CONHECIMENTO FORMAÇÃO INOVAÇÃO
Salão UFRGS 2018

SALÃO DE INICIACAO CIENTIFICA XXX SIC

15 A 19 OUTUBRO CAMPUS DO VALE

UFRGS PROPESQ

Avaliando a sensibilidade do método de Bayes Factor para seleção de modelo quanto à escolha da distribuição a priori

Autor: Lauren Vieira

Orientador: Gabriela Cybis

Introdução

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.

Métodos

1. **Modelo de Variável Latente**

   Seja $\boldsymbol{Y}$ uma matriz com n observações das variáveis de interesse, tal que os indivíduos na amostra estejam conectados por uma filogenia τ. Assumimos que os valores de $\boldsymbol{Y}$ são determinados por uma variável latente $\boldsymbol{X}$, por meio de uma função de ligação $g(\boldsymbol{X})$, tendo $\boldsymbol{X}$ evoluído através de movimento browniano sobre a filogenia τ.

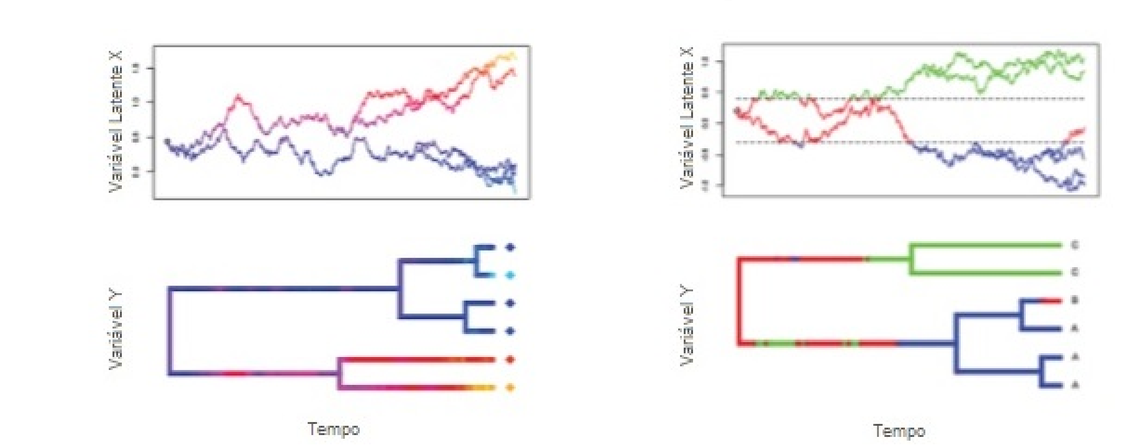

Figura 1: *Evolução da variável latente* ***X*** *através de movimento browniano, que determina os resultados da variável observável* ***Y***

   Desta forma é possível se fazer inferência para este modelo, cuja posteriori é dada por

$$\mathbf{P}(\Sigma^{-}|\boldsymbol{X},\boldsymbol{Y},\tau) \propto \mathbf{P}(X|\tau,\Sigma^{-}) \times \mathbf{P}(Y|X) \times \mathbf{P}(\Sigma),$$

   onde Σ é a matriz de covariâncias de $\boldsymbol{X}$ e $\boldsymbol{Y}$, através de MCMC (Cybis et al. 2015).

2. **Stepping Stone Sampling**

   Assumimos que a variável de interesse segue um modelo $\boldsymbol{M}$ (contínuo, categórico ordinal ou nominal), então podemos estimar numericamente a verossimilhança marginal do modelo,

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

   a partir de uma integral de linha entre a posteriori e a priori, por meio do método de stepping stone sampling (Xie et al. 2011).

3. **Bayes Factor**

   Este método compara dois modelos ($\boldsymbol{M}_1$ e $\boldsymbol{M}_2$) quanto ao seu ajuste aos dados, através da razão entre suas verossimilhanças marginais,

$$BF = \frac{L(\boldsymbol{M}_1|\boldsymbol{Y},\Sigma^{-},\tau)}{L(\boldsymbol{M}_2|\boldsymbol{Y},\Sigma^{-},\tau)}. \quad (1)$$

   Deste modo conforme os valores de BF escolhemos o modelo com melhor ajuste. Assim valores acima de $|\log(BF)| > 2$ definem relações consideradas fortes (Gelman et al.2003).

Estudo de Simulação

Afim de avaliar o método de Bayes Factor para seleção de modelos, foram simulados dados de acordo com dois cenários, no primeiro cenário tínhamos uma variável contínua e uma ordenada (ordinal) e no segundo cenário tínhamos uma variável contínua e uma não ordenada (nominal). Em ambos os casos foram geradas amostras de tamanho $n = 10$, as variáveis discretas com $k = 3$ estados e foram feitas $Re = 50$ replicações.

Em seguida utilizou-se MCMC para estimação da verossimilhança marginal dos modelos. Esta estimação foi feita considerando os diferentes modelos (ordenado e não ordenado) e variando também o valor de **gl** na distribuição a priori (Wishart).

Resultados preliminares

Aqui apresentaremos os resultados obtidos a partir dos dados gerados conforme uma variável categórica nominal.

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.

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

[a] N_3 representa o modelo nominal com Wishart (3); O_2 o modelo ordinal com Wishart (2), O_3 o modelo ordinal com Wishart (3), O_6 o modelo ordinal com Wishart (6). Os dados foram simulados de acordo com o modelo nominal.

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.

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] 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.

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