^p$, $r_{yz.x_i}^p$ and $r_{xz.y_i}^p$ and the permuted eigenvalues $\lambda_{r_i}^p$ and $\lambda_{c_i}^p$.

Step 8: Compute the three partial p values P_z , P_x and P_y . Compute the geometric mean of these p values, P_{gm} . Compute the p values of the eigenvalues P_{λ_r} and P_{λ_c} .

Step 9: Practically, only one method is applied. This could be either the partial correlation or the PCA using the correlation structure or using the covariance structure. In either case reject H_0 if $P \leq \alpha$, else there is no evidence to reject H_0 .

3.1.3 Permutations under the Null hypothesis

The three phylogenetic trees namely X , Y and Z are randomly generated under H_0 to represent species that have evolved independently, given their phylogenetic trees and their association matrix. The triangular association matrix should be randomly generated.

Example 3.1.2. Implementation in R

All the trees are generated randomly with ten tips. For example, tree X is given by

```
(( (1:0.073, 2:0.49):0.76, 3:0.64):0.87, (4:0.58, ((5:0.072, 6:0.55):0.29, 7:0.68):0.8, 8:0.45):0.16):0.085);
```

Tree Y given by

```
((1:1, (2:0.11, ((3:0.71, 4:0.7):0.8, (5:0.26, 6:0.51):0.56):0.63):0.53):0.32, (7:0.68, 8:0.037):0.73);
```

and tree Z given by

```
(( (1:0.93, 2:0.039):0.32, (3:0.58, 4:0.71):0.78):0.9, (( (5:0.14, 6:0.23):0.85, 7:0.2):0.42, 8:0.089):0.36);
```

Figure 3.6 is obtained by plotting these trees. An association matrix has to be supplied such as the one given in figure 3.7.

A function called *nperm* has been developed. The parameters required by this function are: the three phylogenetic trees, a triangular association matrix and the number of permutations desired. This function computes the patristic distances of each of the trees and uses the association matrix to build matrix D . It then computes the pairwise correlation coefficients, observed and permuted partial correlation coefficients and observed and permuted eigenvalues from both covariance and correlation matrices. To test the significance of the observed statistics, the partial geometric p value and the p values due to the eigenvalues from both the correlation and covariance structure are computed. The details of the R code are attached in appendix A.1.1. The syntax is

```
nperm(x, y, z, relation, permut)
```

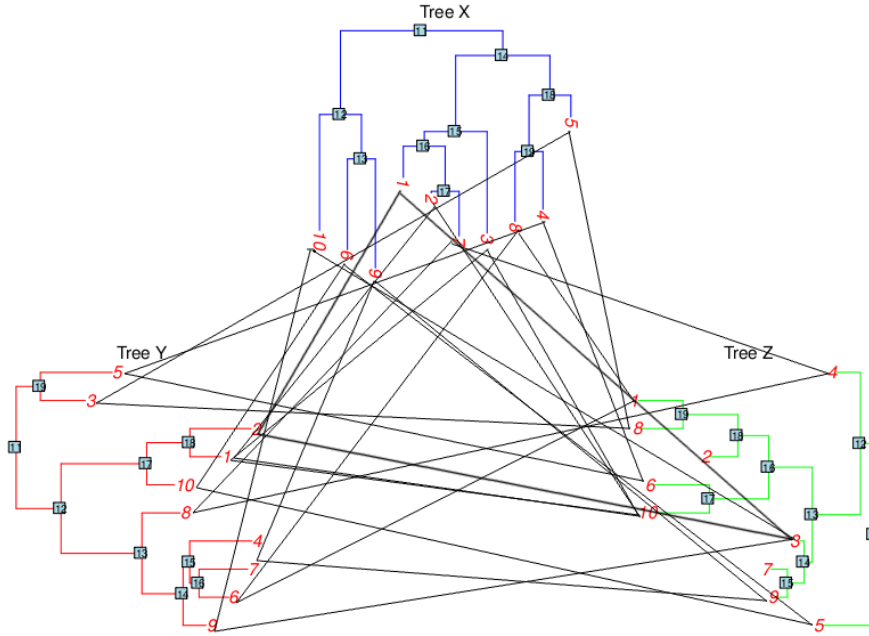


Figure 3.6: Three random phylogenetic trees with 10 tips

$$\begin{array}{c}
 X \quad Y \quad Z \\
 \begin{array}{c}
 1 \\
 2 \\
 3 \\
 4 \\
 5 \\
 6 \\
 7 \\
 8 \\
 9 \\
 10
 \end{array}
 \begin{pmatrix}
 1 & 2 & 3 \\
 2 & 1 & 10 \\
 3 & 1 & 10 \\
 4 & 5 & 6 \\
 5 & 3 & 8 \\
 6 & 10 & 5 \\
 7 & 8 & 4 \\
 8 & 6 & 1 \\
 9 & 7 & 9 \\
 10 & 9 & 3
 \end{pmatrix}
 \end{array}$$

Figure 3.7: An association matrix under H_0 .

where x , y and z are phylogenetic trees in *newick* format.

Density plots are drawn with lines of both the observed statistics and the critical value line for the first 95th percentile of the permuted coefficients and eigenvalues. If the observed statistics lie below the critical value then it means H_0 is not rejected. These results are confirmed by the three p values ($P_{gm} = 0.347$, $P_{\lambda_r} = 0.699$ and $P_{\lambda_c} = 0.860$) calculated in section 4.1.

3.1.4 Type I error

To investigate type I error, data is generated under the H_0 and a random association matrix used. 1000 p values are calculated for each of the statistic and their cumulative distributions checked for uniformity of the p values. The labels of the trees are permuted a large number of times for each p value calculated.

The probability of incorrectly rejecting H_0 is the probability of committing the type I error and is equal to the significance level, α . Hence type I error corresponds to α . It is also desired that α be as small as possible. For the statistic to be unbiased, the p values generated under H_0 should be uniformly distributed.

A function called *simdata* has been developed that enables the user to obtain several p values from the three statistics as desired. The *R*-code is attached in appendix A.1.2 and its syntax is

```
simdata(nsim)
```

where *nsim* is the number of p values to be calculated. This function calls the *nperm* function which permutes the data $N = 10,000$ times or to any value of N as desired. The results can be displayed as a matrix with a number of columns being equal to the number of statistics calculated and rows equal to the number of p values stated in the function. Empirical cumulative distribution functions are then plotted from these values.

Example 3.1.3. Type I error plots

Running the *simdata* function at a set seed of 100, the following type I error plots and tables display the level of uniformity of the distributions of p values when the labels of the different trees are permuted 1000 times and varying the number of p values calculated.

Tables 3.2, 3.3, and 3.4 give the 10, 20 and 50 p values obtained for all the three statistics respectively. Using graphs makes it easier to observe the pattern of the p values and for this purpose figure 3.8 for the 10 p values, figure 3.9 for the 20 p values, figure 3.11 for 100 p values for trees with 10 tips and figure 3.12 for 100 p values for trees with 15 tips are given. From these figures, we can say that the larger the number of p values calculated the more the distribution of the p values tend to have a uniform distribution.

The type I error plots for the 10, 20, and 50 p values do not look random due to the small sample sizes. From the plots of 100 p values, the PCA statistic from the correlation structure has a uniform distribution whereas the others deviate from the straight diagonal line. Calculations for 1000 p values (given in section 4 in figure 4.4) show distributions that are perfectly

uniform for all of the PCA statistics as they make straight diagonal lines. However, the partial correlation coefficient statistic does not seem to have a uniform distribution even for large values of up to 1000 p values (given in section 4 in figure 4.4) .

	<i>partial</i>	<i>eigen_cor</i>	<i>eigen_cov</i>
1	0.40	0.91	0.80
2	0.22	0.43	0.16
3	0.62	0.37	0.96
4	0.41	0.59	0.50
5	0.41	0.88	0.86
6	0.69	0.78	0.95
7	0.30	0.58	0.93
8	0.27	0.83	0.41
9	0.13	0.07	0.06
10	0.08	0.00	0.29

Table 3.2: 10 P values obtained from trees with 10 tips, having permuted the labels 10000 times.

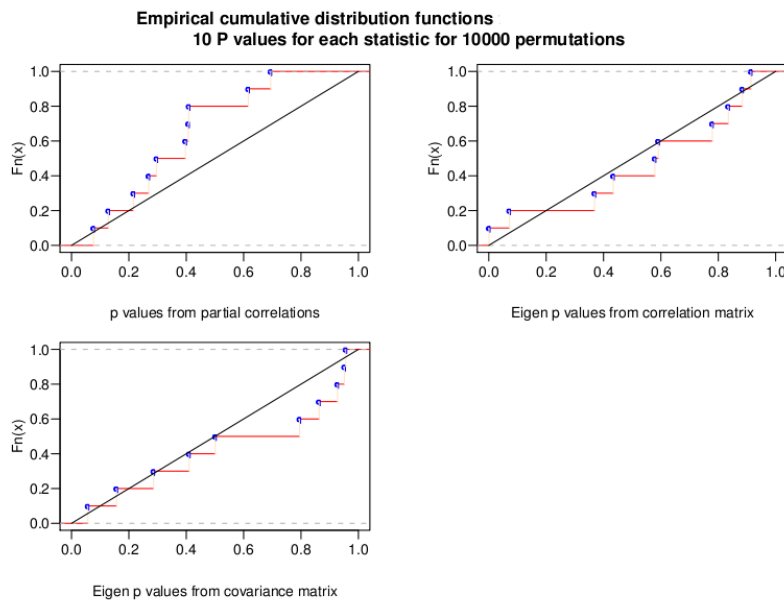


Figure 3.8: Type I error plots for 10 p values for trees with 10 tips. The labels are permuted 10000 times for each p value.

	<i>partial</i>	<i>eigen_cor</i>	<i>eigen_cov</i>
1	0.49	0.72	0.69
2	0.10	0.01	0.51
3	0.45	0.84	0.07
4	0.37	0.68	0.54
5	0.68	0.36	0.66
6	0.70	0.62	0.68
7	0.83	0.31	0.80
8	0.23	0.24	0.04
9	0.26	0.58	0.55
10	0.55	0.90	0.75
11	0.85	0.39	0.57
12	0.35	0.37	0.62
13	0.47	0.15	0.84
14	0.49	0.64	0.68
15	0.33	0.17	0.24
16	0.30	0.97	0.97
17	0.58	0.18	0.70
18	0.68	0.88	0.87
19	0.32	0.68	0.10
20	0.40	0.99	0.01

Table 3.3: 20 P values for trees with 10 tips

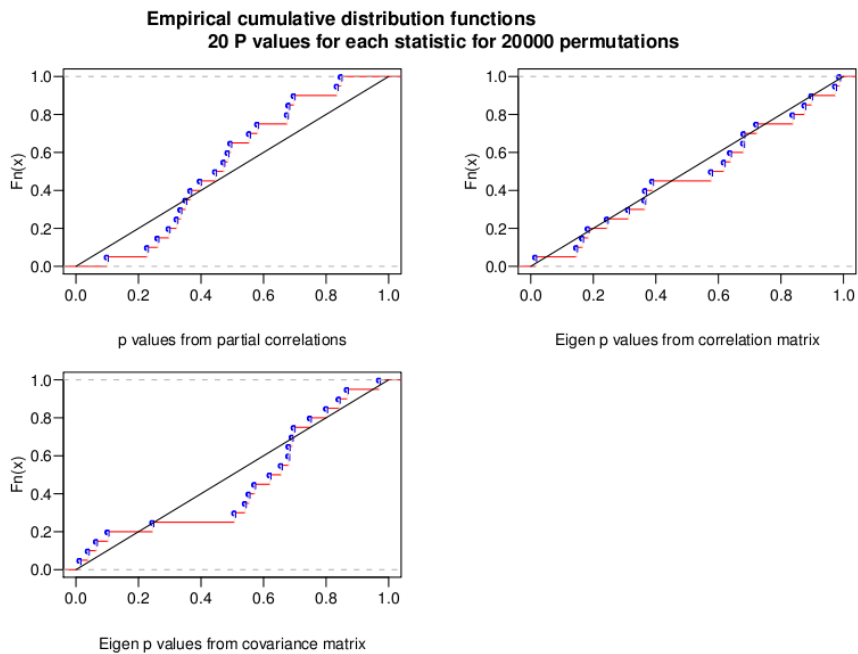


Figure 3.9: Type I error plots for 20 p values calculated from trees with 10 tips.

	<i>partial</i>	<i>eigen_cor</i>	<i>eigen_cov</i>		<i>partial</i>	<i>eigen_cor</i>	<i>eigen_cov</i>
1	0.24	0.42	0.89	26	0.42	0.59	0.08
2	0.07	0.01	0.07	27	0.50	0.41	0.83
3	0.34	0.72	0.91	28	0.37	0.79	0.95
4	0.47	0.78	0.69	29	0.74	0.71	0.56
5	0.55	0.60	0.55	30	0.33	0.88	0.52
6	0.51	0.33	0.17	31	0.57	0.78	0.41
7	0.28	0.24	0.81	32	0.68	0.81	0.49
8	0.51	0.10	0.59	33	0.56	0.64	0.75
9	0.24	0.45	0.32	34	0.25	0.24	0.23
10	0.71	0.57	0.65	35	0.35	0.04	0.57
11	0.58	0.95	0.89	36	0.66	0.76	0.34
12	0.60	0.76	0.91	37	0.33	0.39	0.65
13	0.13	0.04	0.03	38	0.14	0.04	0.83
14	0.69	0.59	0.69	39	0.86	0.78	0.10
15	0.64	0.09	0.30	40	0.23	0.58	0.45
16	0.83	0.31	0.12	41	0.69	0.29	0.61
17	0.69	0.27	0.52	42	0.15	0.22	0.08
18	0.63	0.50	0.71	43	0.85	0.49	0.69
19	0.61	0.94	0.85	44	0.43	0.92	0.70
20	0.64	0.71	0.36	45	0.17	0.35	0.53
21	0.63	0.67	0.24	46	0.19	0.30	0.73
22	0.76	0.62	0.46	47	0.38	0.67	0.91
23	0.38	0.38	0.98	48	0.16	0.14	0.95
24	0.52	0.40	0.79	49	0.15	0.24	0.25
25	0.30	0.71	0.85	50	0.49	0.99	0.80

Table 3.4: 50 P values for trees with 10 tips

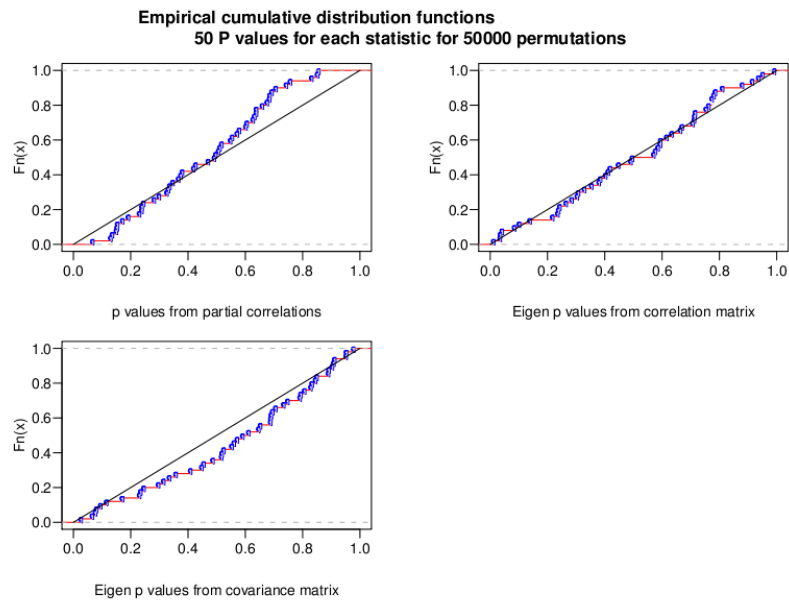


Figure 3.10: Type I error plots for 50 p values on trees with 10 tips

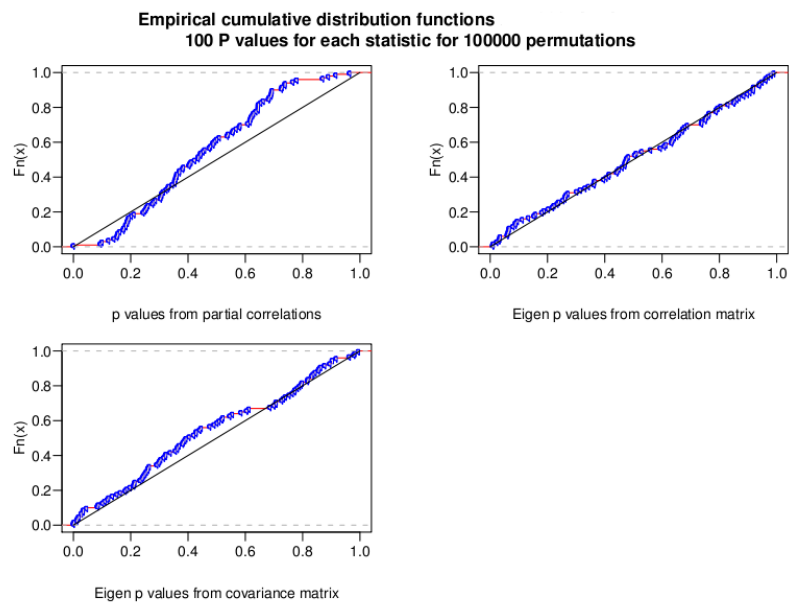


Figure 3.11: Type I error plots for 100 P values for trees with 10 tips, permuting the tree labels 1000 times for each p value.

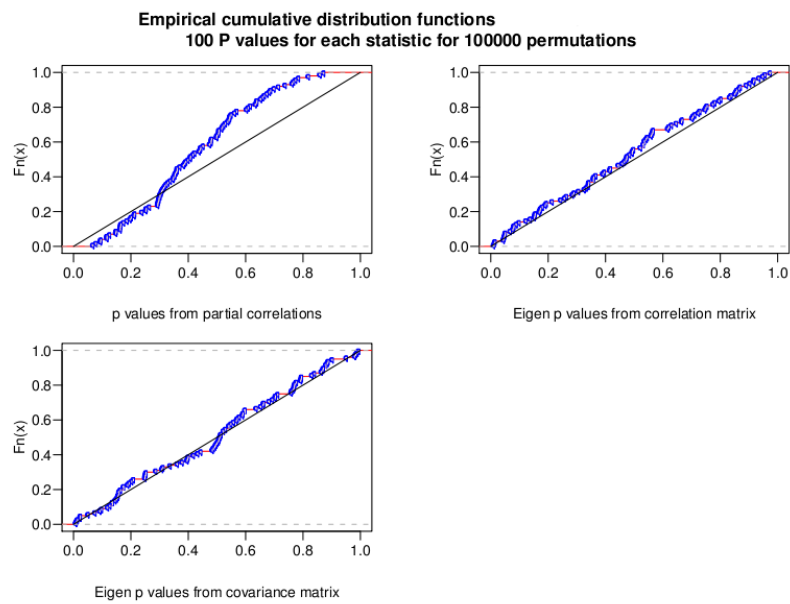


Figure 3.12: Type I error plots for 100 P values for trees with 15 tips, permuting the tree labels 1000 times for each p value.

3.2 Power Simulations

The probability of incorrectly accepting H_0 when it is false is the probability of committing a type II error and is equal to β .

$$\text{Power} = 1 - \text{Pr}(\text{type II error}) = 1 - \beta \quad (3.11)$$

is the ability of the statistic to correctly reject H_0 when it is false and we want this to be large. Table 3.5 summarizes the type I and type II errors.

	H_0 true	H_0 false
Accept H_0	$1-\alpha$	β
Reject H_0	α	$1-\beta$

Table 3.5: A table summarizing type I and type II errors.

To compute the power, the phylogenies have to be generated under the alternative hypothesis H_1 , making them to be exactly the same and having interaction triangles at their corresponding positions. For instance, let the first phylogeny, X , be generated randomly with ten tips. The second and third phylogenies can be made similar to X by either adding say δ and Φ to all branch lengths to get trees Y and Z respectively or by setting trees Y and Z to be exactly the same trees as tree X .

For $n = 10$ triangles, there will be $n(n - 1)/2 = 10(9)/2 = 45$ observations in the matrix D and three columns X , Y and Z representing the trees. The association matrix forms triangles T , where $T_1 : (x_1, y_1, z_1)$, $T_2 : (x_2, y_2, z_2)$, \dots , $T_{10} : (x_{10}, y_{10}, z_{10})$. Thus, the number of rows of the association matrix corresponds to the number of tips of the trees as shown in figure 3.13. The first phylogeny is generated randomly whereas the second and third phylogenies are made to be exact copies of the first.

	X	Y	Z
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9
10	10	10	10

Figure 3.13: An association matrix under H_1 .

To compute the power for the three statistics, two approaches have been implemented, adopted from Hommla et al. (2009) and Legendre et al. (2002).

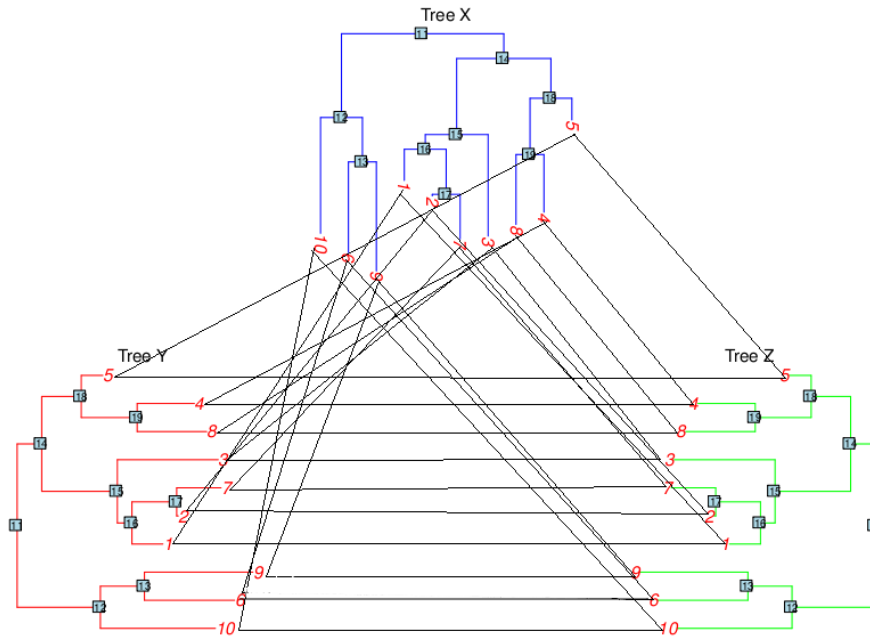


Figure 3.14: Three phylogenetic trees under H_1 .

3.2.1 First approach: Adding random triangles

In this approach, random triangles are added as a percentage of the existing number of triangular associations. The phylogenies are generated each with ten tips, giving ten corresponding association triangles. The random triangles are added at the rate of 10%, 20%, ..., 100% implying that there are 1, 2, ..., 10 more triangles of associations added respectively. Thus the more random triangles we have the more the scenario gets closer to the simulations under H_0 . The same procedure is repeated for phylogenies with 20 tips in order to measure the performance of the method in a large scale phylogenetic system.

For power simulations, 100 or 1000 p values can be calculated, while permuting the labels of the trees a larger number of times, say 10000. A function called *addtriangles* has been written for this purpose and is attached in appendix A.1.3. A sum of the p values greater than α is noted. Plots for rejection rate against the percentage of added triangles are given in section 4.

Example 3.2.1. Adding random triangles

The following triangular association matrices are obtained after adding 10%, 30%, and 50% of random triangles into the original association matrix. The added random triangles are shown in red.

	X	Y	Z
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9
10	10	10	10
11	2	1	8

	X	Y	Z
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9
10	10	10	10
11	4	4	10
12	4	6	8
13	1	7	1

	X	Y	Z
1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9
10	10	10	10
11	6	10	9
12	9	1	4
13	9	10	4
14	1	8	5
15	3	3	6

Figure 3.15 shows the effect of adding random triangles to the original association matrix. The new association matrix has 15 triangles where 10 are at corresponding positions and 5 are random triangles (in red). The more random triangles are added or replaced the more the relationship gets closer to H_0 and the higher the probability of rejecting H_1 .

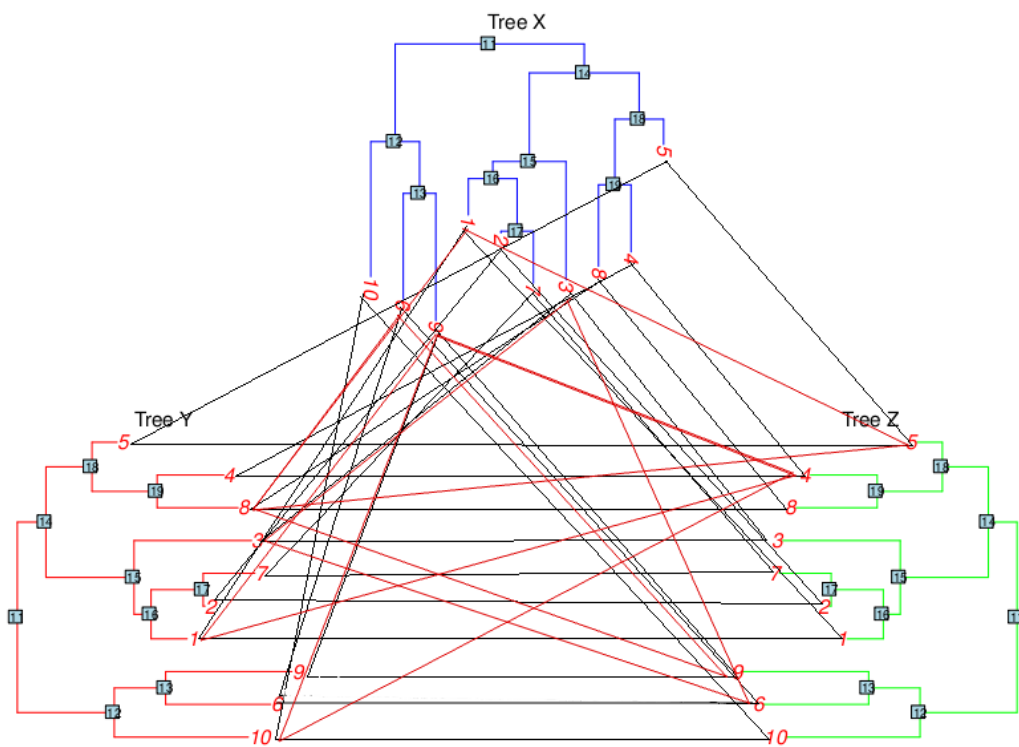


Figure 3.15: The effect of adding random triangles to the original association matrix.

3.2.2 Second approach: Replacing triangles

In this approach, a certain percentage of the triangles are randomly substituted with random triangles which are not allowed to be at corresponding positions. This approach has been done for phylogenies with both 10 and 20 tips by substituting 10%, 20%, . . . , 50% of the association triangles randomly.

Example 3.2.2. *Substituting triangles*

The following triangular association matrices are obtained when 10%, 20%, . . . , 50% of triangles are substituted with random triangles (in red).

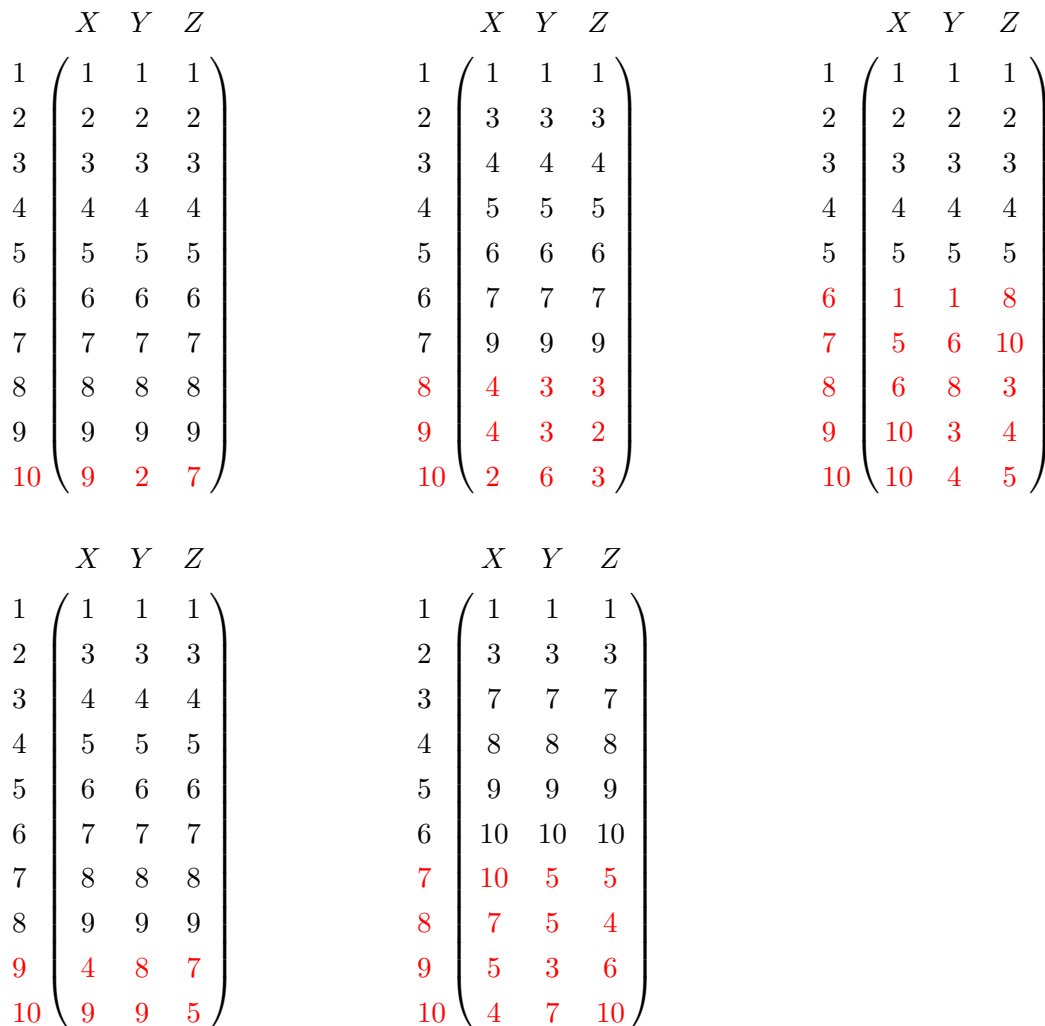


Figure 3.16 shows the effect of substituting 50% of the original triangles that had associations at their corresponding positions with random triangles. The relationship portrayed by this figure is more of random phylogenetic trees than of trees generated under H_1 . Hence the

power to reject H_0 reduces as more triangular associations in corresponding positions are substituted with random triangles.

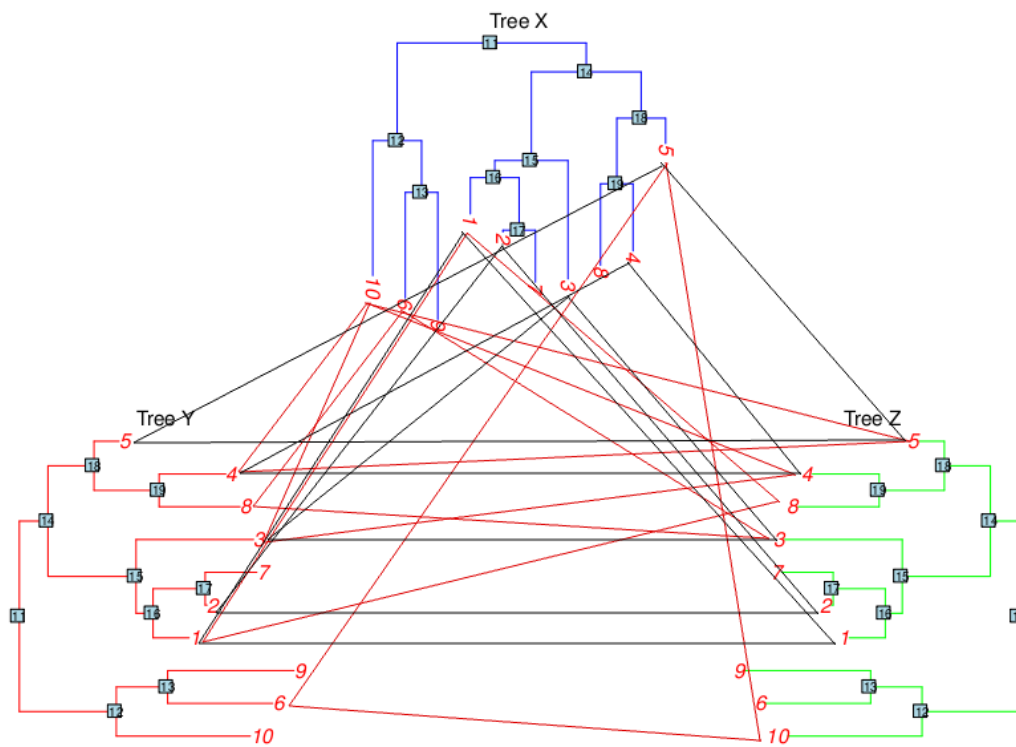


Figure 3.16: The effect of removing the original association triangles whose relationship was at corresponding positions of the trees and replacing them with random triangles.

Chapter 4

Results

4.1 Results under the null hypothesis

From the examples explained in section 3.1.3 and setting a *seed* of 100, the observed partial correlation coefficients obtained were: $r_{xy.z}^{obs} = -0.075$, $r_{yz.x}^{obs} = 0.0203$, $r_{xz.y}^{obs} = 0.1505$, and the observed eigenvalues for the three principal components are given in table 4.1.

In using the PCA, the principal component loadings show that the eigenvalues computed using the covariance matrix are almost twice as large as those computed using the correlation matrix. More important are the proportions of variance that are explained by these components. It is as expected that since these trees are randomly generated and that their triangular associations are not at corresponding positions, the proportions of variance should not be much different. Thus the observed results of 0.388, 0.336 and 0.277 for PC1, PC2 and PC3 are in agreement with the underlying assumption of randomness under H_0 . The standard deviations of the principal components displayed in table 4.1 are also not much different. The first plot in figure 4.2 is a bar plot of these variances whereas the second and third graphs are scree plots for the eigenvalues due to covariance and correlation matrices respectively. The observed Pearson's pairwise correlation coefficients are displayed in figure 4.1.

More plots from the observed output are given in figure 4.2. In the second row of figure 4.2, the first graph is a plot of the rotations of the principal components and the points inside this plot are the observations from matrix D . The plotted points are random since the trees are random as well. The line graph plots the observations in matrix D and also shows how random the patristic distances are, in that the lines criss-cross a lot making it look like a plot of "noisy signals". The box plots show that the means of the X , Y and Z columns are different

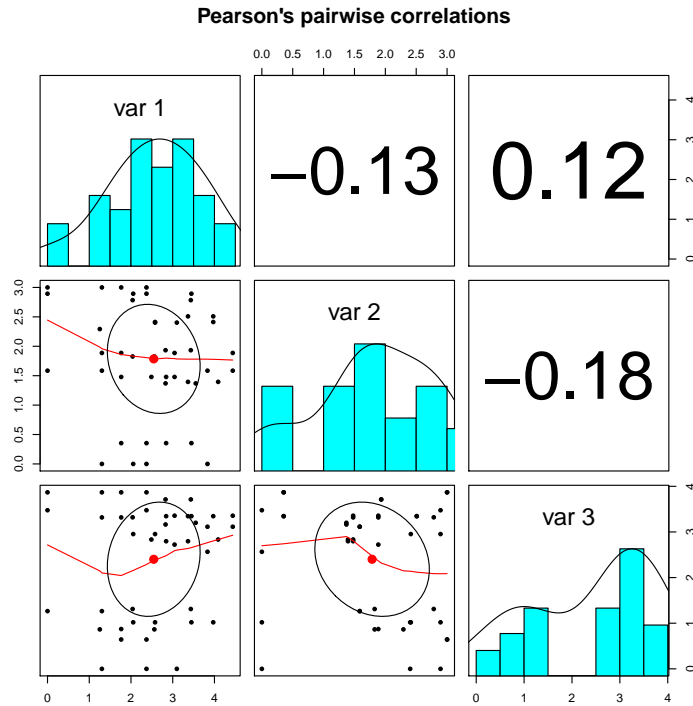


Figure 4.1: Observed pairwise correlation coefficients under H_0 . var 1, var 2 and var 3 represents X , Y and Z respectively.

and the last 3D scatter plot displays the points as being scattered through out the box.

Results from permuted labels

The three trees' labels are permuted 10,000 times and, for each permutation, the pairwise correlation coefficients, partial correlation coefficients and the eigenvalues for the first principal components (PC1) are used. The results show that the geometric mean of the partial p values, $P_{gm} = 0.347$, $P_{\lambda_r} = 0.699$ and $P_{\lambda_c} = 0.860$. All these p values are above 0.05 and above 0.01. Setting a level of significance of, say, 0.05 would result in not rejecting H_0 . The same conclusion would be reached if the significance level is set at 0.01. The 10,000 coefficients are

	PC1	PC2	PC3
Importance of components			
Standard deviations	1.078	1.004	0.911
Proportion of Variance	0.388	0.336	0.277
Cumulative Proportion	0.388	0.723	1.000
Eigenvalues: λ_r^{obs}	1.163	1.007	0.830
Eigenvalue: λ_c^{obs}	1.434	1.198	0.999

Table 4.1: Principal components under H_0 .

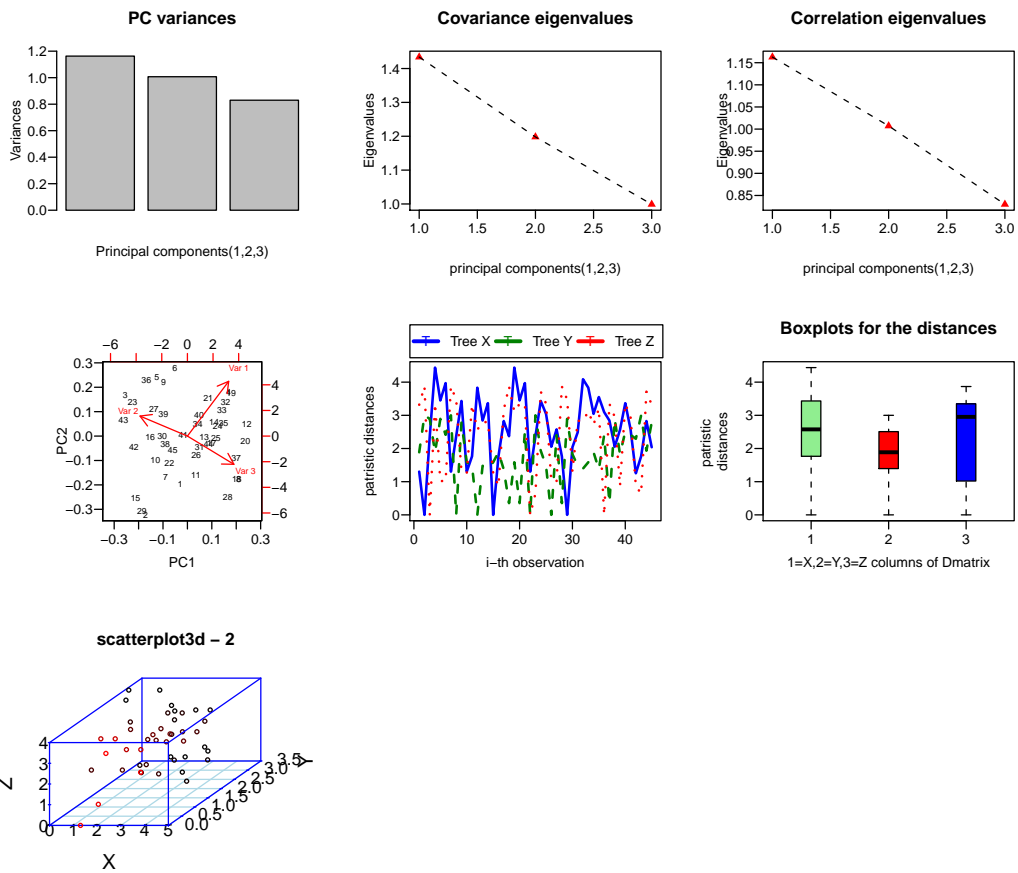


Figure 4.2: Observed Results under H_0 .

then ordered and plotted as density histograms. Lines of the observed coefficients and critical regions are drawn to assist in interaction of the output as given in figure 4.3. Again, the same conclusion can be reached by observing these plots as the observed statistics are outside the critical regions.

Density plots under the null hypothesis

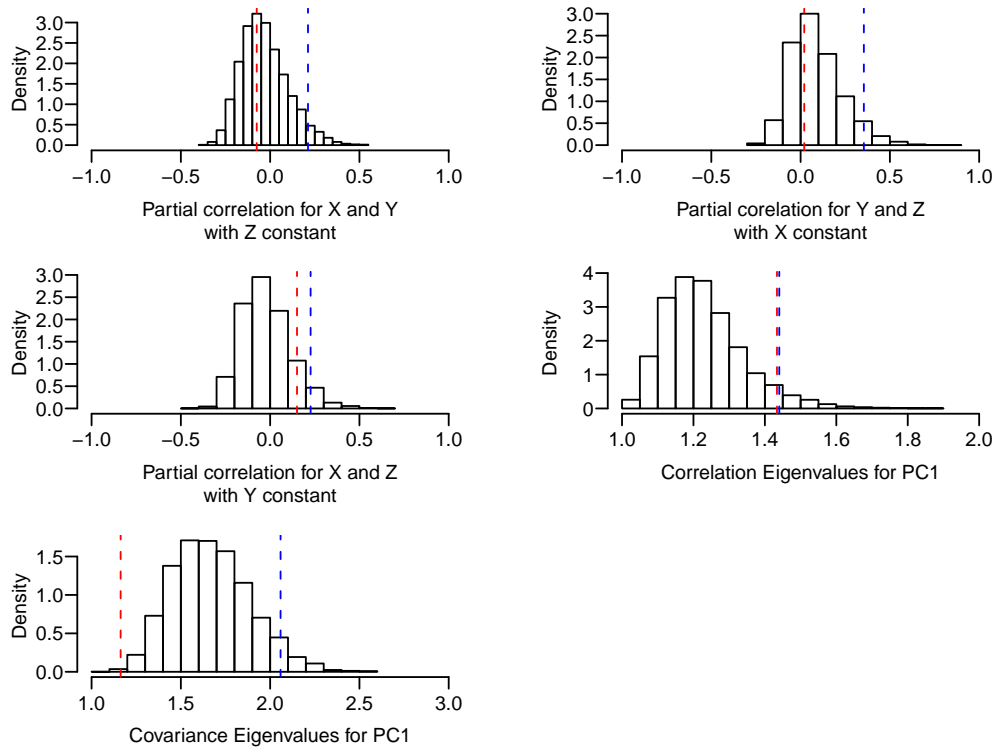


Figure 4.3: Density plots for permuted coefficients and critical values under H_0 . The red line represent the observed value and blue represents the critical line at the 95th percentile.

Results for type I error

In order to calculate the type I error, 1000 p values are calculated for trees with 10 tips and trees with 15 tips. For each of the p value, 1000 permutations on the labels of the trees is performed. Thus there will be 1000 p values each for geometric partial p values, eigenvalues due to correlation matrix and eigenvalues due to covariance matrix. Empirical cumulative distribution function of these results are displayed in figure 4.4. It is evident from these plots that statistics from using the principal component analysis produces uniformly distributed p values, this being not the case when the partial correlations coefficient statistic is used. These plots suggest the use of PCA technique as being a more reliable statistic than its counterpart. The R code used has been attached in appendix A.1.2.

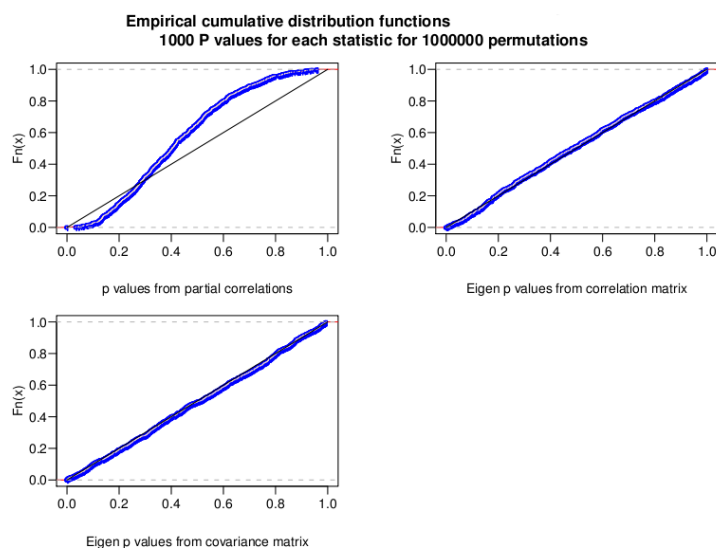


Figure 4.4: Type I error plots for trees with 10 tips performed for 1000 p values permuting the tree labels 10000 times.

Results from using a relatively larger number of tips such as 15 showed that it does not change type I error distribution of the p values. The plots are given in figure 4.5.

4.2 Results under the alternative hypothesis

4.2.1 Results under a perfect H_1 condition

The phylogenetic trees are produced in such way that they are all exactly the same, each with 10 tips and with their triangular relations at their corresponding positions. Appendix A.1.3 shows the R code used.

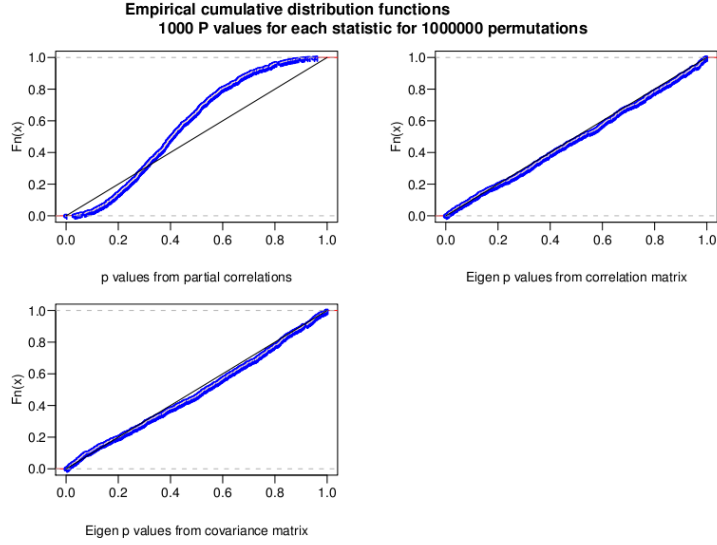


Figure 4.5: Type I error plots for the distribution of 1000 p values for trees with 15 tips, permuting the tree labels 1000 times.

The observed pairwise correlation coefficients are $r_{xy}^{obs} = r_{yz}^{obs} = r_{xz}^{obs} = 1$. This makes the partial correlation coefficients $r_{xy.z}^p$, $r_{yz.x}^p$ and $r_{xz.y}^p$ not computable. Thus this statistic cannot be used unless adjustments to the trees is made. The pairwise plot is given in figure 4.6.

Figure 4.8 displays plots that are strikingly different from the ones obtained under H_0 . The first plot on the first row shows that PC1 explains 100% of the total variation in the data. The standard deviation for PC1 is 1.73 while it is negligibly small for PC2 and PC3. The second and third plots that PC1 has the highest eigenvalue for both covariance and correlation matrices but PC2 and PC3 have almost zero values. In the second row, the rotation of the principal components makes the patristic distances appear clustered together in a line, not being random. All the box plots are similar and the 3D plot in row three has a straight line of points along the $x = y = z$ line indicating dependency of the observed trees.

	PC1	PC2	PC3
Importance of components			
Standard deviations	1.73	0.00	0.00
Proportion of Variance	1.00	0.00	0.00
Cumulative Proportion	1.00	1.00	1.00
Eigenvalues: λ_r^{obs}	3.00	0.00	0.00
Eigenvalue: λ_c^{obs}	3.09	0.00	0.00

Table 4.2: Principal components under perfect conditions of H_1 .

On the other hand, the principal component analysis technique seem to pick up this effect

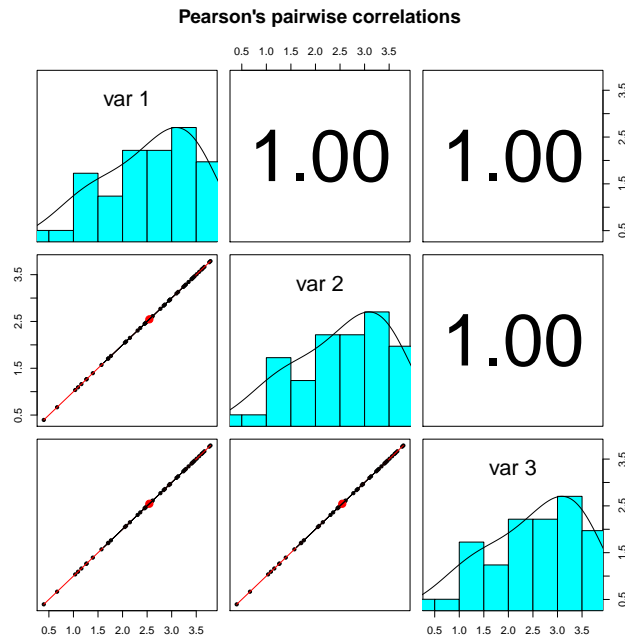


Figure 4.6: Pairwise correlation coefficients produced under a perfect H_1 . Var 1, var 2 and var 3 stand for X , Y and Z from the matrix D .

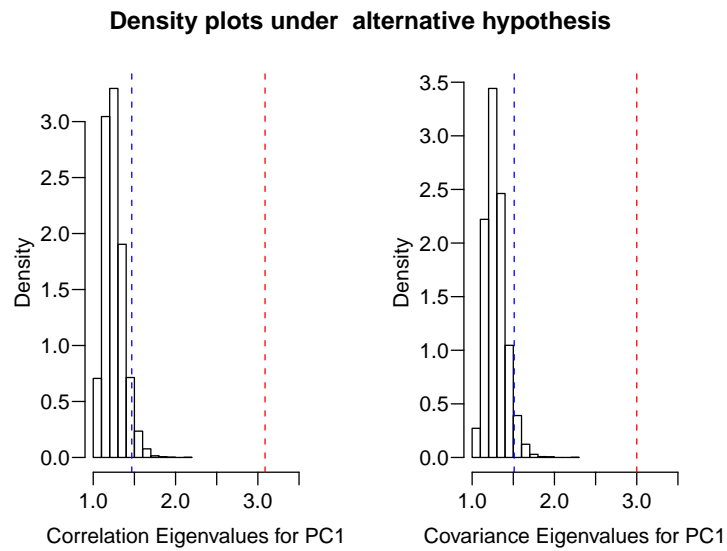


Figure 4.7: Density plots for permuted coefficients and critical values under H_1 . The **red** line represent the observed value and **blue** represents the critical line at the 95th percentile.

quite well and is not affected like the partial correlation method. The observed eigenvalues are in table 4.2 and the density plots given in figure 4.8.

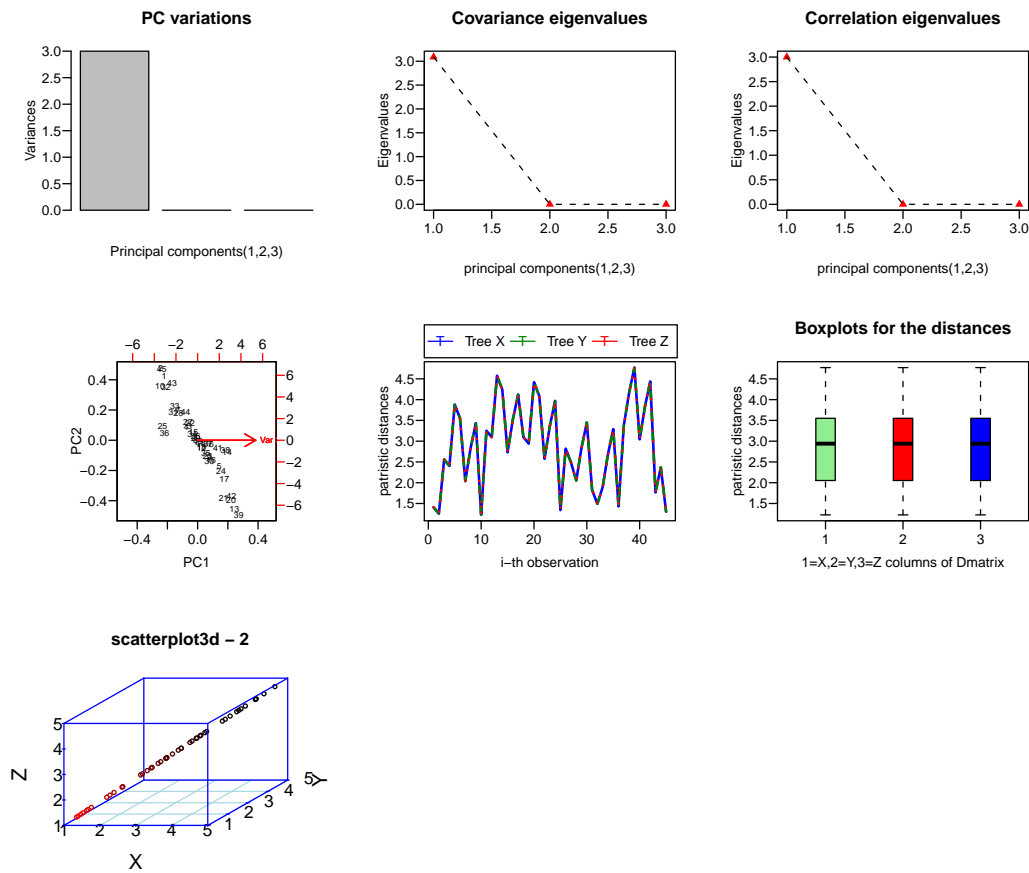


Figure 4.8: Principal components analysis loadings under a perfect H_1 . All the plots reveal an overwhelming evidence of correlations among the phylogenies.

When 10,000 permutations on the labels of the trees is performed, the p values for both P_{λ_r} and $P_{\lambda_c} = 0$, giving an overwhelming evidence to reject H_0 . Figure 4.7 shows the density plots from the PCA method. The labels of the trees have been permuted 10000 times. Since the observed eigenvalues lie in the critical region, this means that there is a strong evidence to reject H_0 and conclude that the trees have a close linear relationship which could indicate cospeciation. This is consistent with the p values of zeros obtained from the calculations above.

4.2.2 Results for adding random triangles

Power simulation

For the trees with 10 tips, 10%, 20%, . . . , 50% and 100% of random triangles were added and 100 p values calculated, each time permuting the labels 10000 times. The same has been done for trees with 20 tips, but this round, 10%, 20%, . . . , 100% of random triangles added to the original association matrix and 1000 p values calculated, each time permuting the labels 1000 times. The R code for adding links has been given in appendix A.1.3.

Figure 4.9 displays the power curves for trees with 10 tips. It is evident from these plots that the power to reject H_0 diminishes as more random triangles are added. This is because, adding random triangles weakens the condition for simulation under H_1 . The more the random triangles are added, the more H_0 condition is approached and therefore the lower the power to reject H_0 . The power curve obtained from the partial correlation coefficient statistic at 0.01 significance level is worse compared to the other statistics as the power drops sharply when adding the random triangles.

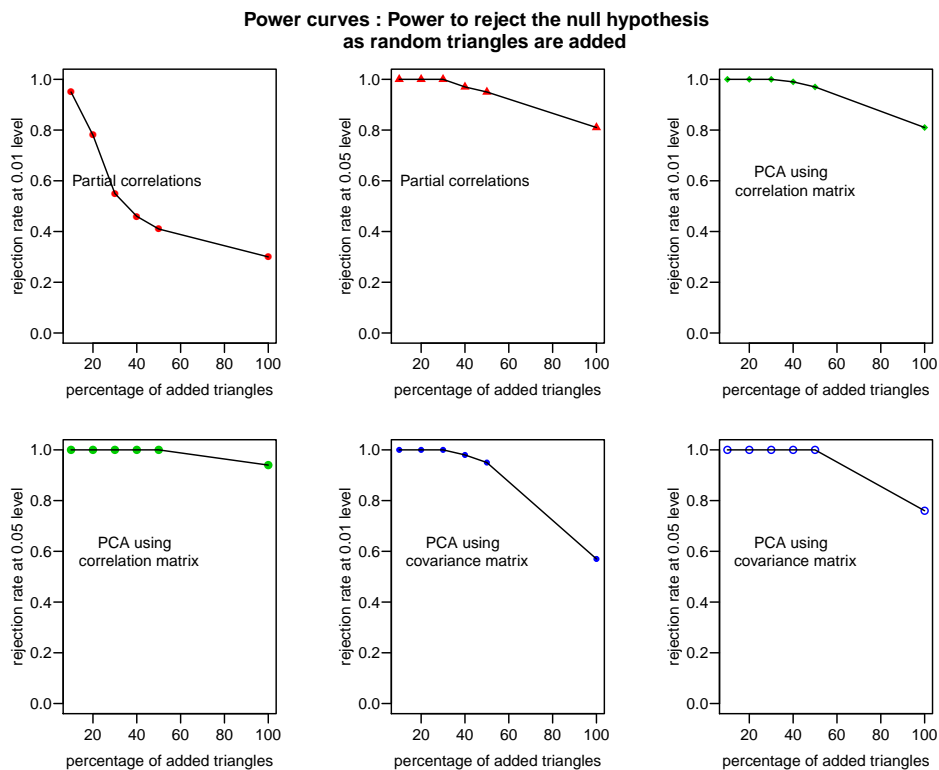


Figure 4.9: Power curves for trees with 10 tips when random triangles are added to the existing association matrix generated under H_1 .

Computations to test the power of the three statistics to reject the null hypothesis for large phylogenetic trees when random triangles are added has been performed on trees with 20 tips. The results are plotted in figure 4.10. This plot shows that the power to reject H_0 is very high for large phylogenetic trees than for small phylogenetic trees. All the three statistics have equally very strong power to reject H_0 .

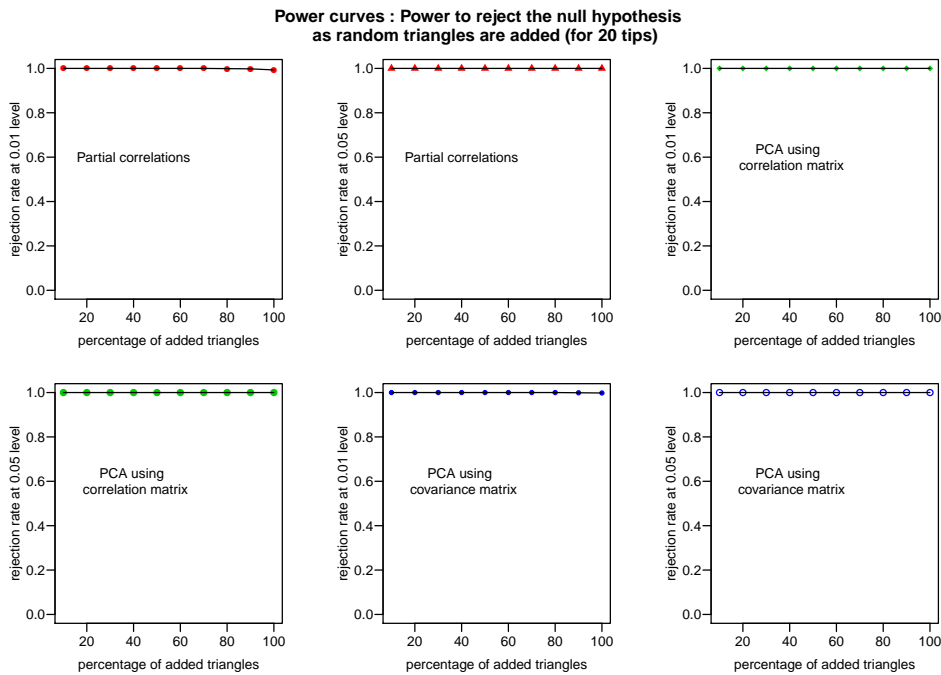


Figure 4.10: Power curves for trees with 20 tips as a result of adding random triangles to the existing association matrix.

4.2.3 Results for replacing triangles

In this approach, some of the existing triangular associations are substituted with random triangles not at corresponding positions. The number of rows of the association matrix does not change but a percentage of the triangles are substituted. Trees with 10 tips and others with 20 tips have been generated. Substitution has been done with 10%, 20%, . . . , 50% of the existing associations replaced with random triangles. An *R* code has been developed and is attached in appendix A.1.4.

Power curves for the trees with 10 tips are given in figure 4.11. In each case, 100 p values have been calculated, each with new trees whose labels are permuted 10000 times. Again, the

performance of the partial correlation statistic is daunting especially at 0.01 significance level since it has a steeper gradient than the rest.

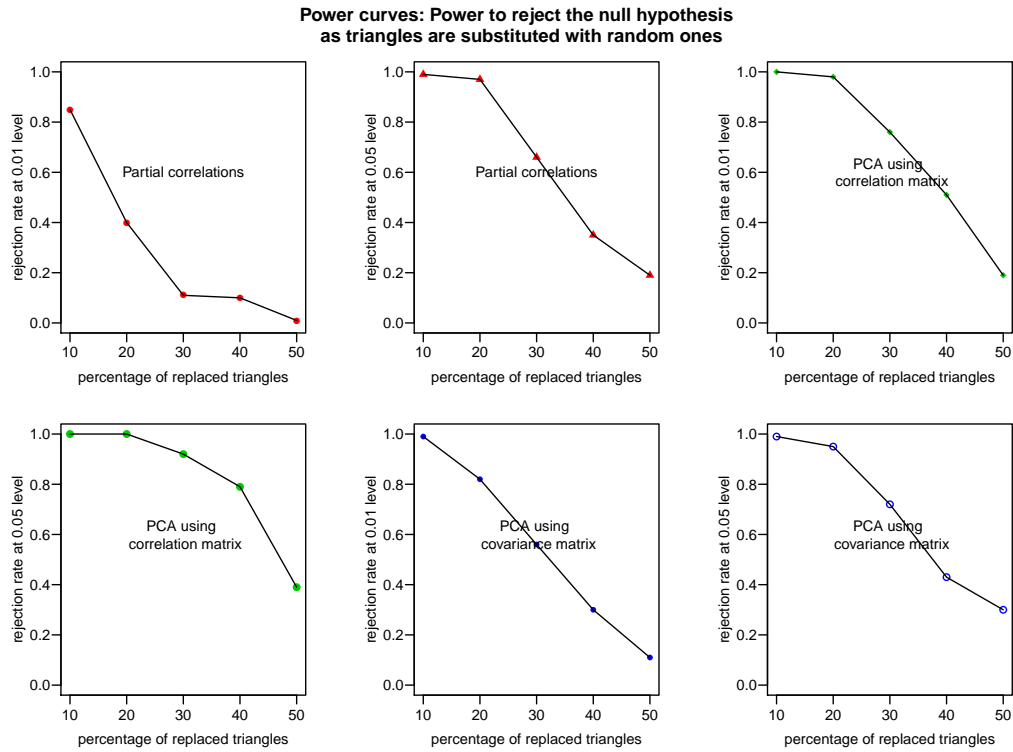


Figure 4.11: Power curves for trees with 10 tips as a result of substituting random triangles into the association matrix.

Similarly, large phylogenetic trees have been generated with all the three having 20 tips. For these trees, 1000 p values are calculated, permuting the labels 1000 times. Power to reject H_0 remain high for all the three statistics for the first 30% of substituting triangles but drops significantly as a 50% of these association triangles become substituted. The results are displayed in figure 4.12.

Power curves: Power to reject the null hypothesis as triangles are substituted with random ones

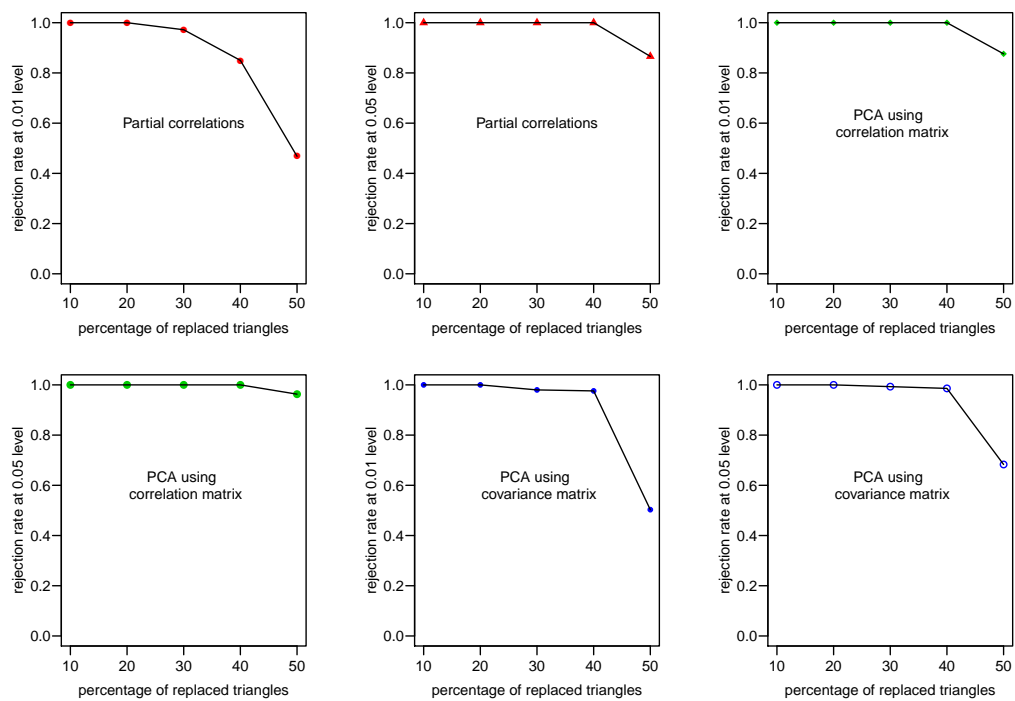


Figure 4.12: Power curves for trees with 20 tips, with random substitution of triangles.

Chapter 5

Conclusion

5.1 Discussion

This study has implemented a permutation algorithm to test the null hypothesis that the three phylogenies are unrelated which indicates that there hosts and parasites evolved independently. The three test statistics developed are useful in testing the cospeciation of the three phylogenies.

All the three test statistics give the same conclusion when applied on the data under the null hypothesis. Taking a significance level, $\alpha = 0.05$, the observed partial correlations are insignificantly small as all the p are above the significance level. A similar conclusion is made at 0.01 significance level. However, when the type I error is performed on the data, the p values from the partial correlation coefficients do not portray a uniform distribution. This implies that the test statistic could be biased. The type I error results are not any different when trees with 10 tips are used and when trees with 15 tips are used.

The empirical distribution density functions for both eigenvalues using correlation and using covariance matrix show that their p values have uniform distribution for both trees with 10 tips and 15 tips, which is a positive result.

When simulations are performed under the alternative hypothesis with trees $X = Y = Z$, and with triangular associations at their corresponding positions, the partial correlation coefficient statistic fails to be useful because the coefficients can no longer be estimated. This is because of the partial correlation formula which uses the pairwise correlation coefficients. Thus in this situation, no test of significance can be performed using this test statistic. The weakness of the partial test statistic makes the use of the eigenvalues preferred since it is not limited by this

fact. Hence, using the PCA technique to compute the two test statistics: eigenvalues under covariance and under correlation matrices is a reliable method. The proportion of variance that is explained by the first principal component is 100% and the p values from both of these statistics is zero. So, there is overwhelming evidence to reject H_0 of no relationship and conclude that these species have cospeciated.

Random triangles are added in one approach and replaced in another, to the existing association matrix whose relation is at corresponding positions. This is done in order to test the power of the three test statistics. The results obtained when 10 tips are used for each tree and when random triangles are added show a relatively poor performance of the partial correlation coefficient statistic performed at 0.01 significance level because the power drops very fast as these trees are added. The gradient is not very sharp for all the other statistics at the 0.01 and 0.05 significance level. However, when large phylogenies are used such as trees with 20 tips, the power improved and in all the three statistics, it remained almost at 100% when random trees were added up to 100%.

When the original triangles are substituted with random triangles for trees with 10 tips, the same characteristic of the partial statistic is observed. The gradient is very steep for the partial statistic at 0.01 significance level. However, when large phylogenies with 20 tips are used, the power improves a great deal and remains high for at least the first 30% of the substitutions in all the three statistics. The power curves obtained when correlation matrix is used is very high and only starts to drop from 50% substitutions. This is case for both 0.01 and 0.05 significance levels.

It is worthy noting that the geometric p value, P_{gm} computed as a summary of the three partial p values, P_x , P_y and P_z , is affected by any of these values being zero. If it happens that any of the partial p values is zero, then the final P_{gm} will also be zero indicating an overwhelming evidence to reject H_0 . This could be the reason why the statistic appear to be biased especially for small phylogenies.

It can be concluded from these findings that the three test statistics can be used to test for cospeciation of multiple phylogenies, with the PCA technique giving the most reliable test statistics. The partial correlation coefficient test statistic is less efficient for small phylogenies as its p values could be biased. When large phylogenies are used, all the three statistics have high power to reject the null hypothesis. To wind it up, the statistic obtained from the eigenvalues computed from the correlation matrix is preferred to all the other statistics that have been computed.

5.2 Further work

Analysing cospeciation of more than two phylogenies can be more complex than what has been done. This is because cospeciation is dependent on both the topologies of the phylogenies and the association matrices. Also, the smaller the number of species per tree the less complex the association matrix will be and the fewer the observations in the matrix D . Although the phylogenies used in the project have same number of species, this is not always the case. There is room to simulate multiple phylogenies with different number of species and test for cospeciation using the approaches discussed.

Different approaches need to be explored. For instance, randomization of clades (internal nodes) approach has been implemented by Hommola et al. (2009) for ditrophic host-parasite associations and could be explored as well for three host-parasite phylogenies. This involves randomizing the clades, thus modifying the phylogenies. The first tree is generated randomly to be the gold standard of the rest. The other trees will have some of their clades replaced by random clades independently. The clades that are randomized are chosen depending on how far they are from the root. The internal nodes can be ordered such that a node with a small value such as 1 or 2 or 3 appears far away from the root, whereas an internal node closer to the root will have the largest value possible. The concept is useful because randomizing the clades that are far from the root would imply that the species have evolved in identical ways for a considerably longer period of evolutionary time than if the randomization is done on clades near to the root. Randomizing clades closer to the roots of the trees would mean the species have diversified much earlier in time and this could make the phylogenies to be less similar. In the other parts of the trees where replacement did not take place, the interaction relations are assigned among the trees at corresponding positions. An R -code would have to be developed that can handle this approach effectively.

Although this project has considered the association among three phylogenies, the approach can be extended to more than three. Also, large phylogenies with more than 20 tips can be generated and explored using the same approaches. The same statistics can be tested on their performance by computing the type I error and simulating power under the alternative hypothesis.

Different multivariate statistical approaches rather than the PCA and partial correlation coefficients could be examined and permutation test applied. Other test besides permutation, such as using the Bayesian statistical methods, can be developed to test cospeciation for the multiple hosts-parasites phylogenies.

The statistical methods that have been developed in this project are based on simulated data. There is a need to apply the statistics to real data, which is not available right at the moment.

Appendix A

R programs

A.1 R-functions

Five functions have been developed to help in computations in *R*. These are:

1. `nperm`
2. `simdata`
3. `addtriangles`
4. `replacetriangles`

A.1.1 The *nperm* function

```
#####  
## This is an nperm function which #  
## calculates the observed and permuted #  
## eigenvalues and partial correlation coefficients#  
#####  
  
nperm<-function(x,y,z,relation,permut){  
# x,y,z are the phylogenetic trees.  
# permut is the number of permutations.  
if(is(x)!="phylo")  
stop(paste(sQuote("x"),"is not of class",  
sQuote("phylo")))
```

```

if(is(y)!="phylo")
stop(paste(sQuote("y"), "is not of class",
sQuote("phylo")))
if(is(z)!="phylo")
stop(paste(sQuote("z"), "is not of class",
sQuote("phylo")))

# Generate the patristic distances for each tree.
X<-cophenetic.phylo(x)
Y<-cophenetic.phylo(y)
Z<-cophenetic.phylo(z)

s=1
  n=nrow(relation)
  Dmatrix=matrix(NA,nrow=n*(n-1)/2,ncol=3)
  for(i in 1:(n-1)){
    for (j in (i+1):n){
      Dmatrix[s,1]=X[relation[i,1],relation[j,1]]
      Dmatrix[s,2]=Y[relation[i,2],relation[j,2]]
      Dmatrix[s,3]=Z[relation[i,3],relation[j,3]]
      s=s+1
    }
  }

## Computing pairwise correlations on the Dmatrix
cor_xy<-cor(Dmatrix[,1],Dmatrix[,2],method="pearson")
cor_yz<-cor(Dmatrix[,2],Dmatrix[,3],method="pearson")
cor_xz<-cor(Dmatrix[,1],Dmatrix[,3],method="pearson")
## The Observed partial correlations are
obs.pcor_xy.z<-(cor_xy-cor_yz*cor_xz)/sqrt((1-cor_yz^2)*(1-cor_xz^2))
obs.pcor_yz.x<-(cor_yz-cor_xy*cor_xz)/sqrt((1-cor_xy^2)*(1-cor_xz^2))
obs.pcor_xz.y<-(cor_xz-cor_xy*cor_yz)/sqrt((1-cor_xy^2)*(1-cor_yz^2))

## work on principal component analysis
pr.r = prcomp(Dmatrix, scale = TRUE)
covar = cov(Dmatrix)
corrm = cor(Dmatrix)
# Next compute the eigen values and vectors:
eig = eigen(covar)

```

```

eig2 = eigen(corrM)
# Now compare:
obs.val = eig$values
obs.val2 = eig2$values
ss<-summary(pr.r)

# Permute labels permut times
sizeX <- nrow(X)
sizeY <- nrow(Y)
sizeZ <- nrow(Z)

DM=list()
val= list()
val2= list()
eigenvalues_corr<-c()
eigenvalues_cov<-c()

permcov_xy<-vector(length=permut)
permcov_yz<-vector(length=permut)
permcov_xz<-vector(length=permut)
## The permuted partial correlations are
permpcov_xy.z<-vector(length=permut)
permpcov_yz.x<-vector(length=permut)
permpcov_xz.y<-vector(length=permut)

# Permute the labels
for (k in 1:permut){
  x <- sample(1:sizeX,sizeX,replace=FALSE)
  X <- X[x,]
  X <- X[,x]
  y <- sample(1:sizeY,sizeY,replace=FALSE)
  Y <- Y[y,]
  Y <- Y[,y]
  z <- sample(1:sizeZ,sizeZ,replace=FALSE)
  Z <- Z[z,]
  Z <- Z[,z]

# Form the new permuted Dmatrix

```

```

s=1
n=nrow(relation)
Dmatrix=matrix(NA,nrow=n*(n-1)/2,ncol=3)
for(i in 1:(n-1)){
  for (j in (i+1):n){
    Dmatrix[s,1]=X[relation[i,1],relation[j,1]]
    Dmatrix[s,2]=Y[relation[i,2],relation[j,2]]
    Dmatrix[s,3]=Z[relation[i,3],relation[j,3]]
    s=s+1
  }
}
DM[[k]]=Dmatrix
permcov_xy[k]<-cov(Dmatrix[,1],Dmatrix[,2],method="pearson")
permcov_yz[k]<-cov(Dmatrix[,2],Dmatrix[,3],method="pearson")
permcov_xz[k]<-cov(Dmatrix[,1],Dmatrix[,3],method="pearson")
## The permuted partial correlations are
permpcov_xy.z[k]<-(permcov_xy[k]-permcov_yz[k]*permcov_xz[k])
/sqrt((1-permcov_yz[k]^2)*(1-permcov_xz[k]^2))
permpcov_yz.x[k]<-(permcov_yz[k]-permcov_xy[k]*permcov_xz[k])
/sqrt((1-permcov_xy[k]^2)*(1-permcov_xz[k]^2))
permpcov_xz.y[k]<-(permcov_xz[k]-permcov_yz[k]*permcov_xy[k])
/sqrt((1-permcov_yz[k]^2)*(1-permcov_xy[k]^2))

val[[k]] = eigen(cov(DM[[k]]))
eigenvalues_corr[[k]]<-val[[k]]$values[1]
val2[[k]] = eigen(cov(DM[[k]]))
eigenvalues_cov[[k]]<-val2[[k]]$values[1]
}
# calculate the p values
pvalue_z<-sum(permpcov_xy.z>=obs.pcov_xy.z,na.rm=TRUE)/permut
pvalue_x<-sum(permpcov_yz.x>=obs.pcov_yz.x,na.rm=TRUE)/permut
pvalue_y<-sum(permpcov_xz.y>=obs.pcov_xz.y,na.rm=TRUE)/permut
partial_pvalue_gm<-(pvalue_z*pvalue_x*pvalue_y)^(1/3)

eigenp_corr<-sum(eigenvalues_corr>=obs.val2[1],na.rm=TRUE)/permut
eigenp_cov<-sum(eigenvalues_cov>=obs.val[1],na.rm=TRUE)/permut

return(c(partial_pvalue_gm,eigenp_corr,eigenp_cov))

```

```

### uncomment this portion to be able to plot the results
#list(covar=obs.val,corr=obs.val2,Dmatrix=Dmatrix,summary=ss,
#pr.r=pr.r,obs.pcor_xz.y=obs.pcor_xz.y,obs.pcor_yz.x=obs.pcor_yz.x,
#obs.pcor_xy.z=obs.pcor_xy.z,correlation=obs.val[1],covariance=
#obs.val2[1],relation=rel,eigenp_cov=eigenp_cov,eigenp_corr=
#eigenp_corr,partial_pvalue_gm=partial_pvalue_gm,
#permpcor_xy.z=permpcor_xy.z,permpcor_yz.x=permpcor_yz.x,
#permpcor_xz.y=permpcor_xz.y,eigenvalues_corr=eigenvalues_corr,
#eigenvalues_cov=eigenvalues_cov)
}
#####
## To come up with the Eigenvalues plots #
## displayed under the null hypothesis, #
## uncomment the list, and comment the return #
##in the nperm function above #
#####

graphics.off()
library(psych)
pairs.panels(permuted.results$Dmatrix,main=
"Pearson's pairwise correlations")
dev.print(pdf, file="Null_pairs_observed_plots.pdf")

X11(width=8.5,height=8)
par(las=1,mgp=c(1.99,0.6,0),mai=c(0.5,0.5,1,0.3))
opar<-par(mfrow=c(3,3))
par(mar=c(5, 4, 4, 2) + 0.1)
par(oma=c(0,0,4,0))

plot(permuted.results$pr.r,xlab="Principal components(1,2,3)",
main="PC variances",ylim=c(0,1.2))

plot(permuted.results$covar,main="Covariance eigenvalues",ylab=
"Eigenvalues",pch=17,
col=2,sub="principal components(1,2,3)",xlab="")
lines(permuted.results$covar,lty=2)

```

```

plot(permuted.results$corrm,main="Correlation eigenvalues",
ylab="Eigenvalues",pch=17,col=2,sub="principal
components (1,2,3)",xlab="")
lines(permuted.results$corrm,lty=2)
biplot(prcomp(permuted.results$Dmatrix, scale = TRUE))

library(matlab)
matplot(permuted.results$Dmatrix, type="l", lwd=2,
col=multiline.plot.colors(),xlab="i-th observation",
ylab="patristic distances")
par(xpd=TRUE)
lambda <- .025
legend(par("usr")[1],(1 + lambda) * par("usr")[4] -
lambda * par("usr")[3], c("Tree X", "Tree Y","Tree Z"),
xjust = 0, yjust = 0,ncol = 3,
lwd=3, lty=1, pch = "T", col = multiline.plot.colors())

library(lattice)
boxplot(permuted.results$Dmatrix,main="Boxplots for
the distances", col=c("lightgreen","red","blue"),
ylab="patristic distances",boxwex = 0.25,
xlab="1=X,2=Y,3=Z columns of Dmatrix")

library(scatterplot3d)
scatterplot3d(permuted.results$Dmatrix[,1],
permuted.results$Dmatrix[,2],permuted.results$Dmatrix[,3],
highlight.3d=TRUE,col.axis="blue", col.grid="lightblue",
main="scatterplot3d - 2", pch=21, xlab="X",
ylab="Y", zlab="Z")
dev.print(pdf, file="PCA_null.pdf")

#####
## To plot the density histograms #
#####

X11(width=6,height=5)
par(las=1,mgp=c(1.99,0.6,0),mai=c(0.8,0.6,0.4,0.3))
par(mfrow=c(3,2))

```

```

par(mar=c(5.1,4.1,0.1,2.1))
par(oma=c(0,0,4,0))

hist(permuted.results$permpcor_xy.z, sub="with Z constant",
freq=FALSE, xlim=c(-1,1),main="",
xlab="Partial correlation for X and Y")
abline(v=permuted.results$obs.pcor_xy.z, col = 2, lty = 2)
perc<-quantile(sort(permuted.results$permpcor_xy.z),
probs = 95/100)
abline(v=perc, col = "blue", lty = 2)

hist(permuted.results$permpcor_yz.x, main="",freq=FALSE,
xlim=c(-1,1),xlab="Partial corelation for Y and Z",
sub="with X constant")
abline(v=permuted.results$obs.pcor_yz.x, col = 2, lty = 2)
perc<-quantile(sort(permuted.results$permpcor_yz.x),
probs = 95/100)
abline(v=perc, col = "blue", lty = 2)

hist(permuted.results$permpcor_xz.y, main="",freq=FALSE,
xlab="Partial correlation for X and Z",xlim=c(-1,1),
sub="with Y constant")
abline(v=permuted.results$obs.pcor_xz.y, col = 2, lty = 2)
perc<-quantile(sort(permuted.results$permpcor_xz.y),
probs = 95/100)
abline(v=perc, col = "blue", lty = 2)

hist(permuted.results$eigenvalues_corr, main="",freq=FALSE,
xlab="Correlation Eigenvalues for PC1",xlim=c(1,2))
abline(v=permuted.results$correlation, col = 2, lty = 2)
perc<-quantile(sort(permuted.results$eigenvalues_corr),
probs = 95/100)
abline(v=perc, col = "blue", lty = 2)

hist(permuted.results$eigenvalues_cov, main="",freq=FALSE,
xlab="Covariance Eigenvalues for PC1",xlim=c(1,3))
abline(v=permuted.results$covariance, col = 2, lty = 2)
perc<-quantile(sort(permuted.results$eigenvalues_cov),

```

```

probs = 95/100)
abline(v=perc, col = "blue", lty = 2)
title(main="Density plots under the null hypothesis",outer=TRUE)
dev.print(pdf, file="null_partialcorrhist.pdf"

```

A.1.2 The *simdata* function

```

#####
## simdata is a function that calculates      #
## several p values such as 1000 for each of #
## the three statistics.                    #
#####

simdata<-function(nsim){
mat.results=matrix(nrow=nsim,ncol=3,dimnames =
list(c(NULL),c("partial_p","eigen_p_cor","eigen_p_cov")))
for (i in 1:nsim){
# randomly generate tree X
x=rtree(10)
labs=x$tip.label
number_labs=as.numeric(substr(labs,2,nchar(labs)))
x$tip.label=number_labs
# randomly generate tree y
y=rtree(10)
labs=y$tip.label
number_labs=as.numeric(substr(labs,2,nchar(labs)))
y$tip.label=number_labs
# randomly generate tree Z
z=rtree(10)
labs=z$tip.label
number_labs=as.numeric(substr(labs,2,nchar(labs)))
z$tip.label=number_labs

# assign a triangular association matrix
nsample<-sample(1:10,30,replace=TRUE)
rel<-matrix(nsample,nrow=10,ncol=3)
mat.results[i,]=nperm(x,y,z,rel,10000)
}

```

```

list(mat.results=mat.results)
}
null<-simdata(1000)

#####
## Plotting the ecdf of the pvalues under null #
#####

graphics.off()
X11(width=8,height=6)
par(las=1,mgp=c(1.99,0.6,0),mai=c(0.8,0.6,0.4,0.3))
par(mfrow=c(2,2))
par(mar=c(5.1,4.1,0.1,2.1))
par(oma=c(0,0,4,0))
l1<-seq(0,1, by=0.1)
l2<-seq(0,1, by=0.1)

F1 <- ecdf(null$mat.results[,1])
plot(F1, verticals=TRUE, col.points='blue',
xlim=c(0,1),col.hor='red',xlab="",
col.vert='bisque',sub="p values from
partial correlations",main="")
lines(l2,l1, col="black",lwd=1)

F2 <- ecdf(null$mat.results[,2])
plot(F2, verticals=TRUE, col.points='blue',
col.hor='red', xlim=c(0,1),
xlab="",col.vert='bisque',sub="Eigen p values
from correlation matrix",main="")
lines(l2,l1, col="black",lwd=1)

F3 <- ecdf(null$mat.results[,3])
plot(F3, verticals=TRUE, col.points='blue',
col.hor='red', xlim=c(0,1),
xlab="",col.vert='bisque',sub="Eigen p values
from covariance matrix",main="")
lines(l2,l1, col="black",lwd=1)

```

A.1.3 The *addtriangles* function

```
#####  
## A function that adds #  
## random triangles into the existing #  
## association matrix, #  
## calculates 100 p values for all the #  
## 100 new trees generated and the #  
## new triangles replaced #  
## It must call the nperm function #  
## which permutes the labels 10000 times. #  
# change addn=1 to addn=2,3,4,5,6,7,8,9,10... #  
#####  
  
addtriangles<-function(nsim){  
mat.results=matrix(nrow=nsim,ncol=3,dimnames =  
list(c(NULL),c("partial","eigen_p_cor", "eigen_p_cov")))  
for (i in 1:nsim){  
## Tree X  
x=rtree(10)  
labs=x$tip.label  
number_labs=as.numeric(substr(labs,2,nchar(labs)))  
x$tip.label=number_labs  
# generate tree y= tree x  
y=x  
# generate tree z= tree x  
z=x  
repl<-rep(1:10,3)  
rel<-matrix(repl,nrow=10,ncol=3)  
addn=1  
check=F  
while(check==F){  
nsample<-sample(1:nrow(rel),ncol(rel)*addn,replace=TRUE)  
if(max(nsample)-min(nsample)!=0){  
check=T  
}  
}  
nsample<-matrix(nsample,nrow=addn,ncol=ncol(rel))
```

```

rel<-as.matrix(rbind(rel,nsample))
mat.results[i,]=nperm(x,y,z,rel,10000)
    }
list(mat.results=mat.results)
}
add1<-addtriangles(100)
#####
# calculate how many of these are
# above the significance level.
#####
length(add1$mat.results[,1][add1$mat.results[,1]>0.05])
length(add1$mat.results[,1][add1$mat.results[,1]>0.01])
length(add1$mat.results[,2][add1$mat.results[,2]>0.05])
length(add1$mat.results[,2][add1$mat.results[,2]>0.01])
length(add1$mat.results[,3][add1$mat.results[,3]>0.05])
length(add1$mat.results[,3][add1$mat.results[,3]>0.01])

```

A.1.4 The *replacetriangles* function

```

#####
## A function that replaces                                     #
## random triangles into the existing                          #
## association matrix,                                       #
## calculates 100 p values for all the                        #
## 100 new trees generated and the                            #
## new triangles replaced                                     #
## It must call the nperm function                            #
## which permutes the labels 10000 times.                    #
## change deln=1 to addn=2,3,4,5                             #
#####

replacetriangles<-function(nsim){
mat.results=matrix(nrow=nsim,ncol=3,dimnames =
list(c(NULL),c("partial","eigen_p_cor", "eigen_p_cov")))
for (i in 1:nsim){
x=rtree(10)
labs=x$tip.label
number_labs=as.numeric(substr(labs,2,nchar(labs)))

```

```

x$tip.label=number_labs
## generate tree y = tree x
y=x
## generate tree z = tree x
z=x
repl<-rep(1:10,3)
rel<-matrix(repl,nrow=10,ncol=3)
deln= 1
del<-rel[sample(1:nrow(rel),deln),]
check=F
while(check==F){
replacen<-sample(1:nrow(rel),ncol(rel)*deln,replace=TRUE)
if(max(replacen)-min(replacen)!=0){
check=T
        }
    }
replacen<-matrix(replacen,nrow=deln,ncol=ncol(rel))
rel<-rel[-del,]
rel<-as.matrix(rbind(rel,replacen))
mat.results[i,]=nperm(x,y,z,rel,10000)
    }
list(mat.results=mat.results)
}

```

Appendix B

Supplementary Tables

B.1 Extra tables

Table B.1 shows the rotation obtained from the PCA results under the null hypothesis given in section 4.1 where the standard deviations were $PC1 = 1.078$, $PC2 = 1.004$, $PC3 = 0.911$

Table B.1: The rotation of the principal components.

	PC1	PC2	PC3
1	0.7143	-0.0341	-0.6990
2	-0.2837	0.8990	-0.3337
3	0.6398	0.4367	0.6325

The association matrix that was used to generate the above results is given in table B.2.

	X	Y	Z
1	4	9	4
2	1	1	5
3	4	3	6
4	6	10	4
5	7	1	1
6	10	10	10
7	8	8	10
8	1	3	6
9	6	9	2
10	9	4	3

Table B.2: Association matrix used under H_0 .

Table B.3 displays the matrix D that was obtained and used to compute the statistics in section 4.1.

	X	Y	Z
1	1.30	1.89	3.32
2	0.00	3.00	3.87
3	2.37	2.89	0.00
4	4.44	1.89	3.36
5	3.45	2.89	1.02
6	3.97	2.41	1.02
7	1.30	3.00	3.87
8	2.37	0.00	3.48
9	3.43	2.78	1.31
10	1.30	1.59	1.26
11	1.77	1.48	3.32
12	3.83	0.00	2.57
13	2.84	1.48	2.80
14	3.37	1.39	2.80
15	0.00	1.59	1.26
16	1.77	1.89	0.87
17	2.83	1.37	3.16
18	2.37	0.35	3.87
19	4.44	1.59	3.11
20	3.45	0.35	3.35
21	3.97	2.51	3.35
22	1.30	0.00	0.00
23	2.37	3.00	0.64
24	3.43	1.93	3.71
25	3.04	1.48	3.36

	X	Y	Z
26	2.05	0.00	1.02
27	2.58	2.40	1.02
28	1.77	0.35	3.87
29	0.00	2.89	3.48
30	2.04	1.83	1.31
31	2.49	1.48	2.83
32	4.09	1.39	2.83
33	3.83	1.59	3.11
34	3.04	1.89	2.72
35	3.55	1.37	3.20
36	3.10	2.40	0.00
37	2.84	0.35	3.35
38	2.05	2.89	2.95
39	2.56	1.83	0.86
40	3.37	2.51	3.35
41	2.58	2.41	2.95
42	1.25	2.29	0.86
43	1.77	3.00	0.64
44	2.83	1.93	3.71
45	2.04	2.78	3.32

Table B.3: Matrix D for results in section 4.1.

Table B.4 displays p values obtained under H_0 for trees with 10 tips, permuting the labels of each trees 1000 times. The type I error plots are given in figure 3.11.

	<i>partial</i>	<i>eigen_cor</i>	<i>eigen_cov</i>		<i>partial</i>	<i>eigen_cor</i>	<i>eigen_cov</i>
1	0.49	0.83	0.30	51	0.21	0.64	0.15
2	0.87	0.66	0.20	52	0.41	0.85	0.90
3	0.14	0.01	0.01	53	0.77	0.22	0.43
4	0.19	0.42	0.26	54	0.62	0.64	0.32
5	0.41	0.76	0.36	55	0.09	0.15	0.26
6	0.49	0.46	0.52	56	0.16	0.10	0.29
7	0.56	0.03	0.03	57	0.73	0.67	0.32
8	0.45	0.87	0.50	58	0.24	0.22	0.77
9	0.28	0.45	0.61	59	0.00	0.03	0.00
10	0.38	0.16	0.10	60	0.65	0.40	0.92
11	0.96	0.09	0.27	61	0.54	0.47	0.39
12	0.32	0.01	0.48	62	0.35	0.91	0.84
13	0.73	0.94	0.55	63	0.31	0.32	0.82
14	0.36	0.95	0.60	64	0.74	0.67	0.24
15	0.56	0.92	0.03	65	0.44	0.80	0.21
16	0.66	0.07	0.73	66	0.62	0.79	0.21
17	0.42	0.47	0.34	67	0.28	0.26	0.76
18	0.58	0.63	0.58	68	0.44	0.64	0.05
19	0.48	0.51	0.78	69	0.69	0.97	0.03
20	0.35	0.20	0.51	70	0.25	0.08	0.83
21	0.69	0.96	0.82	71	0.16	0.18	0.37
22	0.17	0.23	0.38	72	0.19	0.34	0.18
23	0.33	0.76	0.23	73	0.62	0.19	0.26
24	0.74	0.27	0.09	74	0.89	0.41	0.21
25	0.64	0.94	0.39	75	0.26	0.61	0.98
26	0.37	0.36	0.90	76	0.47	0.85	0.25
27	0.36	0.96	0.80	77	0.68	0.68	0.49
28	0.33	0.02	0.19	78	0.44	0.30	0.88
29	0.38	0.36	0.40	79	0.31	0.04	0.71
30	0.18	0.09	0.25	80	0.31	0.79	0.16
31	0.37	0.90	0.75	81	0.62	0.33	0.11
32	0.68	0.06	0.42	82	0.51	0.61	0.84
33	0.30	0.48	0.00	83	0.40	0.69	0.32
34	0.46	0.86	0.99	84	0.35	0.74	0.24
35	0.78	0.24	0.52	85	0.63	0.59	0.80
36	0.18	0.47	0.43	86	0.61	0.74	0.96
37	0.65	0.66	0.08	87	0.26	0.27	0.73
38	0.63	0.53	0.90	88	0.31	0.92	0.14
39	0.19	0.25	0.69	89	0.58	0.92	0.85
40	0.49	0.81	0.36	90	0.19	0.65	0.70
41	0.15	0.13	0.86	91	0.64	0.36	0.45
42	0.54	0.98	0.12	92	0.50	0.56	0.36
43	0.12	0.07	0.30	93	0.92	0.47	0.38
44	0.69	0.45	0.71	94	0.36	0.75	0.83
45	0.25	0.48	0.74	95	0.10	0.06	0.86
46	0.36	0.99	0.98	96	0.65	0.51	0.12
47	0.20	0.44	0.77	97	0.15	0.06	0.01
48	0.69	0.77	0.30	98	0.18	0.11	0.56
49	0.65	0.26	0.04	99	0.29	0.41	0.44
50	0.44	0.75	0.02	100	0.57	0.89	0.79

Table B.4: 100 p values for trees with 15 tips, permuting the labels 1000 times.

Bibliography

- Ahmad, F., Aslam, M. and Razaq, M. (2004). Chemical ecology of insects and tritrophic interactions, *Journal of Research (Science)* **15**: 181–190.
- Becerra, J. X. (1997). Insects on Plants: Macroevolutionary chemical trends in host use, *Journal of Science* **276**: 253–56.
- Brooks, D. R. and McLennan, D. A. (1991). *Phylogeny, ecology and behaviour*, The university of Chicago Press, Chicago and London.
- Choi, K. and Gomez, S. M. (2009). Comparison of phylogenetic trees through alignment of embedded evolutionary distances, *Journal of BMC Bioinformatics* **10**: 423.
- Crawley, M. J. (2007). *The R Book*, John Wiley & Sons.
- Durbin, R., Eddy, S. R., Krogh, A. and Mitchson, G. (2009). *Biological Sequence Analysis*, Vol. Thirteenth printing of *Probability models of proteins and nucleic acids*, Cambridge university press.
- Everitt, B. S. and Dunn, G. (2001). *Applied Multivariate Data Analysis*, Vol. 2nd Edition, Hodder Arnold.
- Ewens, W. J. and Grant, G. R. (2001). *Statistical Methods in Bioinformatics*, Statistics for Biology and Health, Springer.
- Fahrenheit, H. (1913). Ectoparasiten und Abstammungslehre, *Journal of Zoology* **41**: 371–374.
- Fourment, M. and Gibbs, M. J. (2006). Patristic: A program of calculating patristic distances and graphically comparing the components of genetic changes, *Journal of BMC Evolution Biology* **6**: 1.

- Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. and Dudoit, S. (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Healthy, Springer Science+Business Media, Inc.
- Hafner, M. S., Sudman, P. D., Villablanca, F. X., Spradling, T. A., Demastates, J. W. and Nadler, S. A. (1994). Disparate rates of molecular evolution in cospeciating hosts and parasites, *Journal of Science* **265**: 1087–90.
- Hommola, K., Smith, J. E., Qiu, Y. and Gilks, W. R. (2009). A Permutation Test of Host-Parasite Cospeciation, *Journal of Molecular Biology Evolution* **26**: 1457–1468.
- Huelsenbeck, J. P., Rannala, B. and Larget, B. (2000). A Bayesian framework for the analysis of cospeciation, *Journal of Evolution* **54**: 352–364.
- Huelsenbeck, J. P., Ronquist, F., Nielsen, R. and Bollback, J. P. (2001). Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology, *Science Journals: Review* **294**: 2310–2314.
- Jackson, J. E. (1991). *A user's guide to principal components*, New York: Wiley-Interscience.
- Jobson, J. D. (1991). *Applied Multivariate Data Analysis*, Vol. 1 of *Regression and Experimental Design*, Springer-Verlag.
- Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*, Vol. 3rd Edition, Prentice-Hall International Editions.
- Jolliffe, I. T. (1986). *Principal components analysis*, New York: Springer-Verlag.
- Klassen, G. J. (1992). A history of the macroevolutionary approach to studying host-parasite associations, *Journal of parasitology* **78**: 573–87.
- Lapointe, F.-J. and Legendre, P. (1992). Statistical significance of the matrix correlation coefficient for comparing independent phylogenetic trees, *Journal of Systematic Biology* **41**: 378–384.
- Legendre, P., Desdevis, Y. and Bazin, E. (2002). A statistical test for host-parasite coevolution, *Journal of Systematic Biology* **51**: 217–234.
- Manly, B. F. J. (2005). *Multivariate Statistical Methods*, Vol. 3rd Edition of *A primer*, Chapman and Hall/CRC.

- Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Vol. Probability and Mathematical Statistics of *Monographs and Textbooks*, Academic Press, Inc, San Diego, CA92101, University of Leeds, UK.
- Micha, S. G., Kistenmacher, S., Mölck, G. and Wyss, U. (2000). Tritrophic interactions between cereals, aphids and parasitoids: Discrimination of different plant-host complexes by *Aphidius Rhopalosiphii*, *European Journal of Entomology* **97**: 539–543.
- Moran, N. A., VanDohlen, C. D. and Baumann, P. (1995). Faster evolutionary rates in endosymbiotic bacteria than in cospeciating hosts, *Journal of Molecular Evolution* **41**: 727–31.
- Page, R. D. M. (1990b). Temporal congruence and cladistic analysis of biogeography and cospeciation, *Journal of Systematic Zoology* **39**: 205–26.
- Page, R. D. M. (1996b). Temporal congruence revisited: Comparison of mitochondrial DNA sequence divergence in cospeciating pocket gophers and their chewing lice, *Journal of Systematic Biology* **45**: 151–67.
- Page, R. D. M. (2003). *Tangled trees: Phylogeny, cospeciation, and coevolution*, The University of Chicago Press, Chicago and London.
- Paradis, E. (2006). *Analysis of phylogenetics and evolution with R*, Use R, Springer, New York.
- Paterson, A. M. and Gray, R. D. (1997). *Host-Parasite cospeciation, host switching and missing the boat*, Vol. Probability and Mathematical Statistics of *Host-Parasite Evolution: General Principles and Avian Models*, Oxford: Oxford Press, University of Leeds, UK.
- Paterson, A. M., Palma, R. L. and Gray, R. D. (1999). How frequently do avian lice miss the boat?, *Journal of Systematic Biology* **48**: 214–23.
- Pearson, K. (1901). On lines and planes of closet fit to a system of points in space, *Journal of Philosophical Magazine* **2**: 557–572.
- R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: <http://www.R-project.org>
- Timm, N. H. (2002). *Applied Multivariate Analysis*, Springer Texts in Statistics, Springer-Verlag New York, Inc.