Universidade: presente!
UFRGS PROPESQ
XXXI SIC
Salão UFRGS 2019 — Conhecimento o formação o inovação
21 . 25 . OUTUBRO . CAMPUS DO VALE

O Modelo de Variável Latente e seu comportamento em diferentes cenários

Autor: Lauren Vieira
Orientador: Gabriela Cybis

Introdução

Dentre os diferentes métodos de estimação de correlação, o Modelo de Variável Latente é pensado para estimação de correlações filogenéticas. Este modelo utiliza uma variável latente, não observável, que evolui na árvore filogenética e determina os valores da variável fenotípica de interesse. A estimação do modelo é feita por MCMC.

Apresentamos aqui os resultados parciais estudo de simulação no qual avaliamos o comportamento do modelo mediante diferentes cenário. Nestes cenários foram avaliados pares de variáveis (binárias, contínua, ordenadas bem como suas combinações) geradas com correlação $\rho = \{0, 0.5, 0.75, 0.9\}$ e observamos 200 replicações de tamanho $\boldsymbol{n} = \{5, 10, 20\}$. Utilizamos para estimação a prori conjugada Wishart com $\nu = 2$ graus de liberdade.

Métodos

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.

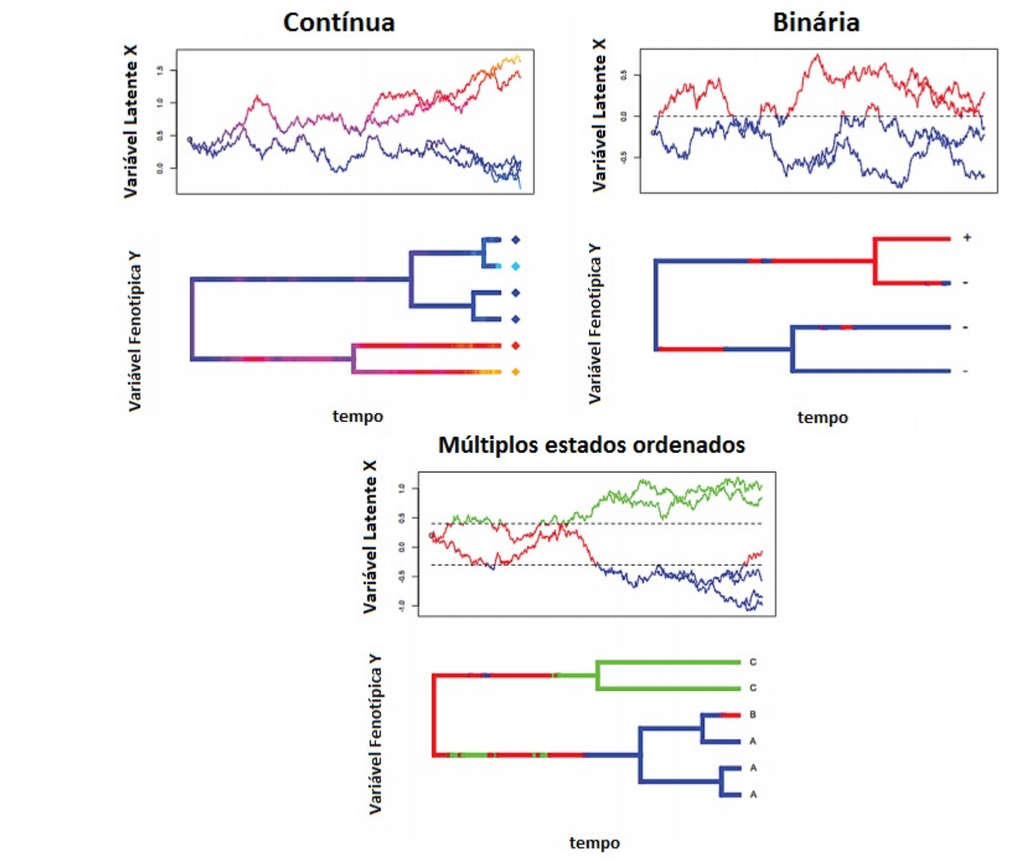

Figura 1: *Evolução da variável latente $\boldsymbol{X}$ através de movimento browniano, que determina os resultados da variável observável $\boldsymbol{Y}$*

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

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

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

Estudo de Simulação

Avaliamos o modelo de variável latente comparando suas estimativas de correlação às obtidas através do método frequentista de Pearson. Para ambos os métodos foram utilizados os mesmos conjuntos de dados em cada replicação.

As estimativas foram obtidas a partir de amostras de tamanho $\boldsymbol{n} = \{5, 10, 20\}$ e a inferência do modelo foi feita a partir de MCMC, utilizando uma filogênia fixa e priori pouco informativa. Foram feitas $Re = 200$ em cada senário e para validação dos resultados optou-se por um tamanho de amostra efetivo (ESS, *Effective Sample Size* da posteriori acima de 100 e das variáveis latentes acima de 70.

0.5cmPara fins de comparação utilizamos como estimador a média à posteriori e a médias das correlações pelo método de Pearson. Optamos também por utilizar como medida de significância a proporção de intervalos contendo 95% das estimativas obtidas em cada cadeia da replicação que não continham o valor 0 (zero).

Resultados Preliminares

Alguns resultados preliminares obtidos a partir da estimação da correlação entre uma variável contínua e uma variável binária, indicam que com o aumento do tamanho de amostra diminui a variabilidade das estimativas de correlação e as aproxima das verdadeiras.

Na tabela 1 a coluna *Média*, contém a média das correlações filogenéticas estimadas, seguida do desvio padrão das estimativas obtidas (*Sd*) e da proporção de replicações que tiveram $Sig > 0.975$ ou $Sig < 0.025$ (*Sig*). As demais colunas apresentam a média das estimativas de correlação obtidas pelo método de Pearson (*r*) e proporção de estimativas que teve $P - Valor > 0.975$ ou $P - Valor < 0.025$ (*P-valor*).

Tabela 1: *Comparação entre diferentes as estimativas de correlação, dados os diferentes tamanhos de amostra em diferentes cenários ($\rho = \{0, 0.5, 0.75, 0.9\}$)*

	Tamanho 5					Tamanho 10					Tamanho 20				
ρ	Média	Sd	Sig	r	P-Valor	Média	Sd	Sig	r	P-Valor	Média	Sd	Sig	r	P-Valor
0	-0.03	0.35	0.08	-0.07	0.04	-0.005	0.44	0.07	0.01	0.021	-0.04	0.35	0.01	-0.02	0.014
0.5	0.30	0.30	0.28	0.843	0.05	0.20	0.45	0.20	0.19	0.031	0.54	0.32	0.55	0.43	0.009
0.75	0.41	0.26	0.41	0.64	0.02	0.33	0.43	0.32	0.31	0.024	0.70	0.22	0.77	0.53	0.012
0.9	0.47	0.20	0.46	0.73	0.01	0.41	0.41	0.39	0.40	0.018	0.80	0.14	0.83	0.66	0.001

O modelo se destaca pela não estimação de correlações espúrias, tendo sempre um viés menor que o método de Pearson, quando as variáveis são geradas e um cenário de independência. Ainda o aumento do tamanho de amostra para $\boldsymbol{n} = 20$, já aparenta ser suficiente para que as estimativas de ambos os métodos se aproximem nos demais cenários.

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.