SME0809 - Inferência Bayesiana - Prova 2

High-Dimensional Multivariate Bayesian Variable and Covariance Selection in Linear Regression (Zhao et al. 2021)

Grupo 13 - Francisco Miranda - 4402962 - Heitor Carvalho - 11833351

Dezembro 2021

Introdução

Com o desenvolvimento de técnicas de alto processamento na biologia molecular, a caracterização molecular em alta escala tornou-se um lugar comum, com o advento de técnicas como:

- genome-wide measurement of gene expression
- single nucleotide polymorphisms
- CpG methylation status
- pharmacological profiling for large-scale cancer drug screen.

A análise de associações conjuntas entre múltiplos fenótipos correlacionados e atributos moleculares de alta dimensionalidade é desafiadora.

Quando múltiplos fenótipos e informação genômica de alta dimensionalidade são analisados conjuntamente, a abordagem bayesiana permite especificar de maneira flexível as relações complexas entre os conjunto de dados altamente estruturados.

O pacote `BayesSUR` combina diversos modelos que foram propostos para a regressão multidimentional com resposta múltipla e introduz um novo modelo, que permite diferentes *prioris* na seleção de variáveis dos modelos de regressão e para diferentes pressupostos a respeito da estrutura de dependência entre as respostas.

Metodologia

- múltiplas opções de seleção de variáveis
- a matriz de covariância pode ser diagonal, densa ou esparsa.
- engloba três classes de modelos de regressão linear de múltipla resposta:
- HRR
- dSUR e SSUR
- MRF

O modelo de regressão é escrito como:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{U} \tag{1}$$

$$\text{vec}(\boldsymbol{U}) \sim \mathcal{N}(\boldsymbol{0}, C \otimes \mathbb{I}_n)$$

onde:

- $\boldsymbol{Y}$ é uma matriz $s \times s$ das variáveis resposta com matriz de covariância C;
- $\boldsymbol{X}$ é uma matriz $n \times p$ de preditores para todas as respostas;
- $\boldsymbol{U}$ é a matriz dos resíduos;
- vec$(\cdot)$ denota a vetorização da matriz;
- $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denota uma distribuição normal multivariada com vetor de médias $\boldsymbol{\mu}$ e matriz de covariâncias $\boldsymbol{\Sigma}$;
- $\mathbf{0}$ denota um vetor coluna com todos os elementos nulos,
- $\otimes$ é o produto de Kronecker e $\mathbb{I}_n$ a matriz identidade de ordem n.

A seleção de variáveis é realizada através de uma matriz indicadora binária latente $\boldsymbol{\Gamma} = \{\gamma_{jk}\}$.

Uma *priori* "spike-and-slab" é utilizada para encontrar um subconjunto esparso relevante de preditores que expliquem a variabilidade de $\boldsymbol{Y}$: condicional em $\gamma_{jk} = 0 \quad (j = 1, ..., p, \text{e } k = 1, ..., s)$

Definem-se $\beta_{jk} = 0$ condicionado em $\gamma_{jk} = 1$ seguem uma distribuição normal difusa:

$$\beta_\gamma \sim \mathcal{N}(\mathbf{0}, W_\gamma^{-1}) \tag{2}$$

Onde $\beta = \text{vec}(\mathbf{B})$, $\gamma = \text{vec}(\boldsymbol{\Gamma})$, β_γ consiste somente nos coeficientes selecionados (i.e. $\gamma_{jk} = 1$), assim W_γ é a sub matriz de W formada pelos coeficientes selecionados correspondentes.

A matriz de precisão, W, é geralmente decomposta em coeficientes de encolhimento e uma matriz que governa a estrutura de covariância dos coeficientes de regressão. É utilizado aqui $W = w^{-1}\mathbb{I}_{sp}$, o que significa que todos os coeficientes de regressão são independentes a priori, com uma *hiperpriori* no coeficiente de encolhimento w, i.e. $w \sim \mathcal{IG}\text{amma}(a_w, b_w)$.

A matriz indicadora binária latente $\boldsymbol{\Gamma}$ tem três opções de *priori*:

- Bernoulli independente hierárquica
- hotspot prior
- MRF prior

A matriz de covariância C também possui três *prioris*:

- Gama inversa independente
- Wishart inversa
- hiper-inversa Wishart

São considerados no total nove possíveis modelos dentre as combinações de C e $\boldsymbol{\Gamma}$

	$\gamma_{jk} \sim Bernoulli$	$\gamma_{jk} \sim$ hotspot	$\gamma_{jk} \sim$ MRF
$C \sim indep$	HRR-B	HRR-H	HRR-M
$C \sim IW$	dSUR-B	dSUR-H	dSUR-M
$C \sim HIW$	SSUR-B	SSUR-H	SSUR-M

Regressão Hierárquica Relacionada (HRR)

A Regressão Hierárquica Relacionada assume que C é uma matriz diagonal, o que se traduz em independência condicional entre múltiplas variáveis resposta.

Uma *priori* gama inversa é especificada para a covariância dos resíduos, i.e

$$\sigma_k^2 \sim \mathcal{IG}\text{amma}(a_\sigma, b_\sigma)$$

Quando combinada com as *prioris* em (2), é conjulgado com a verossimilhança do modelo (1). Podemos então amostrar a estrutura de seleção de variáveis $\boldsymbol{\Gamma}$ marginalmente com respeito a C e $\boldsymbol{B}$.

HRR com uma *priori* Bernouli independente

Para uma *priori* simples de seleção do modelo de regressão, os indicadores binários latentes seguem uma *priori* de Bernoulli:

$$\gamma_{jk}|\omega_{jk} \sim \mathcal{B}er(\omega_{jk}) \quad (j = 1, ..., p, \text{e } k = 1, ..., s) \tag{3}$$

Com uma priori hierárquica Beta em ω_j, i.e. $\omega_j \sim \mathcal{B}eta(a_\omega, b_\omega)$, que quantifica a probabilidade de cada preditor ser associado com qualquer uma das variáveis resposta.

HRR com uma *priori* hotspot

É proposta a decomposição da probabilidade do parâmetro de associação ω_{jk} em (3), onde o_k é responsável pela esparsividade de cada modelo de resposta e π_j controla a propensão de cada preditor a ser associado a múltiplas respostas simuntaneamente:

$$\begin{gathered}\gamma_{jk}|\omega_{jk} \sim \mathcal{B}er(\omega_{jk}) \quad (j = 1, ..., p, \text{e } k = 1, ..., s) \\ \omega_{jk} = o_k \times \pi_j \\ o_k \sim \mathcal{B}eta(a_0, b_0) \\ \pi_j \sim \mathcal{G}amma(a_\pi, b_\pi)\end{gathered} \tag{4}$$

Regressão não relacionada aparentemente esparsa (SSUR)

Para modelar a matriz de covariância C é especificado uma *priori* hiper-Inversa Wishart, o que significa que as variáveis resposta têm por trás um grafo $\mathcal{G}$ que codifica a dependência condicional entre as respostas.

Um grafo esparso corresponde à matriz esparsa de precisão C^{-1}. Do ponto de vista computacional, é impraticável especificar uma priori hiper-inversa Wishart diretamente em C^{-1}. É realizada uma transformação em C para fatorar a verossimilhança. A distribuição hiper inversa de Wishart i.e $C \sim \mathcal{HIW}_\mathcal{G}(\nu, \tau\mathbb{I}_s)$ transforma-se na variância escalar σ^2_{qt} e no vetor de correlação associado $\boldsymbol{\rho}_{qt} = (\rho_{1,qt}, \rho_{2,qt}, ..., \rho_{t-1,qt})^T$, com:

Amostragem MCMC e inferência *a posteriori*

Referências

Zhao, Zhi, Marco Banterle, Leonardo Bottolo, Sylvia Richardson, Alex Lewin, and Manuela Zucknick. 2021. "BayesSUR: An R Package for High-Dimensional Multivariate Bayesian Variable and Covariance Selection in Linear Regression." *Journal of Statistical Software* 100 (11): 1–32. https://doi.org/10.18637/jss.v100.i11.

Apêndice: códigos

```
knitr::opts_chunk$set(echo = TRUE)
```