Estudo da Sensibilidade do Bayes Factor para seleção de modelos

Lauren Alves Vieira [1]

Gabriela Bettella Cybis[2]

Resumo: Métodos bayesianos filogenéticos são uma ferramenta central na biologia evolutiva. Dentre estes o Modelo de Variável Latente estima correlações entre características fenotípicas (contínuas e categóricas ordinais ou nominais), controlando para história evolutiva entre os indivíduos amostrados. Nas aplicações deste modelo é comum a escolha de prioris pouco informativas, geralmente adotando a distribuição conjugada Wishart Inversa para matriz de covariâncias do modelo.

Nossos resultados prévios evidenciaram uma possível sensibilidade do método de seleção de modelos quanto a escolha da priori, de modo que modelos com maior número de graus de liberdade (**gl**), pareciam ser favorecidos. Com o intuito de avaliar esse efeito da priori sobre a seleção do modelo, foi conduzido o estudo apresentado abaixo.

Palavras-chave: *Variável latente, Bayes Factor, Priori*.

1 Introdução

O estudo de correlações evolutivas é um dos grandes focos da biologia evolutiva, com aplicações nas mais diversas áreas. Neste contexto, está a estimação de correlações nos processos evolutivos de traços fenotípicos. Entretanto para estimar adequadamente estas correlações devemos separá-las das correlações induzidas pela história evolutiva compartilhada entre os indivíduos, que pode ser inferida através de dados. 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 fenotípico enquanto controla para história evolutiva compartilhada dos indivíduos ou espécies em estudo.

A diferenciação entre correlações inerentes ao processo de evolução dos fenótipos e correlações geradas pela história evolutiva é necessária para 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

[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*

são um indício de ligação gênica. De modo similar, pressões seletivas entre características como hábitos alimentares e traços morfológicos em grupos de mamíferos também podem ser estudadas por meio de correlações evolutivas.

Para estimação deste tipo de correlação é comum o uso de uma abordagem Bayesiana. Dentre outros modelos para este tipo de relação o Modelo de Variável Latente utiliza uma transformação bijetora que relaciona uma variável latente a uma variável observável. A forma entre essas variáveis depende do tipo de variável em estudo (Contínua, Binária, Categórica Ordinal ou Nominal). Em alguns casos, como no estudo de hábitos alimentares de morcegos, é difícil se escolher um modelo para os dados, uma vez que não existam informações prévias ou mesmo indícios que estes dados são Ordinais ou Nominais. Para verificação do ajuste do modelo aos dados é comum o uso do método de Bayes Factor (Gelman et al.2003), que compara pares de modelos quanto ao seu ajuste a um mesmo conjunto de dados.

Em trabalhos prévios que visavam avaliar as propriedades estatísticas do Modelo de Variável Latente, obtivemos resultados que evidenciavam que o método de Bayes Factor é afetado pela escolha da priori. Neste estudo consideramos amostras de características Ordinais, para as quais utilizamos o Modelo de Variável Latente, considerando um modelo para os dados ora ordinal e ora nominal, comparando os resultados através de Bayes Factor. Se percebeu uma provável sensibilidade do método a escolha de priori. Para melhor compreender este comportamento realizamos o breve estudo descrito neste trabalho.

2 Metodologia

Modelo Filogenético de Variável Latente

A história evolutiva de conjuntos de indivíduos pode ser representada através de uma árvore filogenética (ou filogênia) τ, que nada mais é que um grafo acíclico onde os N nós externos (vértices de grau 1, também chamados folhas) representam os indivíduos da amostra no tempo atual, também possui apenas um nó de grau 2 chamado raiz, que representa o ancestral comum mais recente a todos os indivíduos. Esta estrutura conta ainda com $N - 2$ nós internos (vértices de grau 3), que descrevem as bifurcações evolutivas decorrentes da separação das diferentes linhagens. As arestas (ou galhos) que ligam estes nós representam o tempo evolutivo decorrido até a ocorrência de uma bifurcação, de modo que a o tamanho das arestas é proporcional a esta quantidade. É possível modelar a evolução de variáveis fenotípicas através de um processo estocástico que inicia na raiz da filogenia e evolui ao longo dos galhos da árvore até as folhas onde os valores foram observados. A figura 1 apresenta um exemplo de filogênia.

O modelo filogenético de variável latente descreve a evolução de uma variável observável Y sobre uma filogênia τ, determinada por uma variável X não observável, chamada de variável latente cuja evolução temporal ao longo de τ segue o modelo de movimento browniano. Assim ao final deste processo

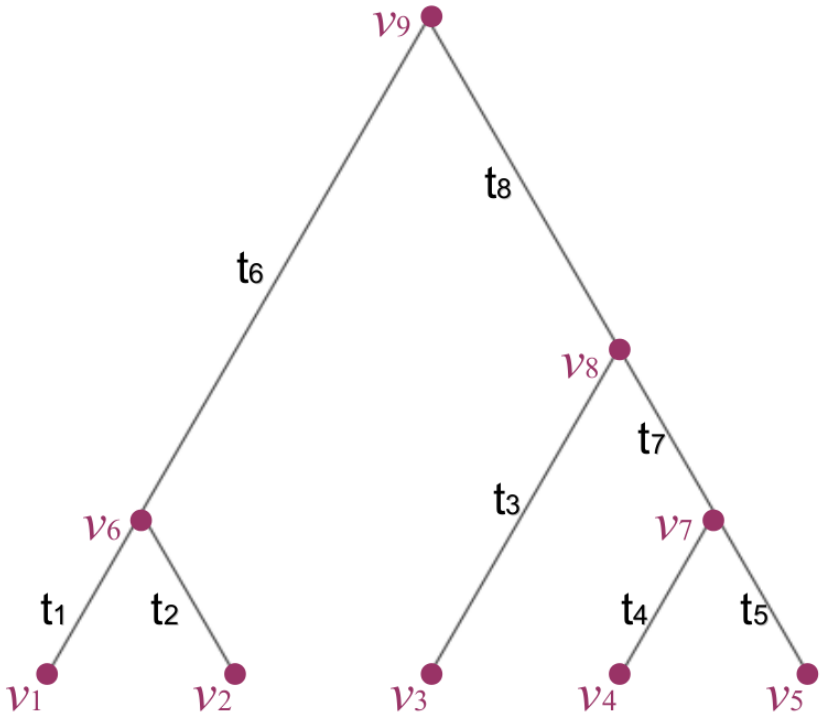

Figura 1: Exemplo de Árvore filogenética com $N = 5$.

a variável Y é determinada por meio de uma função de ligação $g(X)$, a partir dos valores de X. Quando, por exemplo, a variável Y é binária seu valor é determinado pela posição de X em relação a um limiar, já quando Y é contínua temos $Y = X$. No caso de Y multivariado, com estados não ordenados, cada componente de Y é determinada por mais de uma componente de X, porém se Y é multivariado e seus k estados possuem algum tipo de ordenamento, então seus valores são determinados pela posição de X quanto a $k-1$ limiares. Este modelo foi inspirado pelo modelo limiar filogenético. A matriz de precisão Σ^{-1} do movimento browniano multivariado que descreve a evolução de X é utilizada como um proxi para estimar a correlação evolutiva entre as variáveis componentes de Y (Felsenstein 2005).

Para o cálculo da função de verossimilhança deste modelo, consideramos uma extensão dos dados Z, tal que $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$, onde $\mathbf{Y} = (Y_0, ..., Y_N)$ são os valores observados da variável D-dimensional de interesse Y nos N indivíduos da amostra (folhas da filogênia), 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 τ que descreve a evolução de X é um processo já longamente explorado na literatura (Felsenstein, 1988), e sua densidade $P(\mathbf{X}|\Sigma^{-1}, \tau)$ 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 τ. Desse modo, temos

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

Se Y é uma variável binária, 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. Logo, 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 nos externos. Se Y é uma variável categórica com k estados ordenados suas entradas são determinadas de acordo com k intervalos definidos na variável latente X a partir de $k-1$ limiares independentes. Já se estas categorias não são ordenadas, então a cada 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 neste caso 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.

Neste modelo a inferência é feita em uma perspectiva Bayesiana, de modo que calculamos a distribuição à posteriori como

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

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

Bayes Factor

O método de Bayes Factor compara duas hipóteses independentes, aplicadas a modelos como

$$\text{O modelo } \boldsymbol{M}_1 \text{ é o correto } \times \text{ O modelo } \boldsymbol{M}_2 \text{ é o correto,}$$

avaliando qual hipótese é mais verossímil, dada a amostra observada. Para isto calcula-se a razão entre as verossimilhanças marginais das hipóteses

$$BF = \frac{L(\boldsymbol{M}_1|\boldsymbol{Y}, \Sigma^{-}, \tau)}{L(\boldsymbol{M}_2|\boldsymbol{Y}, \Sigma^{-}, \tau)}.$$

Assim valores mais próximos de zero são indicativos de que se deve rejeitar a hipótese nula (Gelman 2003).

Para aplicar tal método a estimação da verossimilhança marginal é usualmente feita através de métodos numéricos como *Stepping Stone Sampling* que estima a log-verossimilhança marginal de um modelo (Xie et al 2011), por exemplo

$$\mathbf{P}(\boldsymbol{M}) \propto \int \mathbf{P}(X, Y|\Sigma, \tau, \boldsymbol{M}) \cdot \mathbf{P}(\Sigma) d\Sigma.$$

Esta estimação é feita através de uma integral de linha entre a distribuição a priori e a distribuição a posteriori da hipótese de interesse

$$L(\boldsymbol{M}|\boldsymbol{Y}) = \sum P(\boldsymbol{M}|\boldsymbol{Y}, \Sigma^{-1}, \tau)^{\beta} \cdot P(\boldsymbol{M}|\Sigma^{-1}, \tau),$$

Tabela 1: Resultados de Bayes Factor comparando modelos com diferentes ordenamentos e número de graus de liberdade da priori conjugada Wishart.

	Comparações					
	$N_3 \times O_2$	$N_3 \times O_3$	$N_3 \times O_6$	$O_2 \times O_3$	$O_2 \times O_6$	$O_3 \times O_6$
Média	-13.055	-12.99	-16.55	0.06159	-3.496	-3.577
Mediana	-12.506	-12.57	-16.1	0.02981	-3.561	-3.608
Máximo	-8.502	-8.2	-11.79	2.1348	-1.576	-2.525
Mínimo	-20.876	-21.13	-24.10	-0.7767	-4.047	-4.345
Desvio	2.8691	2.8421	2.7431	0.4619	0.4274	0.4041

onde $\beta \in [0, 1]$

Estudo de Simulação

Para este pequeno estudo foram geradas amostras de tamanho $n = 10$ de uma variável contínua e de uma variável discreta com $k = 3$ estados não ordenados e com raiz fixa. A partir destas amostras foi feita inferência da verossimilhança marginal de cada modelo condicionado aos dados observados, foram feitas $Re = 50$ repetições do experimento.

Para os dados gerados foi assumido que os estados eram independentes enquanto as correlação entre o primeiro estado e a variável contínua era $r = 0.5$ a correlação entre o segundo estado e a variável contínua era seu simétrico, $r = -0.5$.

Em seguida a estimação da log-verossimilhança marginal dos diferentes modelos, foi feita através do método de Stepping Stone Sampler usando MCMC e comparadas através do método de Bayes Factor

3 Resultados preliminares e Conclusões

A variabilidade das estimativas de verossimilhança marginal obtidas com o modelo nominal (sd = 1.998) é maior que as obtida a partir do modelo ordinal (sd ≤ 1.398). O que reflete o fato de que o espaço paramétrico do modelo ordinal é menor que o do modelo nominal com o mesmo número de categorias, indiferente aos dados observados.

Os resultados na Tabela 1 mostram que o BF favorece o modelo ordenado em todos casos, o que talvez seja associado ao pequeno tamanho de amostra.

Se percebe ainda que o modelo ordenado com maior número de graus de liberdade tende a ser favorecido.

4 Conclusões

Podemos concluir por este pequeno estudo que os resultados não se mantém, quando utilizados dados nominais.

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, 9(2): 969-991. 2015.

[2] Drummond AJ, Rambaut A. Beast: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology, 7:214. 2007.

[3] Felsenstein J. Phylogenies and quantitative characters. Annual Review of Ecology and Systematics, 1:445-71. 1988.

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

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

[6] Xie, Wangang, Lewis, Paul O., Fan, Yu, Kuo, Lynn and Chen, Ming-Hui, Improving marginal likelihood estimation for bayesian phylogenetic model selection. Systematic Biology, 60(2): 150-160. 2011.

[7] Gelman, Andrew, Carlin, John B., Stern, Hal S. and Rubin, Donald B., Bayesian data analysis (2d ed). Chapman and Hall/CRC, 2003