UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Métodos para a avaliação da integração entre caracteres filogenéticos discretos

Maria Luiza Matos Silva

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Maria Luiza Matos Silva

Métodos para a avaliação da integração entre caracteres filogenéticos discretos

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Rafael Izbicki

USP – São Carlos
Novembro de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

Silva, Maria Luiza Matos
S586m Métodos para a avaliação da integração entre caracteres filogenéticos discretos / Maria Luiza Matos Silva; orientador Rafael Izbicki. -- São Carlos, 2024.
61 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Integração. 2. Filogenia. 3. Phylogenetic Logistic Regression. 4. Threshold Model. 5. Cluster. I. Izbicki, Rafael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Maria Luiza Matos Silva

Methods for evaluation the integration between discrete phylogenetic characters

Dissertation submitted to the Institute of Mathematics and Computer Science – ICMC-USP and to the Department of Statistics – DEs-UFSCar – in accordance with the requirements of the Statistics Interagency Graduate Program, for the degree of Master in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Rafael Izbicki

USP – São Carlos
November 2024

AGRADECIMENTOS

Agradeço primeiramente à Deus por toda a proteção, força, luz e discernimento em todos os momentos da minha vida.

Agradeço aos meus pais, Iranildes e Josias por todo o apoio, incentivo e principalmente amor que sempre tiveram. Agradeço também a toda a minha família que sempre esteve presente em minha vida.

Agradeço ao meu orientador Rafael que me acompanha desde a graduação. Por todo ensinamento, suporte, paciência e incentivo em todos os momentos ao longo desses anos.

Agradeço a minha psicóloga Beatriz que foi extremamente fundamental na minha trajetória, auxiliando em todos os momentos, principalmente de medo e incerteza nesses últimos anos.

Agradeço aos meus amigos por compartilharem de tantos momentos de alegrias, aprendizados, dúvidas e tristezas.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“A cada dia basta o seu cuidado.”
(Mt 6,34)

RESUMO

SILVA, M. L. M. **Métodos para a avaliação da integração entre caracteres filogenéticos discretos**. 2024. 63 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Filogenética é a área que busca compreender as relações existentes entre diferentes organismos no que diz respeito ao seu desenvolvimento e evolução. Uma questão fundamental é avaliar a integração e a modularidade de diferentes características de indivíduos. A integração refere-se a associação entre as características e a modularidade trata da investigação de grupos de caracteres que possuem mais dependência com uns do que com outros. Dentro desse campo de estudo, existem uma abundância de trabalhos com dados contínuos, porém há poucos artigos para o caso discreto. Neste trabalho propomos uma abordagem para avaliar a integração entre caracteres filogenéticos discretos, e para isso a metodologia consiste em dois passos. A primeira etapa é calcular a similaridade entre os caracteres, utilizando correlações simples (Pearson e Spearman) e com a utilização da topologia (*Threshold Model* e *Phylogenetic Logistic Regression* - PLR). Na utilização da PLR consideramos os valores absolutos dos coeficientes e o valor-p como medidas de associação. O segundo passo consiste em utilizar a informação obtida anteriormente para construir o *Cluster* hierárquico, a fim de se obter a visualização dos módulos. Utilizamos base de dados simulados dos modelos de Markov e *Threshold*. Para confrontar os resultados de cada técnica, empregamos três métricas: *Rand Index* (RI), *Normalized Mutual nformation* (NMI) e o *Fowlkes Mallows Index* (FMI). Assim, pudemos avaliar como a incorporação da informação sobre a filogenia impacta nas análises por meio da simulação dos dados.

Palavras-chave: Integração, Filogenia, *Phylogenetic Logistic Regression*, *Threshold Model*, *Cluster*.

ABSTRACT

SILVA, M. L. M. **Methods for evaluation the integration between discrete phylogenetic characters**. 2024. 63 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Phylogenetics is the field that aims to understand the relationships between different organisms in terms of their development and evolution. A key question in this area is how to analyze the integration and modularity of different characteristics of individuals. Integration refers to the association between characteristics, while modularity focuses on the investigation of groups of characters that have greater dependence on some than others. Despite the abundance of papers in this field that use continuous data, there are fewer papers that focus on the discrete case. In this paper, we present an approach for evaluating the integration between discrete phylogenetic characters, for this the methodology consisting of two steps. The first step is to calculate the similarity between characters using simple correlations (Pearson and Spearman) and by utilizing topology (*Threshold Model* and *Phylogenetic Logistic Regression* - PLR). In using PLR, we consider the absolute values of the coefficients and the p-value as measures of association. The second step involves using the information obtained in step one to build a hierarchical *Cluster*, in order to visualize modules. We use simulated datasets from Markov and *Threshold* models. To compare the results of each technique, we employ three metrics: *Rand Index* (RI), *Normalized Mutual nformation* (NMI) e o *Fowlkes Mallows Index* (FMI). This allows us to assess how incorporating phylogenetic information impacts the analyses through data simulation.

Keywords: Integration, Phylogeny, *Phylogenetic Logistic Regression*, *Threshold Model*, *Cluster*.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A busca pela compreensão da diversidade biológica e a origem dos organismos tem sido objeto de discussões há bastante tempo, provocando diversos questionamentos. Entender as complexas relações evolutivas entre as espécies não é uma tarefa trivial, e tem mobilizado muitos cientistas de diversas áreas ao longo dos anos.

Algumas obras foram verdadeiros marcos na história. Como por exemplo, *"The origin of species"* de Darwin (1859) que explorou sobre a relação de descendência entre as espécies e *"Morphological Integration"* de Olson e Miller (1999) que impulsionou diversos estudos a cerca desse tema. Com essas e outras descobertas, combinadas com os avanços tecnológicos obtidos no decorrer dos anos, inúmeras outras possibilidades de pesquisa e entendimento se tornaram viáveis.

Para compreender a diversidade biológica, é importante assimilar as informações acerca dos processos de desenvolvimento evolutivo dos seres (MATIOLI; FERNANDES, 2012). Assim, é denominada de filogenética a área que investiga as relações existentes entre diferentes organismos, no que diz respeito ao seu desenvolvimento e evolução.

Os métodos comparativos têm o intuito de analisar as semelhanças e diferenças entre objetos de estudo, de modo a entender suas origens, estruturas, padrões. Esse conceito é utilizado em diversos setores, como antropologia, ecologia, linguística, biologia, entre outros (CORNWELL; NAKAGAWA, 2017).

Na biologia, temos os métodos filogenéticos comparativos (do inglês, *Phylogenetic comparative methods* - PCM). Os PCMs possibilitam a insvestigação da história da evolução dos organismos e, de acordo com Cornwell e Nakagawa (2017), geralmente utilizam uma combinação de dois tipos de dados, que são: estimativas de parentesco entre espécies e valores de características (covariáveis, ou também chamadas de caracteres na biologia) em indivíduos existentes. Além disso, também podem incluir várias outras informações, como por exemplo, a respeito de registros fósseis.

Dois pontos centrais nessa discussão são a integração e a modularidade. A integração morfológica refere-se à associação entre características, ou seja, a correlação entre as variáveis em questão. Já um módulo é uma unidade capaz de evoluir e variar independentemente de outro módulo. Assim, a modularidade trata da investigação de grupos de características que possuem maior dependência com determinados grupos do que com outros (GOSWAMI; POLLY, 2010).

Portanto, integração e modularidade estão relacionadas com a correlação existente entre as variáveis estudadas. Sendo assim, o objetivo deste trabalho é conseguir analisar características que variam conjuntamente, identificando os módulos correspondentes.

1.1 Motivação

Atualmente, existem várias maneiras de se conseguir analisar as questões evolutivas, e a maioria dos estudos desenvolvidos consideram variáveis contínuas. Mitov *et al.* (2020), por exemplo, apresentam um algoritmo eficiente para o cálculo da verossimilhança na inferência filogenética; O'Meara *et al.* (2006) analisam mudanças nas taxas evolutivas; e Goswami e Polly (2010) fazem uma revisão de algumas das técnicas existentes.

Uma das formas de analisar a associação entre características é através da avaliação da correlação. As correlações evolutivas, segundo Harmon (2019), acontecem quando dois caracteres tendem a evoluir conjuntamente em virtude de algum processo, como por exemplo, mutação ou seleção natural.

No entanto, utilizar apenas a correlação para avaliar o comportamento das características pode não ser suficiente. No contexto filogenético, podemos encontrar casos em que duas variáveis podem apresentar uma significativa correlação, mas isso não necessariamente evidencia que elas evoluem conjuntamente.

Para demonstrar esse fenômeno, Harmon (2019) simulou uma árvore com 100 espécies e analisou duas características, denominadas X e Y. As variáveis foram geradas de forma a evoluírem independentemente, mas, como podemos observar na Figura 1, há uma aparente associação entre X e Y, que se deve à filogenia das espécies, ou seja, devido a ancestralidade compartilhada entre elas.

Para exemplificar esse tipo de situação de maneira mais prática, podemos considerar a seguinte situação: temos dois grupos de espécies que compartilham um ancestral comum e estamos observando as características de ter ovo amniótico e ter pelo. Essas variáveis podem estar evoluindo independentemente, mas devido à dependência entre os indivíduos dentro de cada família, eles serão parecidos, e então podemos perceber que as características serão correlacionadas. Isso ocorre devido ao fato das espécies estarem evoluindo sob uma mesma árvore. Logo, os caracteres só apresentam correlação por estarem relacionados a filogenia, e não porque de fato evoluem conjuntamente.

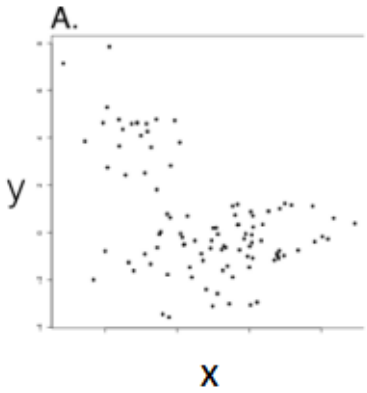


Figura 1 – Exemplo simulado.

Fonte: Harmon (2019).

Assim, uma forma de tentar contornar essa questão é incorporando a topologia em uma medida de associação. Dessa maneira, a avaliação das relações evolutivas entre as características poderá ser mais precisa.

Desta forma, este trabalho busca avaliar a integração entre caracteres filogenéticos discretos. Para isso, a abordagem é dividida em dois passos. O primeiro consiste em calcular a similaridade entre os caracteres, utilizando correlações simples (Pearson e Spearman) e incorporando a topologia (*Threshold Model* e *Phylogenetic Logistic Regression*). O segundo passo consiste em utilizar a informação obtida anteriormente para construir o *Cluster* hierárquico, a fim de se obter a visualização dos módulos.

O trabalho é organizado da seguinte forma: o Capítulo 2 apresenta alguns conceitos básicos relacionados à filogenética, estrutura dos dados, notações e uma revisão de técnicas. No Capítulo 3, discute-se a metodologia proposta neste trabalho. O Capítulo 4 detalha como foram realizadas as simulações dos dados, e no Capítulo 5, é apresentada as métricas utilizadas e a comparação dos resultados resultados obtidos. Por fim, temos o Capítulo 6 com as conclusões.

CAPÍTULO

2

CONCEITOS BÁSICOS

Apresentaremos neste capítulo alguns conceitos de filogenética na Seção 2.1, para facilitar o entendimento de algumas nomenclaturas ao longo desse estudo. Na Seção 2.2, descrevemos a estrutura dos dados a serem analisados, e na Seção 2.3, introduzimos algumas notações utilizadas no decorrer do trabalho. Também incluímos uma revisão das técnicas que serão aplicadas, em que o *Threshold Model* é discutido na Seção 2.4, as técnicas de correlação de Pearson e Spearman são explicadas na Seção 2.6, a *Phylogenetic Logistic Regression* é apresentada na Seção 2.5, e a análise de *Cluster* é detalhada na Seção 2.7. Por fim, a Seção 2.8 descreve as métricas utilizadas para a comparação dos resultados.

2.1 Filogenética

O termo filogenia, de maneira simplificada, diz respeito à história evolutiva das espécies. Assim, a filogenética busca compreender as relações existentes entre diferentes organismos em termos de seu desenvolvimento e evolução.

Em 1950, Willi Hennig propôs a sistemática filogenética, que conseguiu unificar diversos aspectos para a criação de um sistema que retrata os efeitos do processo evolutivo. Com isso, é possível estabelecer relações de parentesco entre os indivíduos com base na filogenia e também realizar a classificação biológica (MATIOLI; FERNANDES, 2012).

Além de retratar a diversidade biológica, a sistemática filogenética também estabelece as relações de parentesco dos indivíduos, de evolução das características morfológicas, comportamentais e outras, segundo Matioli e Fernandes (2012).

Cavalli-Sforza e Edwards (1967) realizaram uma pesquisa sobre a reconstrução de filogenias utilizando dados genéticos. Isso foi um importante passo nesse campo de estudo, em razão de sua abordagem estatística que viabilizou a expansão a outros tipos de dados (PARADIS, 2014).

Ao longo dos anos as contribuições de Felsentein foram bastante significativas. Ele utilizou caracteres contínuos para realizar a estimação de árvores pela verossimilhança (FELSENSTEIN, 1973), tempo depois propôs o cálculo de contrastes filogeneticamente independentes (*phylogenetically independent contrasts - PIC*), discutindo pontos sobre a não independência das unidades (FELSENSTEIN, 1985).

A filogenia compartilhada entre os indivíduos faz com que não sejam independentes, e isso acarreta na falha da suposição de algumas técnicas estatísticas. Nesse sentido, a incorporação da árvore pode ter um papel fundamental nas análises realizadas.

A árvore filogenética é uma maneira de visualizar graficamente as relações de ancestralidade, como mostrado na Figura 2. Uma árvore filogenética é composta por ramos, nós internos e nós terminais. Com base na Figura 2 temos que os ramos estão representados pelos números, os nós internos pelas letras maiúsculas e os nós terminais pelas espécies.

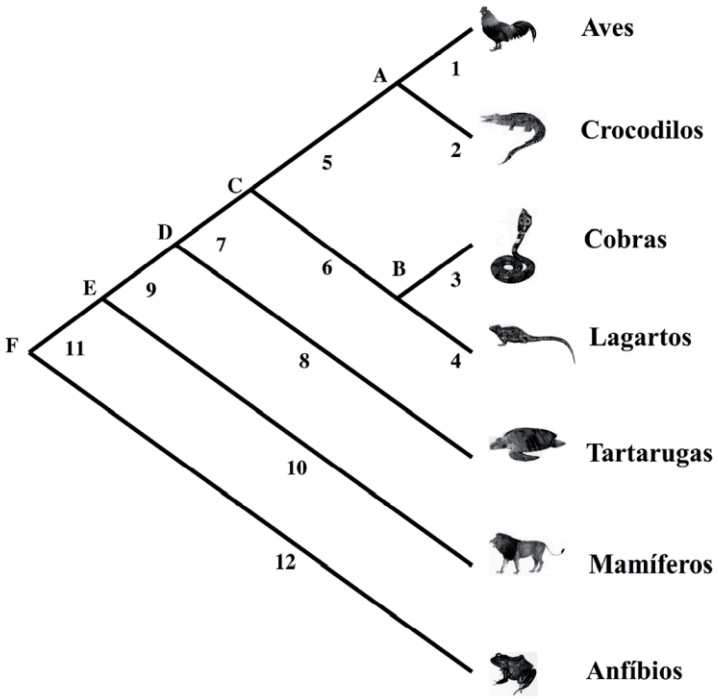


Figura 2 – Exemplo de uma árvore filogenética.

Fonte: Matioli e Fernandes (2012).

2.2 Estrutura dos dados

Os dados filogenéticos representados em uma árvore descrevem a história evolutiva de espécies ou populações (a depender do objetivo estudado). Isso significa que as informações presentes no final da árvore, ou seja, nos nós terminais, correspondem aos elementos atuais, e ao avançar a análise dessa representação em uma linha temporal mais antiga observamos unidades ancestrais extintas.

A Figura 3 simboliza a estrutura de uma árvore filogenética. Nela podemos observar que todos os nós terminais (indivíduos) são advindos de um ancestral comum. Também notamos que a cor azul corresponde a um ramo e o quadriculado vermelho representa um clado, que é um conjunto de organismos originados de um mesmo ancestral comum. Por fim, temos que os nós internos são pontos em que ocorreram especiação, isso quer dizer que a partir desse processo houve uma divisão dos organismos, tornando-os distintos.

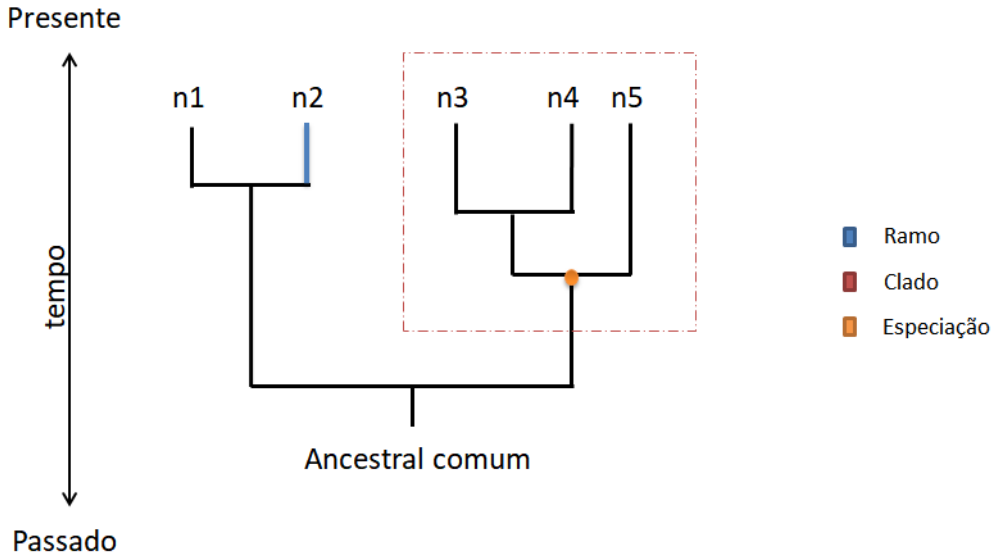


Figura 3 – Esquema de uma árvore filogenética.

Em biologia, as covariáveis são comumente chamadas de características ou caracteres, portanto essas palavras também serão utilizadas como sinônimos para descrever as variáveis neste trabalho.

As características presentes em estudos filogenéticos podem corresponder a informações genéticas ou morfológicas, por exemplo. A título de exemplo podemos pensar na presença de asas, se um indivíduo é vertebrado, o comprimento do crânio, entre outros. Na Tabela 1 podemos observar uma ilustração de como é a estrutura dos dados, em que $n_1, ..., n_5$ são os indivíduos e $X_1, ..., X_7$ correspondem as características, que assumem os valores 0 ou 1, indicando por exemplo, a presença ou ausência da variável em questão.

Tabela 1 – Exemplo de uma tabela de dados com caracteres filogéneticos binários.

	X1	X2	X3	X4	X5	X6	X7
n1	1	1	0	1	0	1	0
n2	1	1	0	0	0	0	0
n3	1	0	1	0	1	0	0
n4	1	0	1	0	0	0	0
n5	1	0	0	0	0	0	0

2.3 Notação

Definiremos as notações gerais utilizadas neste trabalho como forma de padronizar os conceitos matemáticos. Matrizes serão denotadas em letra maiúscula e em negrito e os demais em fonte normal. Temos que:

- $X_{i,j}$ corresponde a variável aleatória discreta que corresponde à característica filogenética j na observação i, com $i \in \{1,...,n\}$ e $j \in \{1,...,p\}$;
- n: representa o número de espécies;
- p: indica o número de covariáveis.

2.4 *Threshold Model*

O *Threshold Model* foi proposto por Sewal Wright em 1934. Nos últimos tempos essa técnica começou a ser empregada em análises comparativas, e o que ela faz é modelar como as características evoluem dada uma topologia. Felsenstein (2012), por exemplo, discorre sobre uma forma de se utilizar o *Threshold Model* entre caracteres discretos, e entre discreto e contínuo.

A ideia principal desse modelo é assumir que existe um caractere contínuo que não é observável, e é denominado de *liability*, sendo capaz de controlar a característica discreta perceptível analisada. Assim, quando a *liability* ultrapassa o *threshold*, que é um valor fixo definido, a característica muda de valor (REVELL, 2014). A Figura 4 exemplifica esse funcionamento.

Na Figura 4 podemos visualizar o *threshold* assumindo valor 0, e quando temos a *liability* excedendo esse ponto o valor do caractere discreto é 1, caso contrário será 0.

Felsenstein (2002) sugere utilizar o *Brownian Motion* (BM) para estimar a *liability*, que será simulada nos ramos da árvore. O BM é um processo de Markov com espaço de estados e transições a tempo contínuo. Essa técnica é muito utilizada para descrever o comportamento de fenômenos da natureza, como por exemplo o movimento da poeira no ar (PINSKY; KARLIN, 2010).

O *Brownian Motion* tornou-se popular em biologia comparativa, e segundo Harmon (2019), pode ser utilizado para modelar a evolução do caractere no tempo. Ao fazer isso estamos

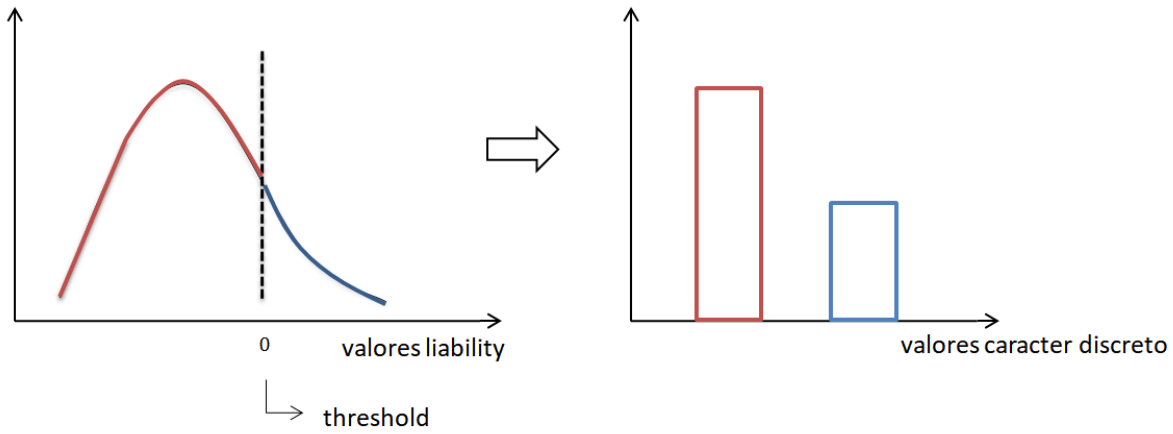


Figura 4 – Esquema da ideia central do *Threshold Model*.

Adaptação: Felsenstein (2005).

analisando o comportamento do valor médio da característica em um instante de tempo t, que é definido como $\bar{z}(t)$.

Um BM pode ser especificado como tendo:

- $E[\bar{z}(t)] = \bar{z}(0)$: isso significa que o valor esperado da média do caractere em qualquer tempo t corresponde ao seu valor no instante inicial,
- $\bar{z}(t) - \bar{z}(0)$ e $\bar{z}(t+\Delta(t)) - \bar{z}(t)$ são independentes : quaisquer 2 incrementos não sobrepostos são independentes,
- $\bar{z}(t) \sim N(\bar{z}(0), \sigma^2 t)$: segue distribuição normal.

Em uma árvore filogenética as espécies têm uma trajetória compartilhada, pois evoluem de ancestrais em comum, fazendo com que elas apresentem semelhanças. O *Brownian Motion* pode ser utilizado para estimar a covariância entre os caracteres das espécies, que é o objeto de estudo deste trabalho.

Para introduzir o BM no contexto filogenético, iremos considerar um exemplo do Harmon (2019). Suponha que uma característica (denotada x) evolua sob uma árvore filogenética com duas espécies (a e b), como na Figura 5.

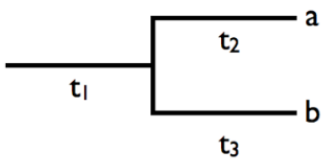


Figura 5 – Árvore para exemplificar um modelo de *Brownian Motion*.

Fonte: Harmon (2019).

O valor médio da característica na espécie a evolui sob o *Brownian Motion* de modo que $\bar{x}_a(t_1+t_2) \sim N(\bar{x}(0), \sigma_B^2(t_1+t_2))$, em que σ_B^2 corresponde a variância das características ao longo do tempo (HARMON, 2019). Para a espécie b ocorre de forma análoga, $\bar{x}_b(t_1+t_3) \sim N(\bar{x}(0), \sigma_B^2(t_1+t_3))$.

Como sabemos, a e b não são independentes, pois estão compartilhando o ramo 1. Desse modo, cada espécie corresponde as mudanças evolutivas denotadas na Equação 2.1:

$$\bar{x}_a = \Delta\bar{x}_1 + \Delta\bar{x}_2, \bar{x}_b = \Delta\bar{x}_1 + \Delta\bar{x}_3. \tag{2.1}$$

em que $\Delta\bar{x}_k$ com $k = 1,2,3$ corresponde a evolução de cada ramo da árvore, e são normalmente distribuídos com média zero e variância respectivamente, $\sigma_B^2 t_1$, $\sigma_B^2 t_2$ e $\sigma_B^2 t_3$.

$\bar{x}_a$ e $\bar{x}_b$ são normais, pois são soma de normais. Se calcularmos a covariância entre os traços dessas espécies estaremos simplesmente calculando a variância do ramo compartilhado:

$$cov(\bar{x}_a, \bar{x}_b) = var(\Delta\bar{x}_1) = \sigma_B^2 t_1. \tag{2.2}$$

Conseguimos também definir uma matriz de variância covariância filogenética, que é denotada por **C** (HARMON, 2019).

$$\mathbf{C} = \begin{bmatrix} \sigma_B^2(t_1+t_2) & \sigma_B^2 t_1 \\ \sigma_B^2 t_1 & \sigma_B^2(t_1+t_3) \end{bmatrix},$$

em que a diagonal principal é representada pelas distâncias totais de cada espécie até a raiz da árvore, e as demais entradas como o tamanho dos ramos compartilhados entre os pares de indivíduos, como visto.

Como na maioria das vezes estamos interessados em estudar várias variáveis de uma vez, a análise multivariada torna-se indispensável. Expandindo o exemplo utilizado até aqui para duas características, também podemos construir uma segunda matriz que diz respeito aos caracteres, chamada de **R**.

$$\mathbf{R} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

em que a diagonal principal corresponde a taxa líquida da evolução para cada caractere, e as demais entradas representam as covariâncias evolutivas entre eles, σ_{12} é justamente o que queremos estimar neste trabalho, pois corresponde a integração entre as características.

As matrizes descritas anteriormente são combinadas pelo produto de Kroeneker, como descrito na Equação 2.3.

$$\mathbf{V} = \mathbf{R} \otimes \mathbf{C} \tag{2.3}$$

em que, $\mathbf{V}$ representa a matriz de variância covariância das características nas espécies.

Por último, definimos os valores médios dos caracteres no instante zero como:

$$\mathbf{A} = \begin{bmatrix} \vec{x_1}(0) \\ \vec{x_2}(0) \end{bmatrix}.$$

em que $\vec{x_i}(0)$ com $i = 1,2$ é um vetor com dimensão $1x2$ que contém os valores médios do caractere no instante de tempo zero para cada espécie.

Com isso, temos que as características seguem distribuição normal multivariada com média $\mathbf{A}$ e matriz de variância covariância $\mathbf{V}$.

Para ajustar o *Threshold Model* foi o utilizado o pacote *phytools* (REVELL, 2024) do R, que considera a abordagem *Markov Chain Monte Carlo* (MCMC) para amostrar as *liabilities* e conseguir estimar a matriz de covariância.

2.5 *Phylogenetic Logistic Regression*

Nelder e Wedderburn (1972) propuseram os Modelos Lineares Generalizados (MLG), popularmente chamados de MLG, como uma forma de incluir em um mesmo conjunto vários modelos estatísticos. Essa classe de modelos permite o uso de diferentes distribuições para a variável resposta, de forma que ela pertença à família exponencial de distribuições, como: Binomial, Poisson, Normal, entre outras.

Para que uma distribuição pertença a família exponencial, sua função de probabilidade pode ser representada como:

$$f(y,\theta,\phi) = \exp \phi [y\theta - b(\theta)] + c(y,\phi), \tag{2.4}$$

em que $b(.)$ e $c(.)$ são funções conhecidas e ϕ é um parâmetro de dispersão (OLSSON, 2002).

A *Phylogenetic Logistic Regression* (PLR) é uma extensão dos modelos lineares generalizados. Um MLG é formado por três principais elementos: componente aleatório, sistemático e a função de ligação. Em um modelo filogenético, há a adição de um quarto elemento, que é a árvore filogenética.

- Componente aleatório: conjunto de variáveis aleatórias independentes que possuem distribuição pertencente a família exponencial;

- Componente sistemático: preditor linear $X\beta$;
- Função de ligação: é uma função que realiza a junção dos elementos citados acima e denominada de $g(.)$.

Quando a variável resposta é binária, ela apresenta os valores 0 ou 1, seguindo a distribuição Bernoulli, a função de probabilidade é dada por:

$$P(Y = y) = \mu^y (1-\mu)^{1-y}. \tag{2.5}$$

em que $\mu = P(Y = 1|\mathbf{X})$.

A PLR é similar a regressão logística padrão. Assim, utilizamos a logito como função de ligação e temos:

$$g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right), \tag{2.6}$$

em que $i = 1, 2, ..., n$.

O modelo PLR proposto por Ives e Jr (2010) assume que a variável resposta Y evolui no decorrer da árvore filogenética. Dessa maneira, a matriz de covariância de Y é representada por:

$$\mathbf{V}(\alpha) = \mathbf{A}^{\frac{1}{2}}\mathbf{C}(\alpha)\mathbf{A}^{\frac{1}{2}}, \tag{2.7}$$

em que, a matriz $\mathbf{A}$ possui diagonal contendo $\mu(1-\mu)$ e

$$\mathbf{C}(\alpha) = exp(-2\alpha(1-\mathbf{W})), \tag{2.8}$$

e $\mathbf{W}$ simboliza a árvore filogenética. Os elementos da diagonal principal dessa matriz - w_{ii}, representam o comprimento da base da árvore até a ponta da espécie i. E os demais elementos w_{ik} correspondem ao comprimento do ramo que as espécies i e k compartilham (IVES; JR, 2010).

Quando não temos a presença de variáveis preditoras, ou seja, nosso modelo terá apenas o b_0 (intercepto), a estimação do parâmetro α representará o sinal filogenético. Segundo Garamszegi (2014), quanto mais alto os valores de α mais forte é o sinal filogenético, em que em uma situação exagerada não haveria variação,pois os elementos da topologia teriam os mesmos valores em uma determinada característica.

2.6 Correlação

Como a correlação expressa o grau de dependência entre variáveis, isso faz com que ela seja, de certa forma, imediata ao pensar no estudo da integração e modularidade. Existem diversas medidas de dependência e as mais comuns são a correlação de Pearson e a covariância.

A covariância é expressa como:

$$s_{j,k} = cov(X_j, X_k) = \frac{\sum_{i=1}^{n}(x_{i,j} - \bar{x_j})(x_{i,k} - \bar{x_k})}{n-1}, \tag{2.9}$$

em que, $\bar{x_j}$ e $\bar{x_k}$ corresponde a média das variáveis X_j e X_k, respectivamente.

A correlação de Pearson é então denotada como uma padronização da covariância, como mostrado na Equação 2.10. Assim, o resultado da correlação é um valor entre 1 e -1, indicando respectivamente, uma perfeita dependência linear positiva e uma perfeita dependência linear negativa.

$$r_{j,k} = corr(X_j, X_k) = \frac{cov(X_j, X_k)}{s_{X_j} s_{X_k}} = \frac{\sum_{i=1}^{n}(x_{i,j} - \bar{x_j})(x_{i,k} - \bar{x_k})}{s_{X_j} s_{X_k}}, \tag{2.10}$$

em que s_{X_j} e s_{X_k} correspondem ao desvio padrão das variáveis X_j e X_k, respectivamente.

Quando não há um relacionamento linear entre as variáveis o coeficiente amplamente utilizado é o de Spearman. Essa medida também varia no intervalo de -1 a 1 como o de Pearson, e sua interpretação é análoga.

Spearman utiliza-se de postos para realizar o cálculo da correlação. Para se obter os postos, é necessário que a ordenação dos dados seja realizada de forma crescente, e então teremos que o menor valor assumido de uma variável receberá posto 1, o segundo menor recebe posto 2 e assim por diante.

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{2.11}$$

em que, d_i corresponde a diferença entre os postos das variáveis analisadas e n é o número de indivíduos.

A medida que o número de variáveis estudadas aumenta, torna-se necessário utilizar matrizes de correlação e covariância para auxiliar na visualização de todas as associações, e também na manipulação dos dados.

A matriz de covariância é composta pelas variâncias na diagonal principal, e nos elementos fora da diagonal temos as covariâncias, como mostrado abaixo:

$$\begin{bmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \dots & \dots & \dots & \dots \\ s_{p1} & \dots & \dots & s_{pp} \end{bmatrix}$$

Já a matriz de correlação, constituí-se de 1 na diagonal principal e as correlações entre as variáveis nas demais células.

$$\begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & \dots & \dots & 1 \end{bmatrix}$$

2.7 *Cluster* Hierárquico

A análise de agrupamento ou também chamada de *Cluster* é muito utilizada na identificação de padrões, e consiste em formar grupos de objetos (indivíduos ou covariáveis) semelhantes, de modo que espera-se encontrar homogeneidade nos objetos dentro de cada grupo e diferenças entre os mesmos (heterogeneidade entre os grupos).

Medidas de similaridade ou dissimilaridade são usadas na construção dos grupos, dentre as métricas de distância temos a Euclidiana (mais conhecida e denotada na Equação 2.12), *Mahalanobis* e *Minkowski*, entre outras. Mas também é possível considerar medidas de associação, como as de correlação, já citadas na Seção 2.6.

$$d_{i,j}^2 = \sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2. \tag{2.12}$$

Os métodos de *Cluster* podem ser divididos em duas grandes frentes, que são os métodos não hierárquicos e os hierárquicos. O primeiro tipo de abordagem precisa da especificação do número de grupos previamente, enquanto o segundo não precisa. O foco deste trabalho está no *Cluster* hierárquico, portanto para mais detalhes sobre o outro tipo de técnica consultar Johnson, Wichern *et al.* (2007) e Izbicki e Santos (2020).

Os agrupamentos hierárquicos podem ser classificados em aglomerativos e divisivos. Nos métodos aglomerativos cada objeto é considerado um *cluster* e de acordo com a similaridade entre os mesmos é realizado o agrupamento, isso ocorre de forma sucessiva até que todos sejam parte de um mesmo grupo. Os métodos divisivos funcionam de forma inversa, partindo de um único grupo e realizando diversas divisões.

Para cada tipo de agrupamento hierárquico, podemos ainda escolher entre os métodos de ligação simples, completa, média e *ward* para construção do procedimento. O método de *ward* realiza a minimização da variância dentro dos grupos (*ESS - error sum of square*) (JOHNSON; WICHERN *et al.*, 2007). Dessa forma, os grupos que apresentarem menores valores de *ESS* são agrupados, e em cada passo isso é novamente analisado.

Para se visualizar os grupos formados pela análise é utilizado o dendrograma, que é uma representação gráfica que exibe como os objetos foram agrupados em cada passo do processo.

2.8 Métricas

As métricas empregadas para as análises incluíram: o *Rand Index* (RI), *Normalized Mutual Information (NMI)* e o *Fowlkes Mallows Index (FMI)*.

- *Rand Index*: Essa medida foi descrita por Rand (1971) e tem como objetivo comparar os objetos agrupados na análise de *Cluster* com seus respectivos grupos verdadeiros.

  Para o cálculo do *RI* é necessário a construção da matriz de confusão, como ilustramos na Tabela 2.

  Tabela 2 – Matriz de confusão.

		Módulo real	
		1	**2**
Módulo cluster	**1**	a	b
	2	c	d

  E então a Equação 2.13 é utilizada para o cálculo da métrica.

$$RI = \frac{a+d}{a+b+c+d}, \tag{2.13}$$

  em que, podemos observar que é realizada a soma da diagonal principal dividida pelo total, como uma proporção de acertos. Os valores obtidos podem variar de 0 a 1, de forma que 1 representa uma perfeita correspondência (HUBERT; ARABIE, 1985).

  Como a marcação dos grupos encontrados no *Cluster* recebem valores arbitrários, ou seja, o que é chamado de grupo 1 poderia receber o valor 0,2,3,... e assim por diante; calculamos esse índice duas vezes, alterando a marcação do que seria módulo 1 e 2. E então para ter o resultado final consideramos o maior valor obtido da medida.

- *Normalized Mutual Information*: A segunda métrica utilizada na avaliação é denominada de Informação Mútua Normalizada - NMI (do inglês *Normalized Mutual Information*).

  De acordo com Cover (1999), a informação mútua entre duas variáveis aleatórias X e Y, com distribuição de probabilidade conjunta $p(x,y)$ e distribuição marginal $p(x)$ e $p(y)$, é descrita como:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}. \tag{2.14}$$

  Uma das maneiras de normalizar essa medida é através da divisão pela média das entropias de X e Y. A entropia quantifica a incerteza e é definida como:

$$H(X) = -\sum_{x \in X} p(x,y) log p(x). \tag{2.15}$$

Assim, NMI é:

$$NMI = \frac{I(X,Y)}{\frac{H(X)+H(Y)}{2}} = \frac{2I(X,Y)}{H(X)+H(Y)}. \tag{2.16}$$

- *Fowlkes Mallows Index*: Por fim, *Fowlkes Mallows* é definido como:

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}}, \tag{2.17}$$

em que, TP: verdadeiro positivo (quando é previsto corretamente a classe), FP: falso positivo (quando a classe de interesse é prevista de forma incorreta) e FN: falso negativo (quando é previsto de maneira incorreta a classe que não temos interesse).

Todas as medidas tem como resultado um valor entre 0 e 1, sendo que 1 indica uma perfeita separação, em outras palavras, significa que o método aplicado conseguiu atribuir os módulos de forma igual.

CAPÍTULO

3

METODOLOGIA

Este estudo tem como objetivo identificar caracteres discretos que covariam. A Figura 6 ilustra o esquema da aplicação da metodologia proposta, que foi dividida em duas etapas. Inicialmente, calcula-se a dissimilaridade entre os caracteres utilizando diferentes técnicas, com a incorporação da topologia e também sem o seu uso. Em seguida, essas estimativas servem como base para a análise de agrupamento, o que auxilia na visualização e na interpretação dos padrões de integração e modularidade entre os caracteres.

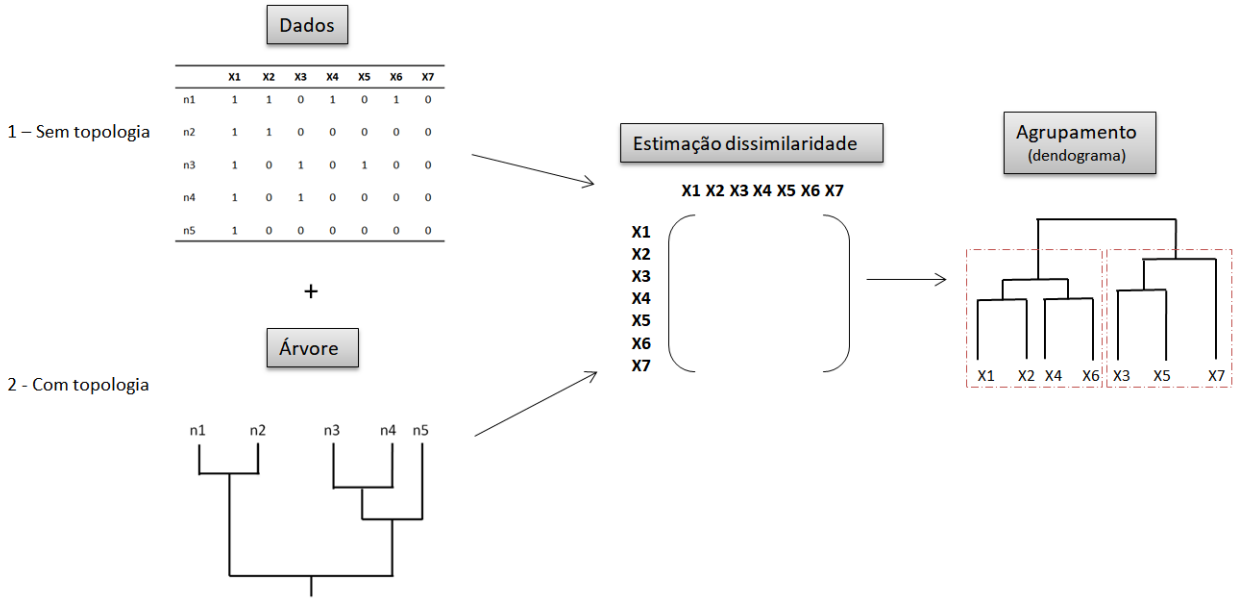


Figura 6 – Esquema do funcionamento do método proposto.

3.1 Estimação da dissimilaridade

A primeira parte, correspondente à estimação da dissimilaridade, pode ainda ser dividida em duas abordagens: correlações simples e aquelas que utilizam a topologia. O *Threshold Model* e a *Phylogenetic Logistic Regression* são métodos que incorporam a informação da árvore filogenética em seus cálculos. Por outro lado, as correlações simples são obtidas através dos coeficientes de Pearson e Spearman, que não consideram a topologia.

Assim, após calcular o *Threshold Model*, Pearson e Spearman, como comentado na Seção 2.4 e Seção 2.6, é necessário obter a matriz de dissimilaridade $\mathbf{M}$:

- *Threshold Model*, Pearson e Spearman:

$$\mathbf{M}_1 = \mathbf{1} - |\mathbf{R}|, \tag{3.1}$$

  em que, $\mathbf{R}$ corresponde à matriz resultante do cálculo do *Threshold Model*, Pearson ou Spearman. E $\mathbf{1}$ é uma matriz de valores 1.

No caso do *Threshold Model*, também consideramos a transformação da matriz de correlação resultante em uma matriz positiva definida, a fim de garantir a estrutura de uma matriz de correlação. Acreditamos que essa alteração poderia resultar em melhores resultados.

Para a *Phylogenetic Logistic Regression*, ajustamos o modelo de regressão considerando os caracteres aos pares, em que um é a variável resposta e o outro é a explicativa. A partir disso, utilizamos os valores dos coeficientes e do valor-p como medidas de associação nessa técnica.

Como os coeficientes podem ser valores maiores do que 1, normalizamos a matriz para assegurar que os valores estivessem na mesma escala que às demais. Dessa forma, calculamos a matriz de dissimilaridade:

- Coeficientes:

$$\mathbf{M}_2 = 3 - |\mathbf{R}|, \tag{3.2}$$

  em que, foi utilizado o valor 3 como constante para que o dendograma permanecesse em uma escala semelhante às demais.

- Valor-p:

$$\mathbf{M}_3 = |\mathbf{R}|, \tag{3.3}$$

  em que, como a matriz de valor-p já é dissimilar, é a única que difere no cálculo da matriz de dissimilaridade.

Com as matrizes de dissimilaridades calculadas, podemos realizar o agrupamento das características.

3.2 Agrupamento

A partir da matriz de dissimilaridade, realizamos o agrupamento dos caracteres através do *Cluster* hierárquico. Adicionalmente, consideramos a indicação do módulo real a que cada traço pertence, permitindo avaliar a performance dos grupos identificados em relação ao comportamento real.

CAPÍTULO

4

EXPERIMENTOS

Há uma dificuldade na obtenção de bases de dados reais para estudos filogenéticos. Em parte pela falta de disponibilidade, e também pela coleta dos dados não ser algo simples. Dessa maneira, para conseguir analisar as técnicas aqui discutidas e proposta, foram utilizadas simulações.

Temos o interesse em analisar o comportamento de diferentes estruturas para o cálculo da similaridade na identificação dos módulos de caracteres, ou seja, conseguir investigar como as diferentes maneiras de se considerar a dependência nas características são influenciadas ou não pela informação da árvore filogenética, e como é capaz de detectar a modularidade existente.

A fim de se conseguir atingir esses objetivos, foram realizados experimentos com base em dois cenários de simulação dos dados. Para o primeiro cenário utilizamos o Modelo de Markov (HARMON, 2019), e para o segundo o *Threshold Model*. Em ambos foi calculado as correlações de Pearson, de Spearman, o *Threshold Model*, Valor-p e os Coeficientes como alternativas na comparação da construção da matriz de similaridade. A matriz com valor-p e a com os coeficientes, foram obtidas através do *Phylogenetic Logistic Regression*. Também utilizamos a transformação da matriz obtida no *Threshold* em positiva definida, como forma de identificar se os resultados alcançados seriam melhores.

Utilizamos a função *rtree* (PARADIS, 2012) do pacote *ape* do *R* para construir as árvores. Essa função gera árvores aleatoriamente a partir do número de nós terminais (espécies) definidos, sendo aqui utilizado 100; por padrão produz comprimentos de ramos utilizando valores uniformes com o *runif*.

4.1 Cenário 1: Modelo de Markov

A simulação dos dados foi primeiramente realizada através do Modelo Oculto de Markov, também chamado de HMM - do inglês *Hidden Markov Model*.

Foram considerados 40 caracteres (características/variáveis), divididos em dois módulos de 20. Dentro do módulo foi fixada uma matriz de transição. Com isso, o estado oculto conseguirá controlar a evolução dos caracteres, e assumirá os seguintes estados:

- inativado: sem modularidade;
- os dois módulos forçam o surgimento dos caracteres;
- o módulo um pressiona o aparecimento e o dois o desaparecimento;
- o módulo um pressiona o desaparecimento e o dois o aparecimento;
- os dois módulos forçam a ausência.

Para a geração dos dados foi utilizado o pacote *pythools*. A função *sim.history* foi empregada para simular o histórico do caracter e a função *sim.multiMk* para se obter os estados de uma característica discreta nos nós terminais da árvore, como definido por Revell (2012). Para essa última etapa são consideradas as seguintes matrizes de transição, correspondendo respectivamente ao aparecimento, desaparecimento e inativação:

	0	**1**
0	0	k
1	0	0

	0	**1**
0	0	0
1	k	0

	0	**1**
0	0	k
1	k	0

A constante k representa a taxa evolutiva, em que foi escolhido o valor igual a 5 para considerar como uma taxa elevada, e 2 para taxa lenta.

4.2 Cenário 2: *Threshold Model*

A simulação dos dados a partir do *Threshold Model* seguiu a configuração básica análoga ao HMM, ou seja, foram considerados dois módulos de 20 características (variáveis) cada e 100 nós terminais (espécies).

Inicialmente simulamos a árvore filogenética, em seguida a matriz dos dados através das *sim.corrs*. O *sim.corrs* faz parte do pacote *pythools* e simula o *Motion Brownian* a partir da árvore e da matriz de correlações reais entre as características.

Foram consideradas diferentes configurações de simulação, ou seja, foram utilizadas diferentes valores de correlação e também duas estruturas na construção da matriz de correlação.

As estruturas de construção da matriz de correlação utilizadas são chamadas de simetria composta e autoregressiva. A primeira é a mais simples, em que temos que independente do tempo a correlação é a mesma, como podemos observar:

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

Já a estrutura autoregressiva (lag 1), possui o valor da correlação diminuindo ao se aumentar a distância entre os pontos, como podemos observar abaixo:

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}$$

Com isso, conseguimos construir a matriz de correlação original, de modo que a correlação entre os módulos é zero, e dentro de cada um deles assume os seguintes valores: 0.2, 0.5 ou 0.8. Abaixo temos uma pequena representação do formato da matriz original para 4 características, mas esse formato pode ser estendido para qualquer valor desejado.

$$\begin{bmatrix} 1 & r & 1 & 0 \\ r & 1 & 0 & 1 \\ 1 & 0 & 1 & r \\ 0 & 1 & r & 1 \end{bmatrix}$$

A matriz de dados é então obtida a partir da informação sobre a árvore (que necessita apenas da indicação do número de observações) e a matriz de correlação original. Em seguida, transformamos todos os caracteres no formato binário, pois o interesse está em utilizar caracteres discretos.

Dessa forma, conseguimos ajustar cada uma das técnicas mencionadas para a comparação na obtenção da matriz de similaridade.

Consideramos uma pequena distinção dentro desse cenário relacionada à construção das *liabilities*. Nessa segunda alternativa, subtraimos os valores das *liabilities* pela média das colunas da matriz da mesma.

4.3 Estimação

A partir dos dados gerados através das simulações descritas anteriormente, fizemos a estimação das correlações. Para Pearson e Spearman foi utilizado a função *cor*.

Já para o *Threshold* a função *treshBayes* do pacote *pythools* foi empregada. Essa função foi construída baseada no método descrito por Felsenstein (2012). Os seguintes parâmetros são definidos: 1000 como o intervalo de amostragem, 50000 gerações de cadeias MCMC e 10000 cadeias descartadas (*burn in*), também é utilizada a árvore filogenética e a indicação de que as variáveis são discretas.

Assim, foi calculada a média das correlações resultantes. É interessante destacar que essa função só permite a estimação de dois caracteres, portanto o código foi construído de modo a fazer repetições e calcular a correlação entre os pares de características.

Por fim, escolhemos utilizar o *Phylogenetic Logistic Regression* como um outro método de estimação das correlações que faz uso da topologia. Aplicamos a função *phyloglm* do pacote *phylolm* (HO; ANÉ, 2014a; HO; ANÉ, 2014b) do R, e consideramos duas medidas para construção da matriz estimada, são elas: o valor-p e o valor absoluto do coeficiente. O modelo de regressão foi ajustado de modo a considerar os pares de características, para que uma fosse a variável resposta e a outra a explicativa.

CAPÍTULO

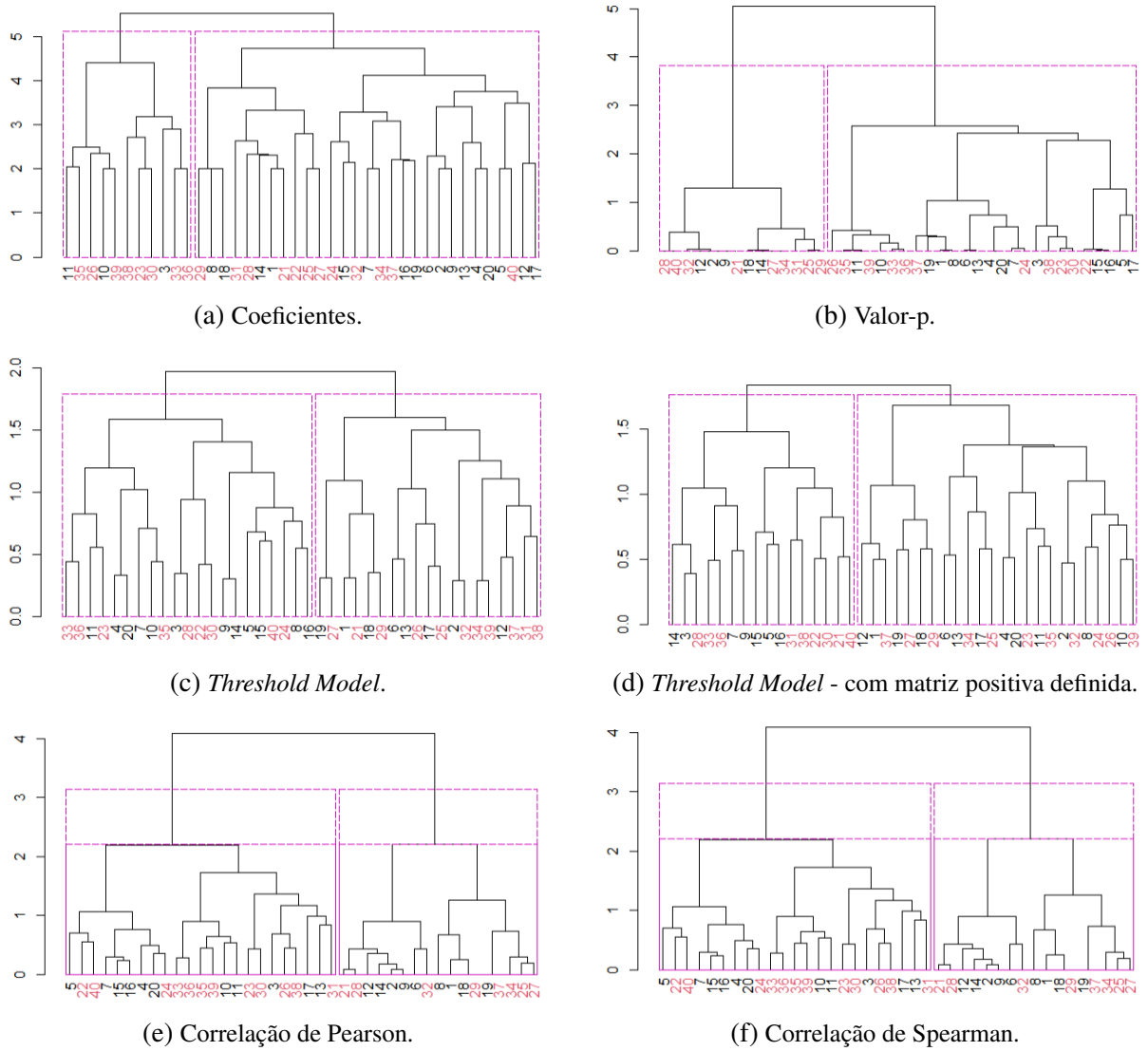
5

RESULTADOS

Utilizando os conjuntos de dados simulados detalhados no Capítulo 4, desenvolvemos dendrogramas para cada um dos métodos empregados. As análises foram baseadas nas matrizes de correlações estimadas, juntamente com a identificação prévia das características nos módulos específicos (os valores reais dos grupos). Posteriormente, procedemos à análise visual dos dendrogramas e realizamos o cálculo de diversas métricas, que permitiram uma avaliação dos resultados obtidos.

5.1 Comparação

Inicialmente, podemos observar os dendrogramas obtidos em cada um dos métodos estudados. A Figura 7 exemplifica um dos cenários, que corresponde aos dendrogramas obtidos para os dados gerados do modelo de Markov com taxa evolutiva igual a 5. As demais figuras podem ser encontradas no Apêndice A.



(a) Coeficientes.

(b) Valor-p.

(c) *Threshold Model*.

(d) *Threshold Model* - com matriz positiva definida.

(e) Correlação de Pearson.

(f) Correlação de Spearman.

Figura 7 – Dendrogramas utilizando os dados simulados do modelo de Markov com taxas evolutivas = 5.

Os dendrogramas possuem a marcação em linha pontilhada representando o agrupamento realizado pelo *Cluster* hierárquico do tipo *hclust*, e as cores dos rótulos (rosa e preto) indicam o verdadeiro módulo ao qual a variável pertence, ou seja, como foi marcado na construção da simulação dos dados (do caracter 1 ao 20 foi considerado um módulo e o restante outro módulo).

Dessa maneira, podemos notar que na Figura 7a o primeiro agrupamento (considerando os rótulos da esquerda para a direita) possui em sua maioria caracteres pertencente a um mesmo módulo, pois possui mais rótulos com a cor rosa.

A Tabela 3 apresenta os resultados obtidos em cada um dos métodos estudados em todas as métricas descritas na Seção 2.8. Em negrito são destacados os que tiveram melhores resultados segundo cada métrica, e o último cenário *Threshold Model** analisado tem as *liabilites* centradas na média como explicado na Seção 4.2.

Tabela 3 – Avaliação dos agrupamentos em cada configuração de cenário analisada.

Simulação			Método de Estimação	RI	NMI	FMI
Markov	Taxa evolutiva	5	*Threshold Model*	0.575	**0.748**	**0.949**
			Threshold Model - matriz postiva definida	0.550	0.448	0.815
			Correlação de Pearson	0.525	0.520	0.856
			Correlação de Spearman	0.525	0.520	0.856
			PLR -Pvalor	0.600	0.342	0.744
			PLR - Coeficientes	**0.625**	0.235	0.671
	Taxa evolutiva	2	*Threshold Model*	0.600	0.448	0.815
			Threshold Model - matriz positiva definida	0.625	**0.748**	**0.949**
			Correlação de Pearson	**0.675**	**0.748**	**0.949**
			Correlação de Spearman	**0.675**	**0.748**	**0.949**
			PLR - Pvalor	0.650	0.266	0.691
			PLR - Coeficientes	**0.675**	0.390	0.777
Threshold Model	Simetria Composta	0.2	*Threshold Model*	0.575	0.182	0.644
			Threshold Model - matriz positiva definida	0.650	**0.448**	**0.815**
			Correlação de Pearson	0.650	0.207	0.656
			Correlação de Spearman	0.650	0.207	0.656
			PLR - Pvalor	0.500	0.342	0.744
			PLR - Coeficientes	**0.700**	**0.448**	**0.815**
		0.5	*Threshold Model*	0.975	0.748	0.949
			Threshold Model - matriz positiva definida	0.975	0.748	0.949
			Correlação de Pearson	0.975	0.748	0.949
			Correlação de Spearman	0.975	0.748	0.949
			PLR - Pvalor	**1.000**	**1.000**	**1.000**
			PLR - Coeficientes	0.975	0.748	0.949
		0.8	*Threshold Model*	**1.000**	**1.000**	**1.000**
			Threshold Model - matriz positiva definida	**1.000**	**1.000**	**1.000**
			Correlação de Pearson	**1.000**	**1.000**	**1.000**
			Correlação de Spearman	**1.000**	**1.000**	**1.000**
			PLR - Pvalor	**1.000**	**1.000**	**1.000**
			PLR - Coeficientes	**1.000**	**1.000**	**1.000**
	Autoregressivo	0.2	*Threshold Model*	0.625	0.235	0.671
			Threshold Model - matriz positiva definida	0.575	**0.748**	**0.949**
			Correlação de Pearson	0.550	0.448	0.815
			Correlação de Spearman	0.550	0.448	0.815
			PLR - Pvalor	**0.750**	0.342	0.744
			PLR - Coeficientes	0.575	0.138	0.633
		0.5	*Threshold Model*	0.525	**0.302**	**0.716**
			Threshold Model - matriz positiva definida	0.550	0.207	0.656
			Correlação de Pearson	0.550	0.266	0.691
			Correlação de Spearman	0.550	0.266	0.691
			PLR - Pvalor	**0.775**	0.235	0.671
			PLR - Coeficientes	0.600	0.159	0.636
		0.8	*Threshold Model*	0.800	0.266	0.691
			Threshold Model - matriz positiva definida	0.975	0.748	0.949
			Correlação de Pearson	0.850	0.342	0.744
			Correlação de Spearman	0.850	0.342	0.744
			PLR - Pvalor	0.850	0.342	0.744
			PLR - Coeficientes	**1.000**	**1.000**	**1.000**
*Threshold Model**	Simetria Composta	0.2	*Threshold Model*	0.775	0.748	0.949
			Threshold Model - matriz postiva definida	**0.925**	0.520	0.856
			Correlação de Pearson	0.525	0.390	0.777
			Correlação de Spearman	0.525	0.390	0.777
			PLR -Pvalor	0.825	0.390	0.777
			PLR - Coeficientes	0.900	**1.000**	**1.000**
		0.5	*Threshold Model*	0.975	0748	0.949
			Threshold Model - matriz positiva definida	**1.000**	**1.000**	**1.000**
			Correlação de Pearson	**1.000**	**1.000**	**1.000**
			Correlação de Spearman	**1.000**	**1.000**	**1.000**
			PLR - Pvalor	**1.000**	**1.000**	**1.000**
			PLR - Coeficientes	0.900	**1.000**	**1.000**
		0.8	*Threshold Model*	0.975	0.748	0.949
			Threshold Model - matriz positiva definida	**1.000**	**1.000**	**1.000**
			Correlação de Pearson	**1.000**	**1.000**	**1.000**
			Correlação de Spearman	**1.000**	**1.000**	**1.000**
			PLR - Pvalor	**1.000**	**1.000**	**1.000**
			PLR - Coeficientes	**1.000**	**1.000**	**1.000**

Através da Tabela 3 podemos observar diferenças interessantes dentro das estruturas de cada tipo de cenário e estrutura utilizada. De modo geral percebemos que ao aumentar o valor da correlação utilizada há um aumento nos valores obtidos nas métricas para a estrutura de simetria composta.

Na simulação de Markov com taxa 5 o *Threshold Model* apresentou melhores resultados em duas das três métricas utlizadas. Por outro lado, com uma taxa evolutiva de menor valor, ocorreu empates entre as correlações simples de Pearson e Spearman com o *Threshold Model* utilizando matriz positiva definida, no NMI e FMI; enquanto no RI o empate foi com o PLR - Coeficientes.

A estrutura autoregressiva obteve o PLR - Pvalor com melhor resultado para o RI, com excessão do valor inicial de 0.8.

Para facilitar a interpretação dos resultados construímos uma tabela resumo, que mostra a quantidade de vezes que cada método de estimação obteve melhor resultado, segundo cada uma das métricas utilizadas. Como observado na Tabela 4, PLR - Coeficientes teve melhores resultados em todas as medidas, seguido do *Threshold Model* - matriz positiva definida e PLR - valor-p. Todos esses métodos tem em comum a utilização da informação filogenética na sua construção. Pearson e Spearman que são os métodos sem a utilização da topologia tiveram um empate, apresentando-se de forma inferior na identificação dos módulos.

Tabela 4 – Quantidade de vezes que cada método de estimação apresentou melhor resultado segundo cada métrica.

	RI	**NMI**	**FMI**
Threshold Model	1	3	3
Threshold Model - matriz positiva definida	4	6	6
Pearson	4	4	4
Spearman	4	4	4
PLR - Valor-p	6	4	4
PLR - Coeficientes	6	6	6

CAPÍTULO

6

CONCLUSÃO

Neste trabalho investigamos a identificação de módulos, ou seja, características que variam conjuntamente com base na estimação das dissimilaridades. Para isso, propomos uma metodologia composta por dois principais passos, o cálculo das dissimilaridades entre os caracteres e a construção do *Cluster* hierárquico como forma de se obter a visualização dos módulos.

Além disso, testamos a metodologia com dados simulados, em que utilizamos simulação de Markov e do *Threshold*. Dentro de cada tipo de simulação foram consideradas diferentes estruturas, como a configuração da matriz (simetria composta e autoregressiva), diferentes valores para a taxa evolutiva (utilizamos 2 e 5) e para as correlações iniciais originais.

Utilizamos correlações simples - Pearson e Spearman, e com a topologia - *Threshold Model* e *Phylogenetic Logistic Regression*, como formas de se calcular a dissimilaridade. Para a PLR consideramos os valores absolutos dos coeficientes e o valor-p como medida de associação. A partir das configurações utilizadas conseguimos perceber diferenças dentro das mesmas. Na estrutura de simetria composta, independente do método de estimação utilizado, foi observado que o aumento dos valores da correlação inicial também levava a um aumento nos valores das métricas, obtendo um melhor desempenho com o valor 0.8.

Ao avaliarmos as técnicas com diferentes métricas (RI, NMI e FMI), foi possível notar que os Coeficientes obtiveram melhores resultados em todas as medidas, seguido do *Threshold* com matriz positiva definida e do Valor-p. Isso significa que essas técnicas conseguiram atribuir os módulos de forma muito próxima dos grupos reais.

O ponto em comum das técnicas que foram destaque, trata-se da utilização da árvore filogenética em sua construção. Isso reforça a importância e o quanto há de acréscimo de informação no uso de topologias, comparada quando não à utilizamos.

6.1 Trabalhos Futuros

Neste trabalho, consideramos estudos de simulação sob determinadas condições. Em futuras discussões, seria interessante incorporar a aplicação em bases de dados reais.

Além disso, este estudo focou em variáveis binárias, e seria relevante explorar características com mais níveis, examinando seu efeito na integração.

Outro aspecto a ser abordado em estudos posteriores é a análise de intervalos de confiança para as métricas utilizadas.

REFERÊNCIAS

CAVALLI-SFORZA, L. L.; EDWARDS, A. W. Phylogenetic analysis. models and estimation procedures. **American journal of human genetics**, Elsevier, v. 19, n. 3 Pt 1, p. 233, 1967. Citado na página 23.

CORNWELL, W.; NAKAGAWA, S. Phylogenetic comparative methods. **Current Biology**, Elsevier, v. 27, n. 9, p. R333–R336, 2017. Citado na página 19.

COVER, T. M. **Elements of information theory**. [S.l.]: John Wiley & Sons, 1999. Citado na página 33.

DARWIN, C. **The Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life**. [S.l.]: Books, Incorporated, Pub., 1859. Citado na página 19.

FELSENSTEIN, J. Maximum-likelihood estimation of evolutionary trees from continuous characters. **American journal of human genetics**, Elsevier, v. 25, n. 5, p. 471, 1973. Citado na página 24.

____. Phylogenies and the comparative method. **The American Naturalist**, University of Chicago Press, v. 125, n. 1, p. 1–15, 1985. Citado na página 24.

____. Quantitative characters, phylogenies, and morphometrics. **Systematics Association Special Volume**, London; Chapman & Hall; 1998, v. 64, p. 27–44, 2002. Citado na página 26.

____. Using the quantitative genetic threshold model for inferences between and within species. **Philosophical Transactions of the Royal Society B: Biological Sciences**, The Royal Society London, v. 360, n. 1459, p. 1427–1434, 2005. Citado na página 27.

____. A comparative method for both discrete and continuous characters using the threshold model. **The American Naturalist**, University of Chicago Press Chicago, IL, v. 179, n. 2, p. 145–156, 2012. Citado nas páginas 26 e 42.

GARAMSZEGI, L. Z. **Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice**. [S.l.]: Springer, 2014. Citado na página 30.

GOSWAMI, A.; POLLY, P. D. Methods for studying morphological integration and modularity. **The Paleontological Society Papers**, Cambridge University Press, v. 16, n. 5, p. 213–243, 2010. Citado na página 20.

HARMON, L. J. **Phylogenetic comparative methods**. [S.l.]: Independent, 2019. Citado nas páginas 20, 21, 26, 27, 28 e 39.

HO, L. S. T.; ANÉ, C. Intrinsic inference difficulties for trait evolution with ornstein-uhlenbeck models. **Methods in Ecology and Evolution**, Wiley Online Library, v. 5, n. 11, p. 1133–1146, 2014. Citado na página 42.

HO, L. s. T.; ANÉ, C. A linear-time algorithm for gaussian and non-gaussian trait evolution models. **Systematic biology**, Oxford University Press, v. 63, n. 3, p. 397–408, 2014. Citado na página 42.

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of classification**, Springer, v. 2, p. 193–218, 1985. Citado na página 33.

IVES, A. R.; JR, T. G. Phylogenetic logistic regression for binary dependent variables. **Systematic biology**, Oxford University Press, v. 59, n. 1, p. 9–26, 2010. Citado na página 30.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina: uma abordagem estatística**. [S.l.]: Rafael Izbicki, 2020. Citado na página 32.

JOHNSON, R. A.; WICHERN, D. W. *et al.* **Applied multivariate statistical analysis**. [S.l.]: Pearson, 2007. Citado na página 32.

MATIOLI, S. R.; FERNANDES, F. M. d. C. **Biologia Molecular e Evolução**. 2. ed. [S.l.]: Holos Editora, 2012. Citado nas páginas 19, 23 e 24.

MITOV, V.; BARTOSZEK, K.; ASIMOMITIS, G.; STADLER, T. Fast likelihood calculation for multivariate gaussian phylogenetic models with shifts. **Theoretical population biology**, Elsevier, v. 131, p. 66–78, 2020. Citado na página 20.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. **Journal of the Royal Statistical Society: Series A (General)**, Wiley Online Library, v. 135, n. 3, 1972. Citado na página 29.

OLSON, E. C.; MILLER, R. L. **Morphological integration**. [S.l.]: University of Chicago Press, 1999. Citado na página 19.

OLSSON, U. **Generalized Linear Models: An Applied Approach**. [S.l.]: Studentlitteratur AB, 2002. Citado na página 29.

O'MEARA, B. C.; ANÉ, C.; SANDERSON, M. J.; WAINWRIGHT, P. C. Testing for different rates of continuous trait evolution using likelihood. **Evolution**, Wiley Online Library, v. 60, n. 5, p. 922–933, 2006. Citado na página 20.

PARADIS, E. **Analysis of Phylogenetics and Evolution with R**. [S.l.]: Springer, 2012. v. 2. Citado na página 39.

____. An introduction to the phylogenetic comparative method. **Modern phylogenetic comparative methods and their application in evolutionary biology: Concepts and practice**, Springer, p. 3–18, 2014. Citado na página 23.

PINSKY, M.; KARLIN, S. **An introduction to stochastic modeling**. [S.l.]: Academic press, 2010. Citado na página 26.

RAND, W. M. Objective criteria for the evaluation of clustering methods. **Journal of the American Statistical association**, Taylor & Francis, v. 66, n. 336, p. 846–850, 1971. Citado na página 33.

REVELL, L. J. phytools: an r package for phylogenetic comparative biology (and other things). **Methods in ecology and evolution**, Blackwell Publishing Ltd, n. 2, p. 217–223, 2012. Citado na página 40.

____. Ancestral character estimation under the threshold model from quantitative genetics. **Evolution**, Wiley Online Library, v. 68, n. 3, p. 743–759, 2014. Citado na página 26.

____. phytools 2.0: an updated r ecosystem for phylogenetic comparative methods (and other things). **PeerJ**, PeerJ Inc., v. 12, p. e16505, 2024. Citado na página 29.

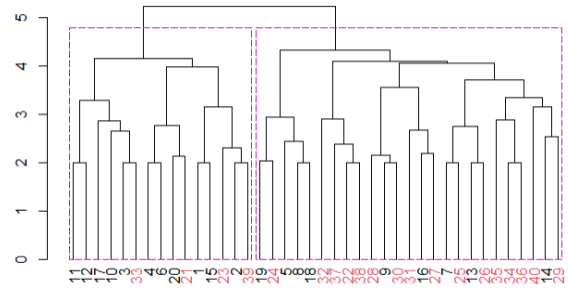
APÊNDICE

A

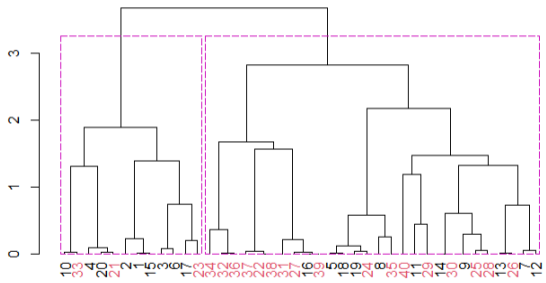
DENDROGRAMAS

A.1 Cenário 1: Modelo de Markov

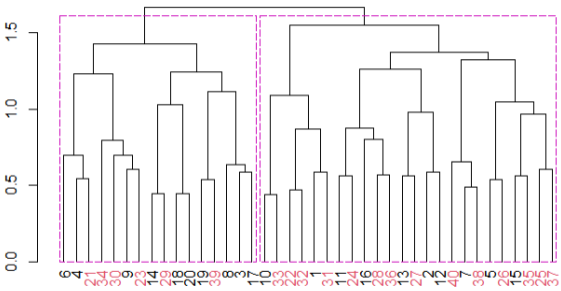
A Figura 8 corresponde aos dendrogramas obtidos para os dados gerados do modelo de Markov com taxa evolutiva utilizando o valor 2.



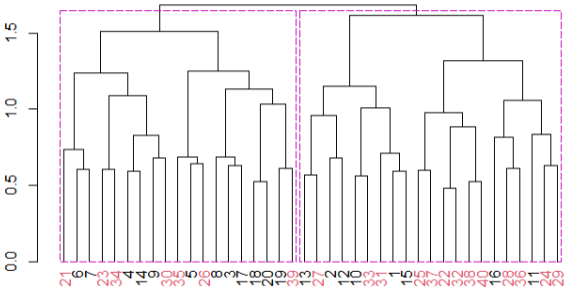
(a) Coeficientes.



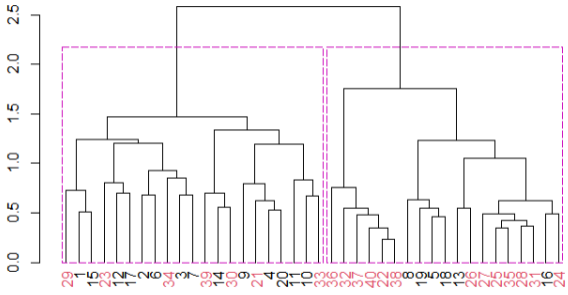
(b) Valor-p.



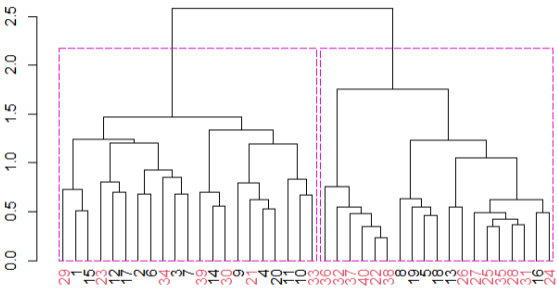
(c) *Threshold Model*.



(d) *Threshold Model* - com matriz positiva definida.



(e) Correlação de Pearson.



(f) Correlação de Spearman.

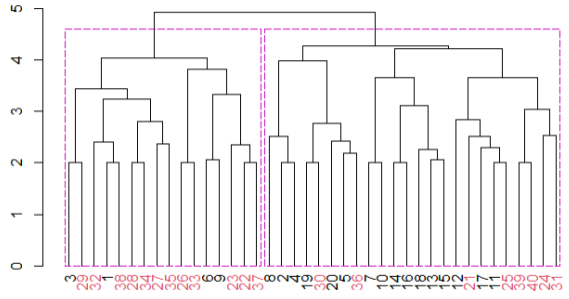
Figura 8 – Dendrogramas utilizando os dados simulados do modelo de Markov com taxas evolutivas = 2.

A.2 Cenário 2: *Threshold Model*

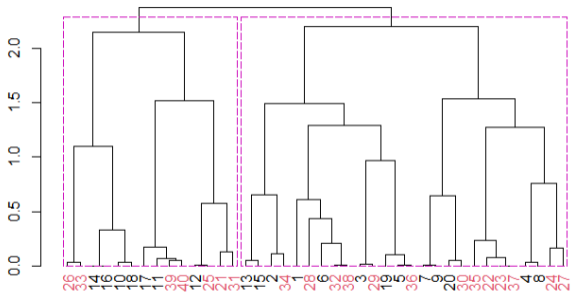
A.2.1 Estrutura de simetria composta

A.2.1.1 Correlação = 0.2

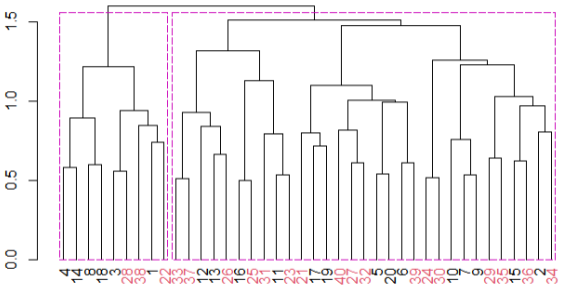
A Figura 9 corresponde aos dendrogramas obtidos para os dados gerados do modelo *Threshold* utilizando a estrutura de simetria composta com correlação igual a 0.2.



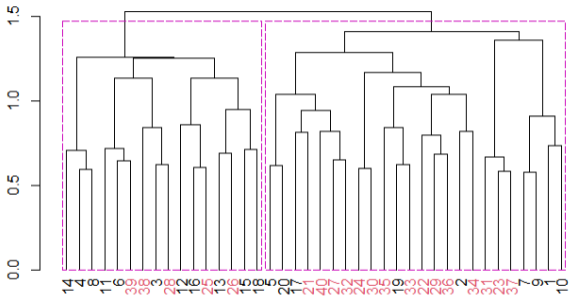
(a) Coeficientes.



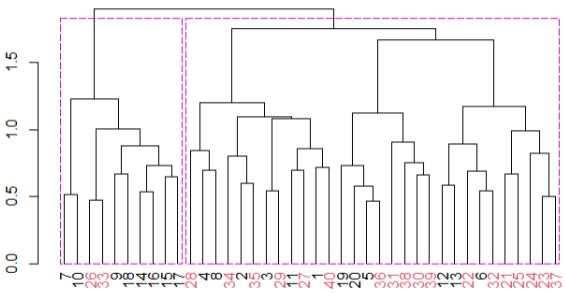
(b) Valor-p.



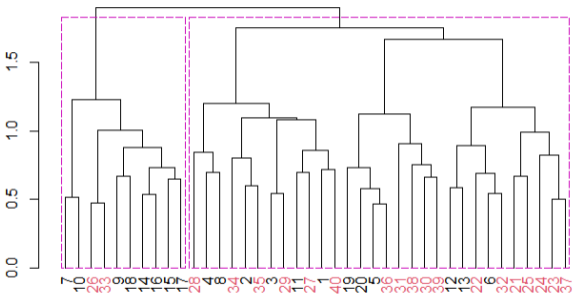
(c) *Threshold Model*.



(d) *Threshold Model* - com matriz positiva definida.



(e) Correlação de Pearson.

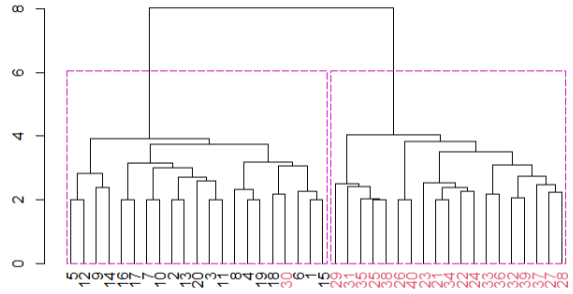


(f) Correlação de Spearman.

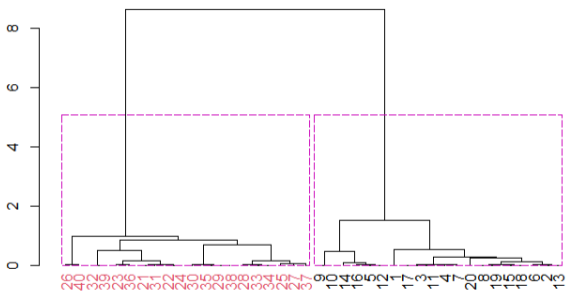
Figura 9 – Dendrogramas utilizando os dados simulados do modelo *Threshold* e utilizando simetria composta com r=0.2.

A.2.1.2 Correlação = 0.5

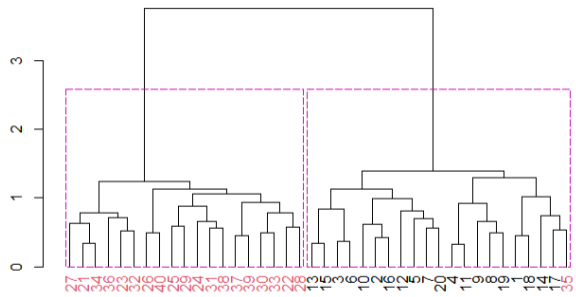
A Figura 10 corresponde aos dendrogramas obtidos para os dados gerados do modelo *Threshold* utilizando a estrutura de simetria composta com correlação igual a 0.5.



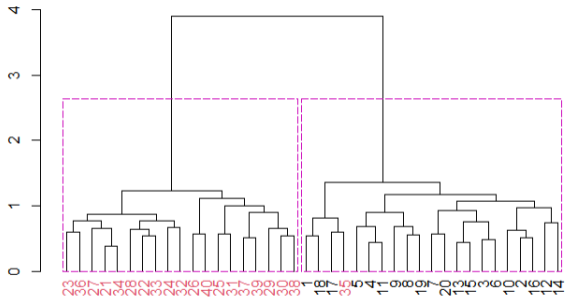
(a) Coeficientes.



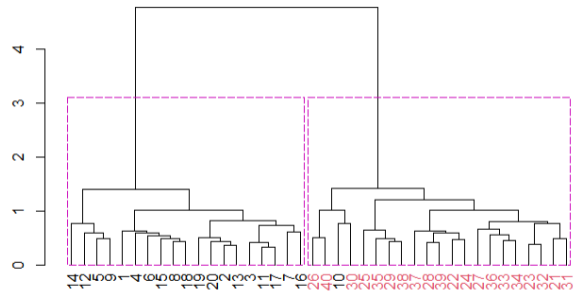
(b) Valor-p.



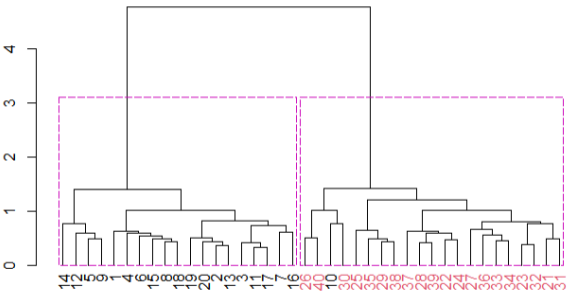
(c) *Threshold Model*.



(d) *Threshold Model* - com matriz positiva definida.



(e) Correlação de Pearson.

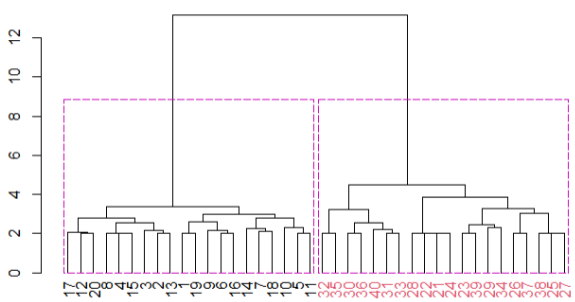


(f) Correlação de Spearman.

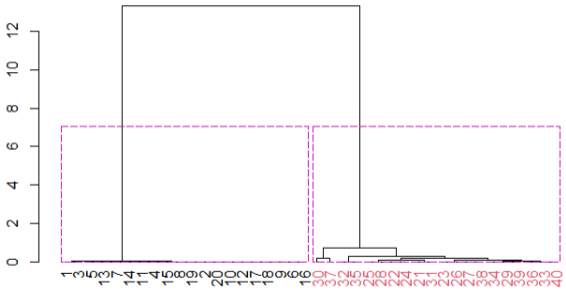
Figura 10 – Dendrogramas utilizando os dados simulados do modelo *Threshold* e utilizando simetria composta com r=0.5.

A.2.1.3 Correlação = 0.8

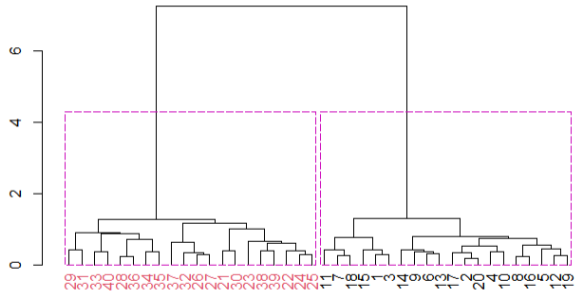
A Figura 11 corresponde aos dendrogramas obtidos para os dados gerados do modelo *Threshold* utilizando a estrutura de simetria composta com correlação igual a 0.8.



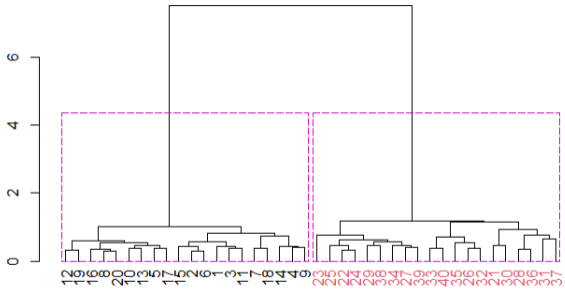
(a) Coeficientes.



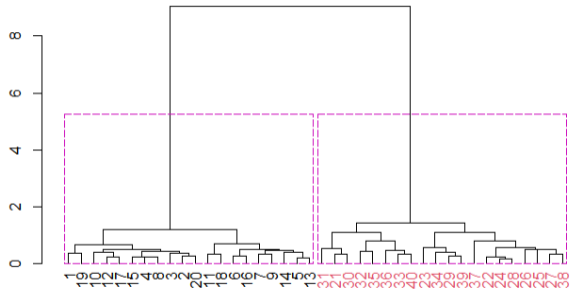
(b) Valor-p.



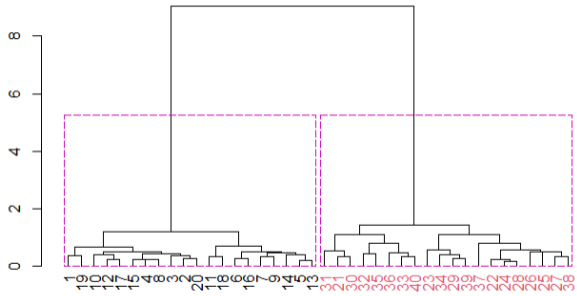
(c) *Threshold Model*.



(d) *Threshold Model* - com matriz positiva definida.



(e) Correlação de Pearson.



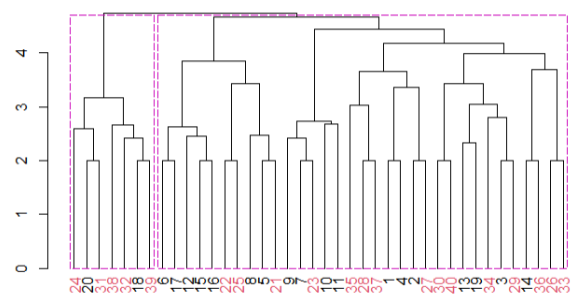
(f) Correlação de Spearman.

Figura 11 – Dendrogramas utilizando os dados simulados do modelo *Threshold* e utilizando simetria composta com r=0.8.

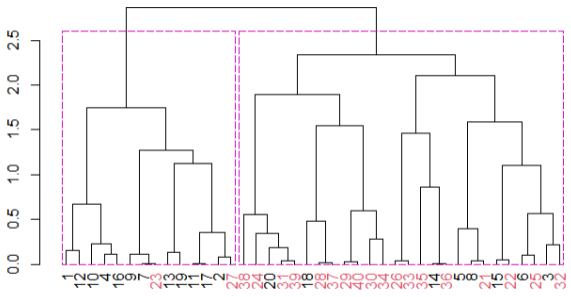
A.2.2 Estrutura autoregressiva

A.2.2.1 Correlação = 0.2

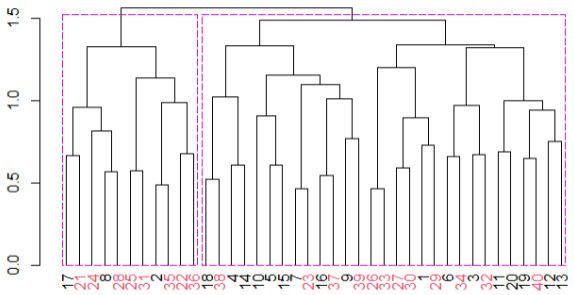
A Figura 12 corresponde aos dendrogramas obtidos para os dados gerados do modelo *Threshold* utilizando a estrutura autoregressiva com correlação igual a 0.2.



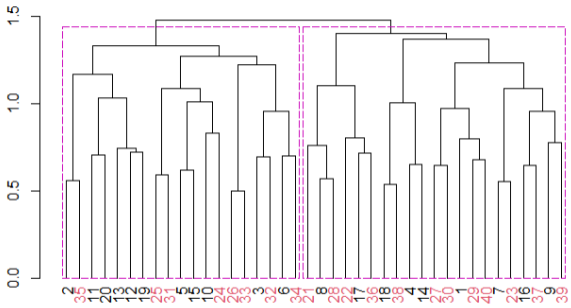
(a) Coeficientes.



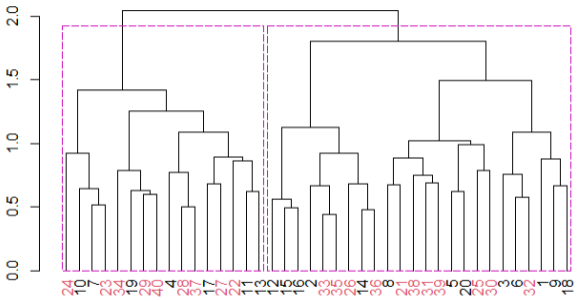
(b) Valor-p.



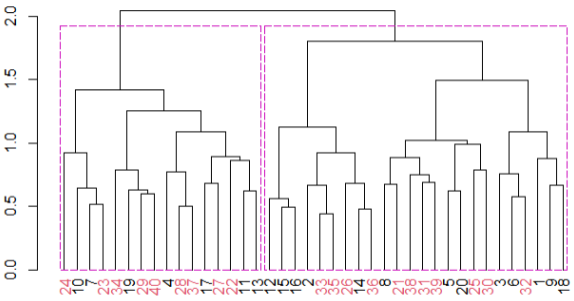
(c) *Threshold Model*.



(d) *Threshold Model* - com matriz positiva definida.



(e) Correlação de Pearson.

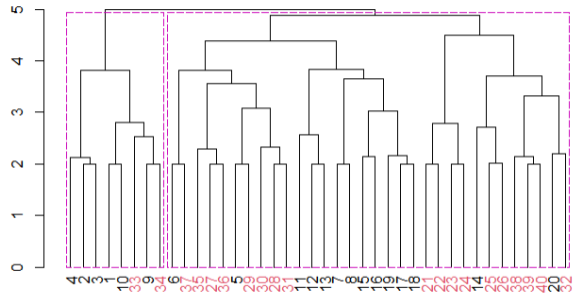


(f) Correlação de Spearman.

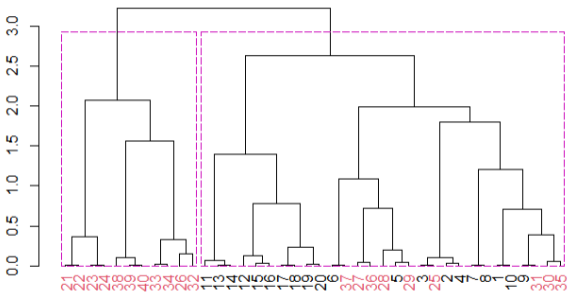
Figura 12 – Dendrogramas utilizando os dados simulados do modelo *Threshold* e utilizando a estrutura autoregressiva com r=0.2.

A.2.2.2 Correlação = 0.5

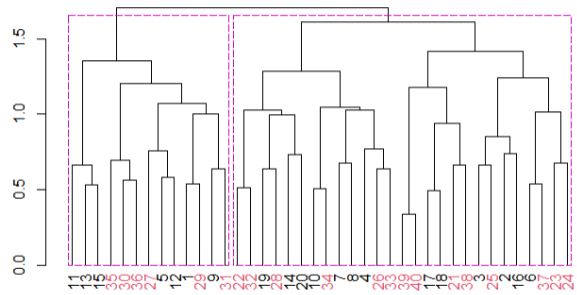
A Figura 13 corresponde aos dendrogramas obtidos para os dados gerados do modelo *Threshold* utilizando a estrutura autoregressiva com correlação igual a 0.5.



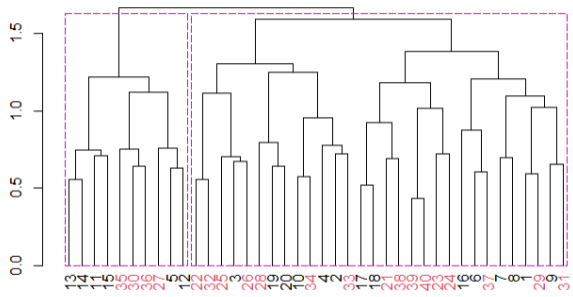
(a) Coeficientes.



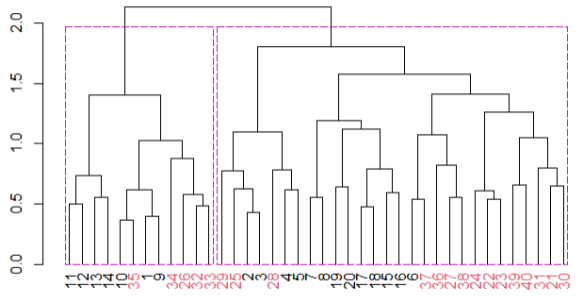
(b) Valor-p.



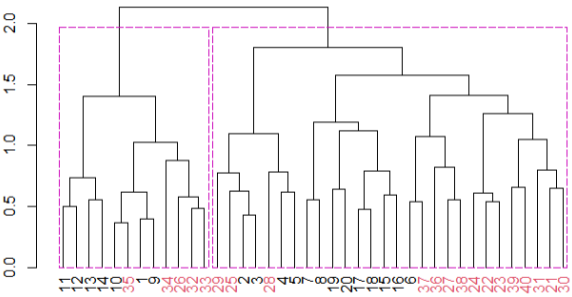
(c) *Threshold Model*.



(d) *Threshold Model* - com matriz positiva definida.



(e) Correlação de Pearson.

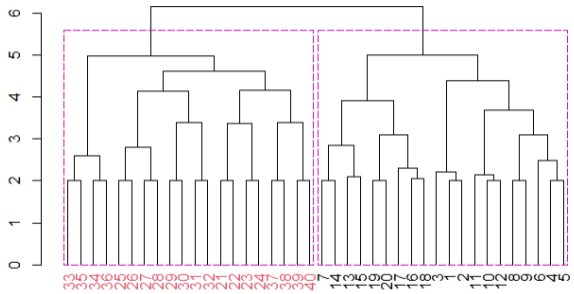


(f) Correlação de Spearman.

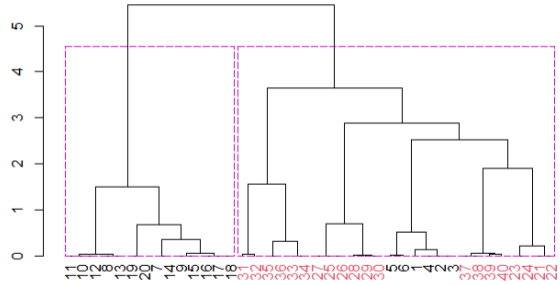
Figura 13 – Dendrogramas utilizando os dados simulados do modelo *Threshold* e utilizando a estrutura autoregressiva com r=0.5.

A.2.2.3 Correlação = 0.8

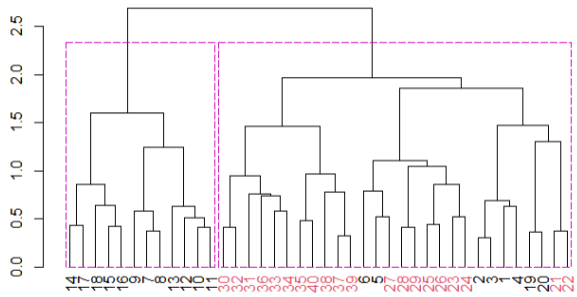
A Figura 14 corresponde aos dendrogramas obtidos para os dados gerados do modelo *Threshold* utilizando a estrutura autoregressiva com correlação igual a 0.8.



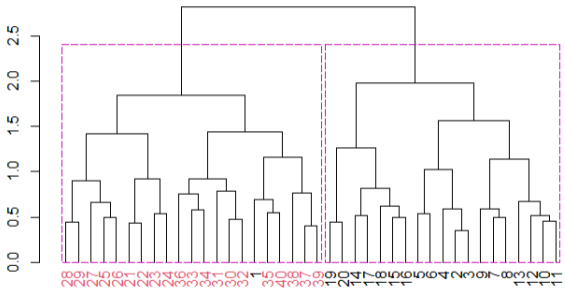
(a) Coeficientes.



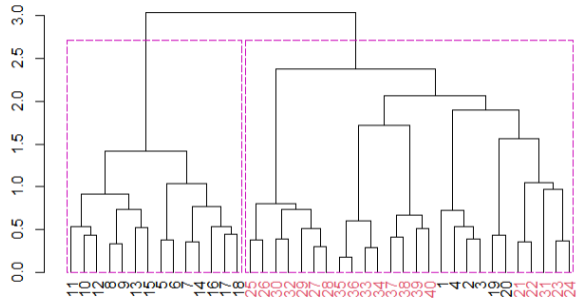
(b) Valor-p.



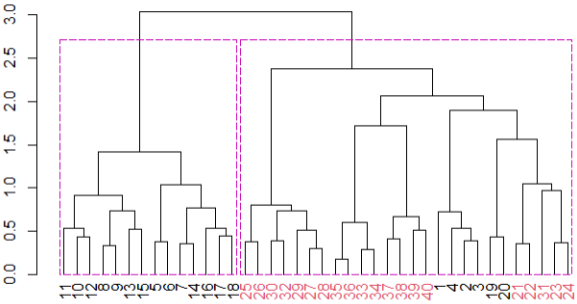
(c) *Threshold Model*.



(d) *Threshold Model* - com matriz positiva definida.



(e) Correlação de Pearson.



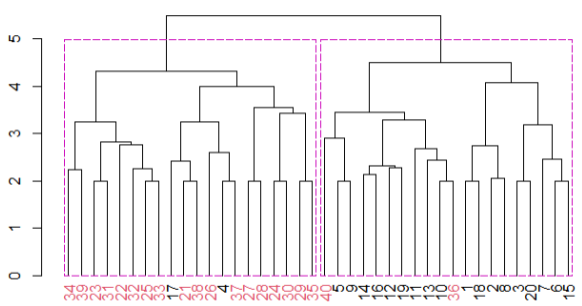
(f) Correlação de Spearman.

Figura 14 – Dendrogramas utilizando os dados simulados do modelo *Threshold* e utilizando a estrutura autoregressiva com r=0.8.

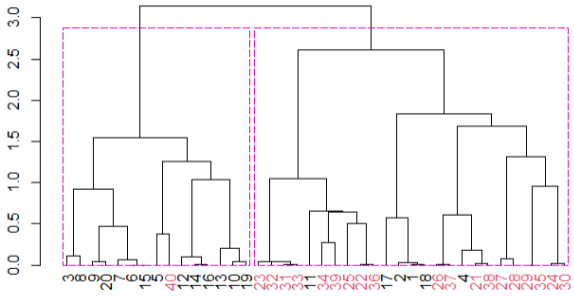
A.2.3 Estrutura de simetria composta - (segunda aplicação)

A.2.3.1 Correlação = 0.2

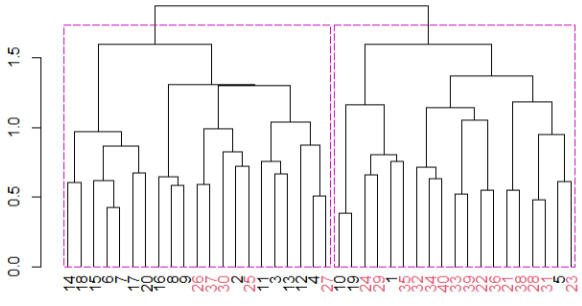
A Figura 15 corresponde aos dendrogramas obtidos para os dados gerados do modelo *Threshold* utilizando a estrutura de simetria composta com correlação igual a 0.2.



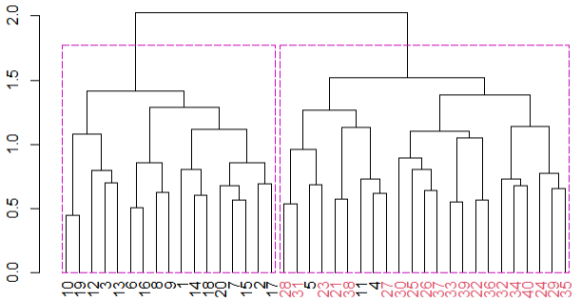
(a) Coeficientes.



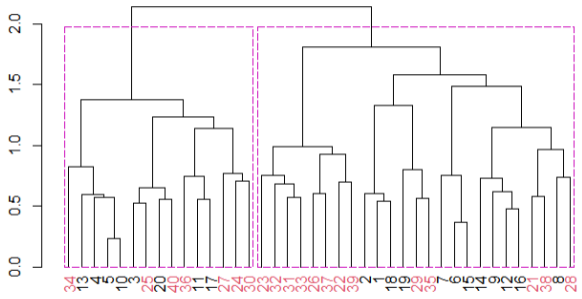
(b) Valor-p.



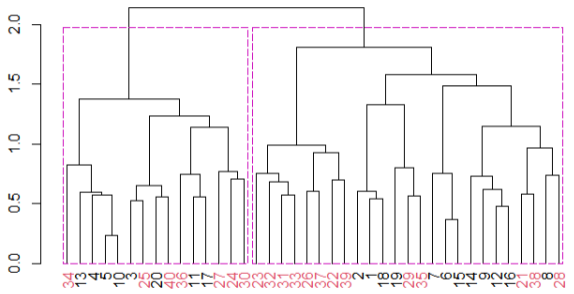
(c) *Threshold Model*.



(d) *Threshold Model* - com matriz positiva definida.



(e) Correlação de Pearson.

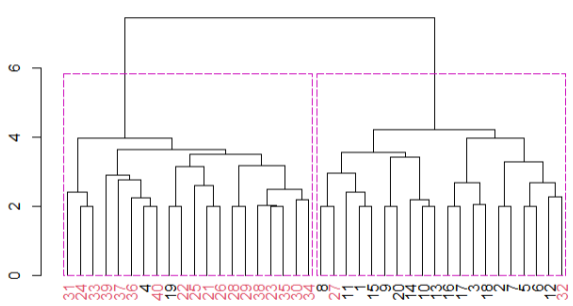


(f) Correlação de Spearman.

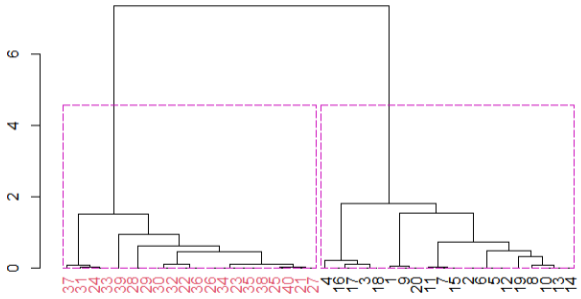
Figura 15 – Dendrogramas utilizando os dados simulados do modelo *Threshold* e utilizando simetria composta com r=0.2.

A.2.3.2 Correlação = 0.5

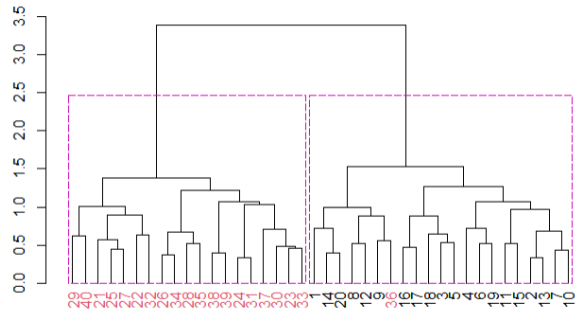
A Figura 16 corresponde aos dendrogramas obtidos para os dados gerados do modelo *Threshold* utilizando a estrutura de simetria composta com correlação igual a 0.5.



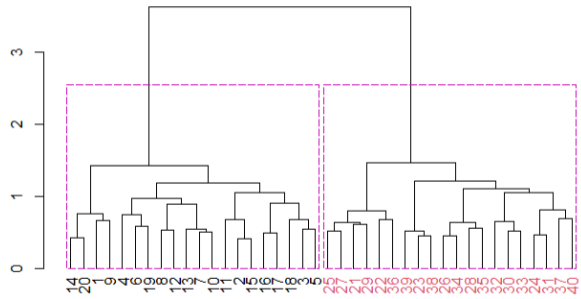
(a) Coeficientes.



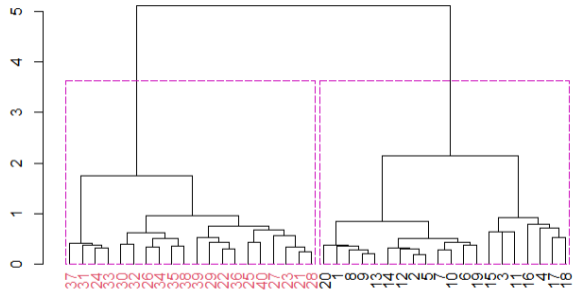
(b) Valor-p.



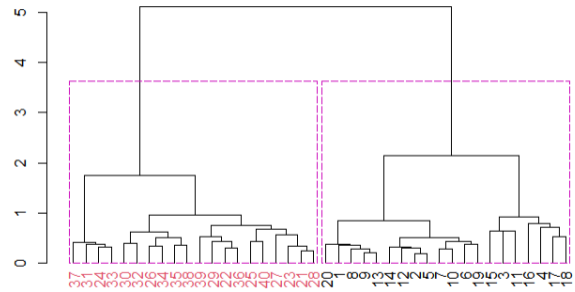
(c) *Threshold Model*.



(d) *Threshold Model* - com matriz positiva definida.



(e) Correlação de Pearson.

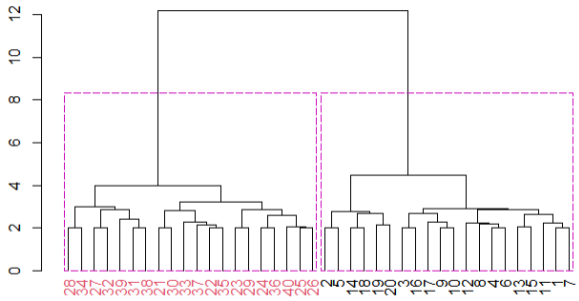


(f) Correlação de Spearman.

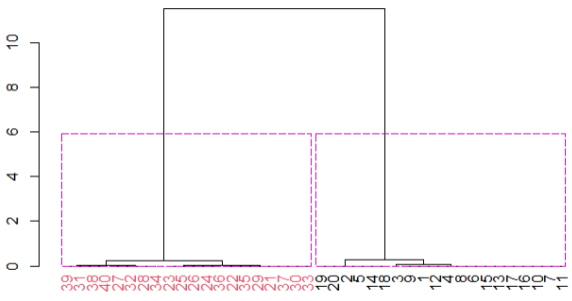
Figura 16 – Dendrogramas utilizando os dados simulados do modelo *Threshold* e utilizando simetria composta com r=0.5.

A.2.3.3 Correlação = 0.8

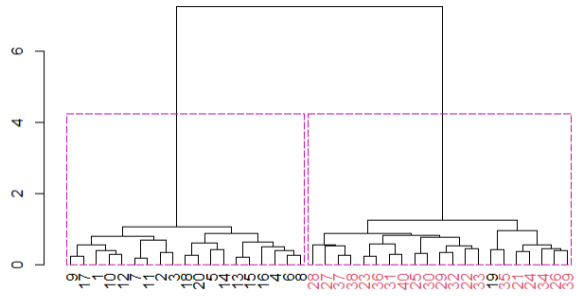
A Figura 17 corresponde aos dendrogramas obtidos para os dados gerados do modelo *Threshold* utilizando a estrutura de simetria composta com correlação igual a 0.8.



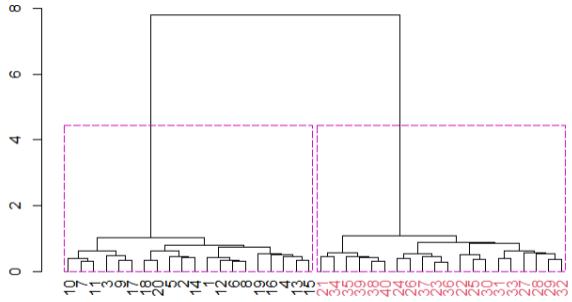
(a) Coeficientes.



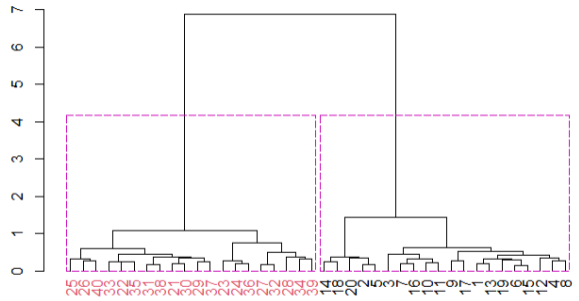
(b) Valor-p.



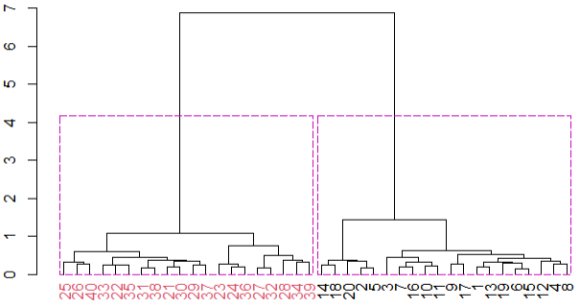
(c) *Threshold Model*.



(d) *Threshold Model* - com matriz positiva definida.



(e) Correlação de Pearson.



(f) Correlação de Spearman.

Figura 17 – Dendrogramas utilizando os dados simulados do modelo *Threshold* e utilizando simetria composta com r=0.8.


ICMC
USP
SÃO CARLOS