Avaliação da reconstrução de caractere em ancestral comum e estimação de correlações pelo modelo filogenético de variável latente

Lauren Alves Vieira [1] [3]

Gabriela Bettella Cybis[2] [3]

Resumo: O estudo de correlações entre variáveis fenotípicas ao longo da evolução é um dos problemas centrais da biologia evolutiva. O modelo filogenético de variável latente (Cybis et al. 2015) é uma opção para a estimação de tais correlações no contexto das filogenias bayesianas. O modelo permite a estimação simultânea de correlações entre variáveis contínuas, discretas ordenadas e discretas sem ordenamento, controlando para a história evolutiva compartilhada das amostras. Neste trabalho nós realizamos uma aplicação do modelo a um conjunto de dados de morcegos, que contem uma variável contínua e uma discreta não ordenada, na qual estimamos a correlação evolutiva entre as variáveis e reconstruímos o valor dessas variáveis no ancestral comum a todas as espécies em estudo. Como nos modelos ordenados a aplicação do modelo depende da escolha de um estado de referência, nós realizamos uma análise de sensibilidade, verificando que em geral estas estimativas são robustas à escolha do referencial.

Palavras-chave: *Variável latente, Filogenias, Inferência bayesiana*.

1 Introdução

O estudo das interações entre genótipos e fenótipos é um dos grandes focos da biologia evolutiva, com aplicações nas mais diversas áreas. Nesse contexto, uma questão de interesse é a estimação de correlações nos processos evolutivos de traços fenotópicos. Entretanto, para adequadamente estimar essas correlações, devemos separá-las das correlações induzidas pela história evolutiva compartilhada entre os indivíduos, que pode ser inferida a partir de dados genéticos. O modelo filogenético de variável latente (Cybis et al 2015) mostra-se como uma opção para estas análises, já que pode ser usado para estimar correlações entre diferentes tipos de dados fenotípicos, enquanto controla para a história evolutiva compartilhada dos indivíduos ou espécies em estudo.

[1]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *laurendiasalves@gmail.com*

[2]UFRGS - Universidade Federal do Rio Grande do Sul. Email: *gabriela.cybis@ufrgs.br*

[3]Um agradecimento especial a Gislene Lopes Gonçalves e Tiago Ferraz, do PPGBM da UFRGS pela disponibilização dos dados.

A separação de correlações inerentes aos processos de evolução dos fenótipos daquelas geradas pela história evolutiva é importante para a identificação de dois fenômenos de interesse biológico: ligação gênica e seleção natural. O estudo da evolução da resistência bacteriana a diferentes antibióticos é um exemplo de problema de interesse epidemiológico em que correlações na evolução de fenótipos são um indício de ligação gênica. De modo similar, pressões seletivas entre características com hábitos alimentares e traços morfológicos em grupos de mamíferos também podem ser estudadas por meio de correlações evolutivas.

Até onde temos conhecimento, o modelo filogenético de variável latente é o único modelo proposto que permite estimar correlações evolutivas entre traços contínuos, discretos binários e discretos com múltiplos estados ordenados ou não. Além disso, a metodologia de inferência bayesiana associada ao modelo e implementada na plataforma BEAST (software de inferência bayesiana para filogenias - Drumond et al. 2007) permite que estas correlações sejam estimadas, mesmo quando não se conhece a história evolutiva de modo preciso, fazendo uso direto de sequências de DNA. Adicionalmente, quando se considera o histórico de análises de correlações no contexto filogenético, nossa metodologia permite a análise de conjuntos de dados relativamente grandes.

Neste trabalho realizaremos a análise de um conjunto de dados de morcegos, disponibilizado por colaboradores (dados ainda não publicados), com o objetivo de estudar a correlação evolutiva entre o número de dentes e hábito alimentar destas espécies. Além disso, estimamos os valores dessa característica no ancestral comum do grupo de morcegos. Como a realização desta análise envolve algumas escolhas de referencial de parâmetros, realizamos uma pequena análise de sensibilidade para verificar o efeito destas escolhas sobre a estimação.

2 Metodologia

Modelo Filogenético de Variável Latente

Representamos a história evolutiva de um conjunto de N indivíduos por meio de um grafo acíclico denominado árvore filogenética F (ou filogenia). A árvore conta com N nós externos (vértices de grau 1), que representam os indivíduos da amostra no tempo presente, e uma raiz (vértice de grau 2), que representa o mais recente ancestral comum aos N indivíduos da amostra e o momento mais antigo no tempo representado na árvore. Além disso, há $N - 2$ nós internos (vértices de grau 3), que representam bifurcações evolutivas causadas pela separação de linhagens. O comprimento da arestas ligando esses nós representa a quantidade de tempo evolutivo até a ocorrência de uma bifurcação. A evolução das variáveis fenotípicas é modelada por um processo estocástico que inicia na raiz da árvore e evolui ao longo das arestas até os nós externos, onde o valor das variáveis nos N indivíduos da amostra é determinado. A

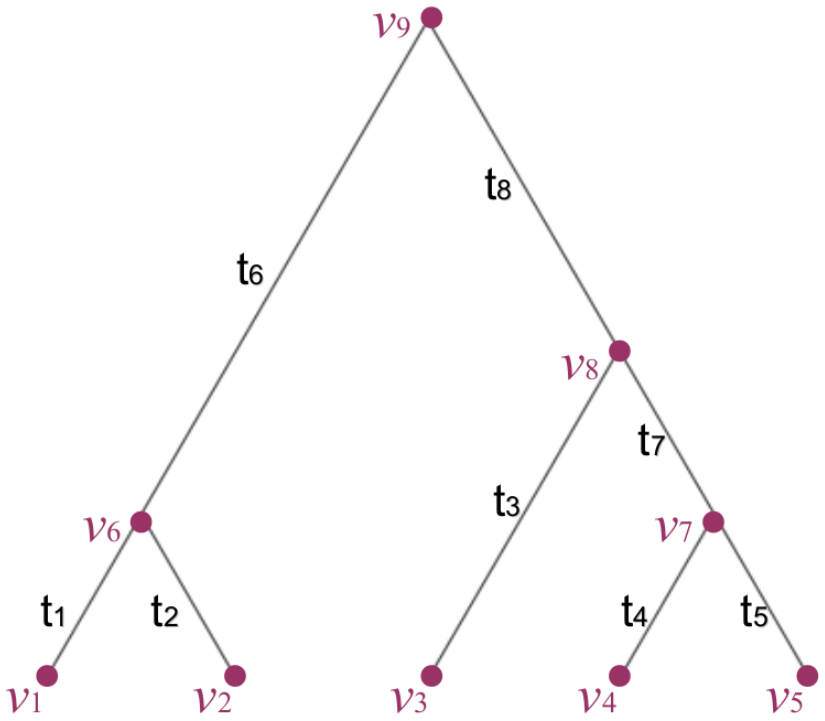

Figura 1: Exemplo de árvore filogenética com $N = 5$.

Figura 1 apresenta um exemplo de filogenia.

O modelo filogenético de variável latente descreve a evolução de uma variável latente X contínua, não observável, ao longo da árvore filogenética F, e a de uma variável de interesse observável Y. A evolução temporal da variável latente X segue um modelo de movimento browniano ao longo da filogenia. Ao final do processo, o valor da variável Y é determinado a partir de X por meio de uma função de ligação $g(X)$. Quando a variável Y é binária, por exemplo, seu valor é determinado pela posição de X em relação a um limiar, e quando Y é contínua temos $Y = X$. No caso de Y multivariado, cada componente de Y é determinada por mais de uma componente de X. A matriz de precisão Σ^{-1} do movimento browniano multivariado que descreve a evolução de X é utilizada como um proxy para estimar a correlação evolutiva entre as variáveis componentes de Y. Este modelo foi inspirado pelo modelo limiar filogenético (Felsenstein 2005).

Para calcular a função de verossimilhança para esse modelo, consideramos uma extensão dos dados de modo que $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$, em que $\mathbf{Y} = (Y_0, ..., Y_N)$ são os valores da variável D-dimensional de interesse Y observados nos N nós externos da filogenia (amostra), e $\mathbf{X} = (X_0, ..., X_N)$ são os valores

da variável latente D-dimensional X nos mesmos nós. O movimento browniano ao longo da árvore F que descreve a evolução de X é um processo já bem explorado na literatura (Felsenstein, 1988), e sua densidade $P(\mathbf{X}|\Sigma^{-1}, F)$ pode ser calculada por meio de um algoritmo iterativo que computa uma série de convoluções de distribuições normais D-variadas ao longo das arestas de F. Desse modo, temos

$$P(X, Y|F, \Sigma^{-1}) = P(X|F, \Sigma^{-1})P(Y|X).$$

No caso de variáveis Y binárias, definimos $P(X|Y)$ como

$$P(Y|X) = \prod_{i=1}^{N}\prod_{j=1}^{D} \left(\mathbf{I}(y_{i,j} = 1)\mathbf{I}(x_{i,j} > 0) + \mathbf{I}(y_{i,j} = 0)\mathbf{I}(x_{i,j} \leq 0)\right),$$

em que $\mathbf{I}(A)$ é a função indicadora de A, e $x_{i,j}$ e $y_{i,j}$ são a j-ésima componente das respectivas variáveis no nó i. Ou seja, em cada coordenada, temos $Y = 1$ se a variável latente é maior do que zero, e $Y = 0$ caso contrário.

Quando Y é contínuo, tomamos $Y = X$, fixando o valor da variável latente nos nós externos. Já quando Y é uma variável categórica com k estados não ordenados, então a uma entrada de Y correspondem $k - 1$ variáveis latentes em X. O valor observado $y_{i,j}$ na componente j da observação i é determinado pela maior das variáveis latentes correspondentes $\{x_{i,j'}, \cdots x_{i,j'+k-2}\}$ de modo que a função link é dada por

$$y_{ij} = g(x_{i,j'}, \cdots x_{i,j'+k-2}) = \begin{cases} s_1 \quad se \quad 0 = sup(0, x_{i,j'}, \cdots x_{i,j'+k-2}) \\ s_l \quad se \quad x_{i,l} = sup(0, x_{i,j'}, \cdots x_{i,j'+k-2}), \end{cases}$$

em que, sem perda de generalidade, tomamos o primeiro estado s_1 como o estado de referência. Neste caso

$$P(Y|X) = \prod_{i=1}^{N}\prod_{j=1}^{D} \left(\mathbf{I}(y_{i,j} = g(x_{i,j'}, \cdots x_{i,j'+k-2})\right).$$

Também podemos naturalmente considerar a extensão em que alguns componentes de Y são discretos e outros contínuos.

Inferência nesse modelo é feita em uma perspetiva Bayesiana, de modo que calculamos a distribuição à posteriori como

$$P(\Sigma|X, Y, F) \propto P(X, Y|F, \Sigma^{-1})P(\Sigma) = P(Y|X)P(X|F, \Sigma^{-1})P(\Sigma),$$

em que utilizamos a distribuição Whishart para distribuição à priori $P(\Sigma)$. Para fazer inferência baseada nesse modelo utilizamos um algorítimo de MCMC.

Para estimara o valor da variável de interesse Y no ancestral comum (raiz da filogenia), consideramos a extensão $\mathbf{Z}^* = (\mathbf{Y}, \mathbf{X}^*)$, em que $\mathbf{X}^* = (X_0, ..., X_{2N-1})$ são os valores da variável latente D-dimensional X em todos os nós da árvore. O algoritmo de MCMC é utilizado para obter a distribuição à posteriori de X_{2N-1}, e a função de ligação $g(X)$ é utilizada para recuperar os valores correspondentes de Y.

3 Resultados

Aplicação

Consideramos um conjunto de dados de 41 diferentes espécies de morcegos cedido por colaboradores (dados ainda não publicados). Os dados consistem de uma árvore filogenética que relaciona as espécies e, para cada espécie, informações sobre o número de dentes, variando entre 20 e 36. Além disso temos dados sobre os hábitos alimentares das espécies, divididos em k=6 categorias frugívoro (11 espécies), insetívoro (10), onívoro (6), carnívoro (3), nectarívoro (10) e hematófago (3). Como não há ordenamento inerente entre estas categorias, empregamos $k - 1 = 5$ dimensões da variável latente X para determinar o hábito alimentar e uma para modelar o número de dentes. Consideramos o hábito frugívoro como o referencial, e realizamos a análise no BEAST para estimação de correlações e reconstrução da raiz.

Tomamos como critério de significância para uma correlação que o seu intervalo de credibilidade 95% (IC) não inclua o zero. Isto é equivalente à probabilidade à posteriori de a covariância ser positiva (sig) ser superior a 0.975 ou inferior à 0.025. Apenas duas das entradas da matriz Σ foram consideradas significativas, as respectivas estimativas para a correlação e valor de sig estão apresentados na primeira linha da Tabela 1. Na reconstrução do número de dentes no ancestral comum, a média a posteriori foi 32.5 com IC de [29.8; 34.3]. Já o hábito alimentar estimado para o ancestral comum, com uma probabilidade à posteriori de 0.765, foi insetívoro.

A Figura 2 apresenta a reconstrução do número de dentes, estimada pela média da distribuição à posteriori, e do hábito alimentar, segundo o estado com maior probabilidade à posteriori, em toda a árvore filogenética destas espécies. Notamos que embora boa parte da evolução destas espécies provavelmente tenha ocorrido no estado frugívoro, há alta probabilidade à posteriori que o ancestral comum (raiz no centro da figura) seja insetívoro.

Análise de Sensibilidade

Notamos que o modelo de variável latente, no caso de variáveis com múltiplos estados não ordenados não é perfeitamente simétrico em relação a todos os estados. O estado de referência tem uma probabilidade à priori superior aos outros estados quando $k = 6$. Além disso, a interpretação de corre-

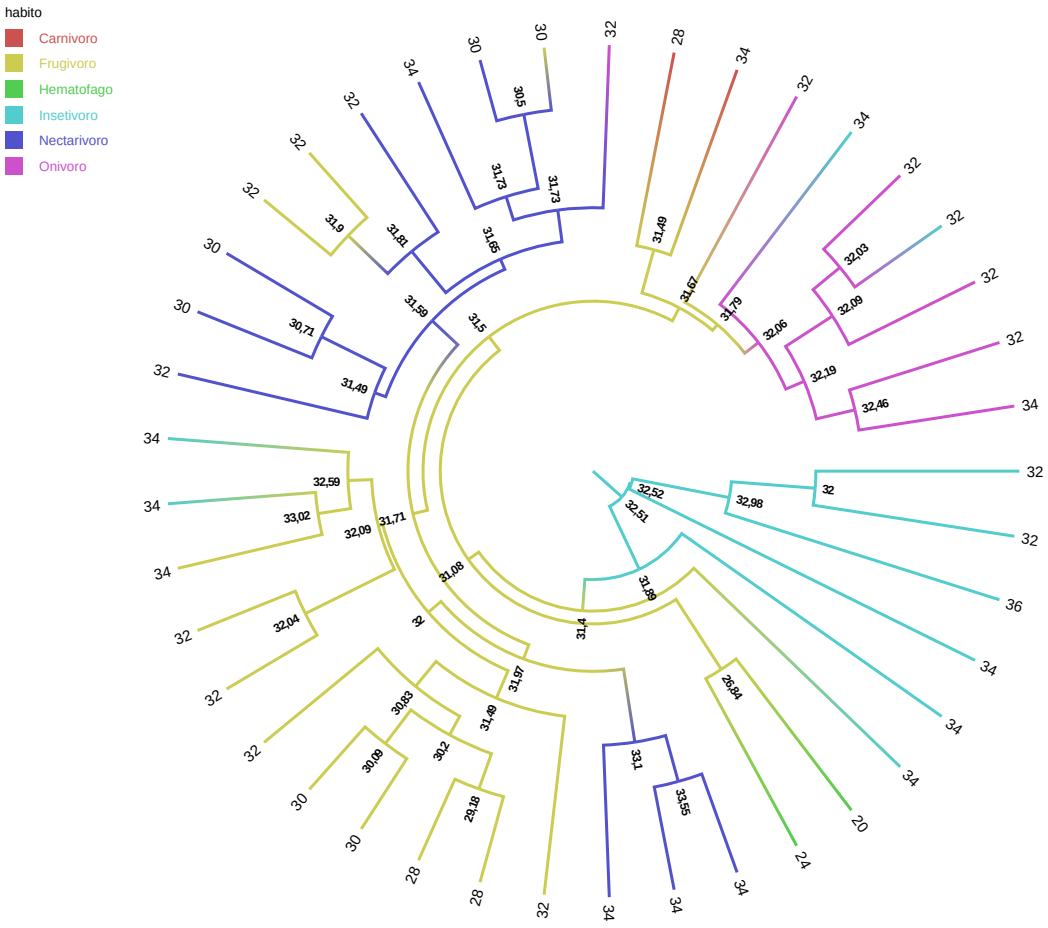

Figura 2: Reconstrução de hábito alimentar (representados por cores) e número de dentes (representados por números na figura) sobre a árvore filogenética para as 41 espécies de morcego consideradas.

Tabela 1: Correlações relevantes entre os hábitos alimentares e o número de dentes

	Dentes×Insetívoro		Dentes×Hematófago	
	Cor	Sig	Cor	Sig
Frugívoro	0.461	0.979	−0.501	0.007
Insetívoro	Referência		−0.575	0.002
Onívoro	0.476	0.960	−0.578	0.145
Nectarívoro	0.471	0.964	−0.548	0.012
Hematófago	0.533	0.994	Referência	

Tabela 2: Estimação do número de dentes e do hábito alimentar dos morcegos no ancestral comum situado na raiz da filogenia.

	Dentes		Hábito Alimentar						
				Probabilidade à Posteriori					
Referência	Média	MAP	MAP	Frugívoro	Insetívoro	Onívoro	Carnívoro	Nectarívoro	Hematófago
Frugívoro	32.502	31.403	Frugívoro	0.122	0.765	0.055	0.012	0.044	0.001
Insetívoro	32.476	30.673	Frugívoro	0.053	0.837	0.039	0.014	0.051	0.006
Onívoro	32.413	32.059	Insetívoro	0.062	0.759	0.123	0.009	0.043	0.003
Carnívoro	32.505	32.737	Insetívoro	0.060	0.794	0.057	0.032	0.053	0.003
Nectarívoro	32.449	31.935	Nectarívoro	0.044	0.757	0.057	0.012	0.125	0.004
Hematófago	32.473	33.529	Insetívoro	0.059	0.821	0.050	0.016	0.039	0.014

lações para este estado é indireta. Assim, buscamos responder à seguinte questão: A escolha do estado de referência afeta a estimação? Para tanto, repetimos a análise deste conjunto de dados considerando os outros estados como referenciais. A Tabela 1 apresenta as estimativas de correlações significativas na análise original, considerando os diferentes estados como referencial. A tabela apresenta as estimativas à posteriori para a correlação (cor) e o valor de sig. Notamos que tanto as estimativas de correlação quanto os valores de sig em geral são consistentes para os diferentes modelos. Uma exceção é a estimativa da correlação entre dentes e insetívoro quando o estado de referência é hematófago. Em todos os outros modelos, há correlação importante entre dentes e hematófago. Como neste caso não há uma variável latente específica ligada ao estado hematófago, o efeito desta correlação neste modelo acaba sendo manifestado em todos os outros estados.

Já a Tabela 2 apresenta as reconstruções para a raiz da filogenia em cada um destes casos. Para o número de dentes, apresentamos dois métodos de estimação a média à posteriori (Média) e o máximo a posteriori (MAP). Notamos que as estimativas pela média são muito mais consistentes considerando os diferentes referenciais. Para o hábito alimentar, apresentamos as estimativas por MAP e considerando a probabilidade à posteriori de cada estado. Pelo método do MAP, notamos variação na estimativa para a riz. Já quando a probabilidade à posteriori é considerada, concluímos que o estado da raiz é insetívoro com probabilidade superior a 0.7 para todos os referenciais. Assim, observamos que o método MAP parece não ser robusto para este tipo de estimação. Para os outros métodos, notamos que a escolha do estado de referência tem pouquíssimo efeito sobre a estimação da raiz.

4 Conclusões

Neste trabalho utilizamos a análise de um conjunto de dados de morcegos para verificar o comportamento da estimação das correlações evolutivas e reconstrução de caractere ancestral com o modelo filogenético de variável latente. Ao realizar a análise de dados considerando diferentes estados de referência para a variável com múltiplos estados não ordenados, percebemos que tanto as estimações de correlações quanto as do valor da variável na raiz aparentam ser robustas à escolha do referencial. Este

é um resultado importante pois destaca a confiança das estimativas obtidas, independente da escolha de estado de referência, que frequentemente é feita de forma arbitrária.

Referências

[1] CYBIS, G.B.; SINSHEIMER, J.S.; BEDFORD, T.; MATHER, A.E., LEMEY; P. and SUCHARD, M.A. Assessing phenotypic correlation through the multivariate phylogenetic latent liability model.*The Annals of Applied Statistics* , v. 9(2), p969-991, 2015.

[2] DRUMMOND AJ; RAMBAUT A. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, v. 7, p. 214, 2007.

[3] FELSENSTEIN J. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, v. 1, p. 445-71, 1988.

[4] FELSENSTEIN J. Using the Quantitative Genetics Threshold Model for Inferences Within and Between Species. *Philosophical Transactions of the Royal Society B*, v. 360, p. 1427-1434, 2005.

[5] KINGMAN, J. F. C. The coalescent. Stochastic processes and their applications, v. 13(3), p. 235-248, 1982.