Análise de Correlações Evolutivas Estimadas pelo Modelo Filogenético de Variável Latente

Vitória Martini Wendt (BIC-UFRGS), Gabriela Bettella Cybis (Orientadora-UFRGS)
vitoriawendt@gmail.com, gabriela.cybis@ufrgs.br

paz no plural

Introdução

O estudo das interações entre genótipos e fenótipos é um dos principais focos da biologia evolutiva. Por se tratar de um problema complexo, ainda existem poucos métodos para estimar correlações entre fenótipos na evolução. Assim, o modelo de Variável Latente apresenta-se como uma opção para estas análises, já que pode ser utilizado para estimar estas correlações considerando **diferentes tipos de dados** - contínuos e discretos, binários ou múltiplos, ordenados ou não - enquanto controla para a **história evolutiva** dos indivíduos. Sua implementação é feita no software bayesiano para filogenias BEAST, já bastante difundido na biologia evolutiva.
O modelo filogenético de Variável Latente ainda pode ser considerado novidade na sua área, não existindo estudos específicos para a avaliação de suas propriedades estatísticas.

Objetivos

1. Avaliar as propriedades estatísticas da estimação de correlações evolutivas pelo modelo filogenético de Variável Latente.
2. Comparar com métodos clássicos utilizados para estimar correlações.
3. Comparar as propriedades estatísticas das correlações estimadas para diferentes tipos de dados.

Modelo Filogenético de Variável Latente

Seja $Y = (y_0, ..., y_N)$ um vetor cujas entradas são os valores observados para a variável de interesse nos N indivíduos da amostra, $X = (x_0, ..., x_N)$ um vetor cujas componentes representam as variáveis latentes do processo evolutivo e F uma filogenia (representação da história evolutiva dos N indivíduos). A variável X evolui ao longo de F por movimento browniano cuja verossimilhança, $P(X|F, \Sigma^{-1})$, pode ser calculada por sucessivas convoluções de normais multivariadas, em que Σ^{-1} é a matriz de precisão do modelo browniano, sendo utilizada como proxy para as correlações entre as componentes de Y. Ao final do movimento o valor da variável latente determinará o valor de Y, de modo que $Y = X$ quando Y contínuo. No caso em que Y é discreto, define-se o valor de Y de acordo com a posição de X em relação a um limiar. Deste modo a posteriori do modelo pode ser obtida como:

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

em que $P(Y|X)$ depende se Y é discreto ou contínuo.
Utilizamos MCMC para realizar inferência neste modelo.

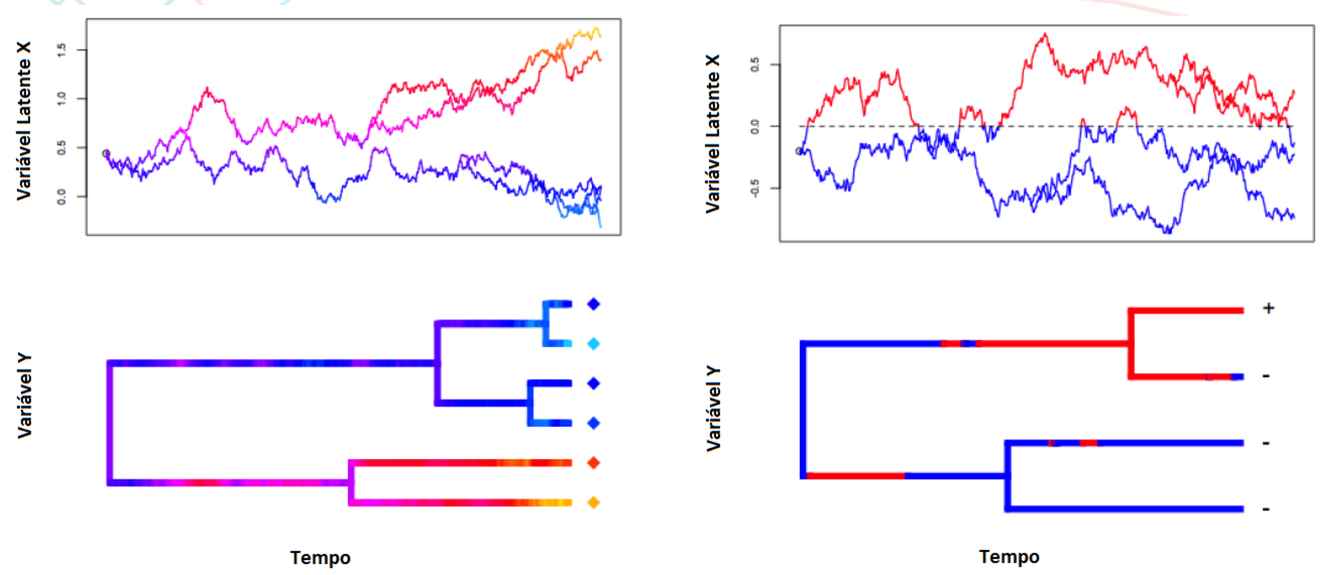

Imagem 1: Realização para Y contínua (esquerda); Realização para Y binária (direita).

Estudo por simulação

Para avaliar as propriedades estatísticas do modelo de Variável Latente, uma rotina de simulação com $re = 1000$ repetições foi construída. Foram considerados $N = 10$ indivíduos e $D = 4$ variáveis fenotípicas, das quais 2 contínuas e 2 discretas binárias.
Seja τ o comprimento da aresta que liga 2 nós na árvore F e, Σ a matriz de correlações entre as variáveis. O valor de X em cada nó da filogenia é simulado por $X_i|x_{i^*}, \tau_i \sim N(x_{i^*}, \tau_i\Sigma)$, em que x_{i^*} é o nó imediatamente ancestral a x_i. Uma função de ligação é utilizada para encontrar Y a partir de X. A matriz de correlações do modelo de variável latente é então obtida através da análise de Y por MCMC no software BEAST.

Resultados

Tabela 1: Propriedades dos estimadores de correlação no cenário de evolução independente.

Tipos de variáveis	Modelo de Variável Latente			Correlação de Pearson		
	Vício	$sd(\hat{r})$	Falso positivo	Vício	$sd(\hat{r})$	Falso positivo
Contínua × Contínua	0.0013	0.3711	0.1818	-0.0018	0.4768	0.1623
Contínua × Binária	0.0289	0.5019	0.1601	-0.0243	0.4252	0.1082
Binária × Binária	0.0011	0.4412	0.0616	0.0077	0.4190	0.0876

- As estimativas para a esperança das correlações são próximas de zero, indicando pouco ou nenhum vício.
- O desvio padrão das correlações quando tratadas duas variáveis contínuas é inferior pelo modelo.
- O teste de correlação do modelo retorna uma taxa de falso positivo para variáveis binárias inferior as demais, sendo inclusive inferior à taxa encontrada pela correlação de Pearson.

Tabela 2: Propriedades dos estimadores de correlação no cenário de evolução dependente.

Tipos de variáveis	r	Modelo de Variável Latente			Correlação de Pearson		
		Vício	$sd(\hat{r})$	Falso negativo	Vício	$sd(\hat{r})$	Falso negativo
Cont. × Cont.	-0.87	0.0493	0.1470	0.0453	0.0616	0.2188	0.1646
Cont. × Bin.	-0.75	0.0844	0.3250	0.4463	0.2038	0.2983	0.6336
Bin. × Bin.	0.87	-0.3094	0.3328	0.7064	-0.3413	0.2893	0.7708

- Os vícios apresentados indicam que para ambos estimadores as correlações simuladas são inferiores à r.
- O desvio padrão das correlações quando tratadas duas variáveis contínuas é inferior pelo modelo.
- Os falsos negativos encontrados pelo teste de correlação do modelo são inferiores àqueles encontrados através das correlações de Pearson.

Conclusão

A partir dos resultados listados ao lado é possível perceber informações importantes quanto a eficácia do modelo filogenético de Variável Latente:

Apresenta pouco ou nenhum vício em um cenário de independência entre as variáveis. Já em um cenário de dependência, apresenta vícios sempre inferiores quando comparado ao método de correlação de Pearson.

Os estimadores de correlações pelo modelo filogenético apresentam menor variância entre variáveis contínuas em ambos cenários de dependência e independência entre as variáveis.

Menor índice de falso positivo entre variáveis binárias pelo teste de correlação do modelo de Variável Latente, indicando certa vantagem na sua utilização para este tipo de análise.

Menor índice de falso negativo independentemente do tipo de variável, quando comparado a correlação de Pearson, indicando um maior poder do teste de correlação pelo modelo de Variável Latente.