SME0809 - Inferência Bayesiana - Prova 2 - Grupo 13

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

Francisco Miranda - 4402962 Heitor Carvalho - 11833351

Dezembro 2021

Contents

1 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:

- Medida de expressão Gênica
- Poliomorfismos de Nucleotídeo Único (SNP)
- Status de Metilação CpG
- Perfil farmacológico para testes em larga escala.

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.

2 Metodologia

O modelo de regressão é escrito como:

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

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

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.

2.1 Seleção de Variáveis

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

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

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*, assim como a matriz de covariância. São considerados no total nove possíveis modelos dentre as combinações de C e $\boldsymbol{\Gamma}$, exibidos na Tabela 1.

Table 1: Modelos disponibilizados pelo pacote `BayesSUR`

	$\gamma_{jk} \sim Bernoulli$	$\gamma_{jk} \sim$ 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

2.2 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 $\mathbf{\Gamma}$ marginalmente com respeito a C e $\boldsymbol{B}$.

2.2.1 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.

2.2.2 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:

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

2.2.3 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:

$$\sigma^2_{qt} \sim \mathcal{IG}amma\left(\frac{\nu - s + t + |S_q|}{2}, \frac{\tau}{2}\right),\ q = 1, ..., Q,\ t = 1, ..., |R_q|, \boldsymbol{\rho}_{qt}|\sigma^2_{qt}\ \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2_{qt}}{\tau}\mathbb{I}_{t-1}\right) \tag{5}$$

onde Q é o número de componentes primos no grafo decomposto $\mathcal{G}$, S_q e R_q são os separadores e os componentes residuais de $\mathcal{G}$, respectivamente.

Como *priori* para o grafo é utilizado uma Bernoulli com probabilidade η em cada vértice $E_{kk'}$ de $\mathcal{G}$ como em:

$$\mathbb{P}(E_{kk'} \in \mathcal{G}) = \eta, \quad \eta \sim \mathcal{B}eta(a_\eta, b_\eta). \tag{6}$$

São admitidas três *prioris* em β_γ.

2.3 Amostragem MCMC e inferência *a posteriori*

Para amostrar da distribuição a *posteriori*, os autores utilizam o algoritmo de busca estocástica evolucionária, que utiliza uma forma particular do Monte Carlo evolucionário (EMC).

Múltiplas cadeias de Markov temperadas são processadas paralelamente e movimentos de troca ou mudança são permitidos dentre as cadeias para melhorar a mistura entre modelos potencialmente diferentes da *posteriori*. A temperatura é adaptada durante a fase de burn-in.

A cadeia principal provém amostras da distribuição a *posteriori* não-temperada, que é utilizada para toda a inferência. Para cada variável resposta, os autores utilizam um amostrador de Gibbs para atualizar o vetor dos coeficientes de regressão $\beta_k (k = 1, ..., s)$, baseado na distribuição a *posteriori* condicional correspondente ao modelo específico, selecionado entre os modelos apresentados anteriormente.

Após L iterações do MCMC, obtêm-se $\boldsymbol{B}^{(1)}, ..., \boldsymbol{B}^{(L)}$ e a estimativa da média a *posteriori* é:

$$\hat{\boldsymbol{B}} = \frac{1}{L - b}\sum_{t=b+1}^{L} \boldsymbol{B}^{(t)}$$

onde b é o número de iterações de *burn-in*. As distribuições condicionais completas a *posteriori* também estão disponíveis no modelo SSUR. Já nos modelos HRR, os coeficientes de regressão e as covariâncias residuais foram integrados para fora e ainda assim a saída do MCMC não pode ser utilizada diretamente para inferencia posterior desses parâmetros.

Contudo, para $\boldsymbol{B}$, a distribuição *posteriori* condicional em $\boldsymbol{\Gamma}$ pode ser obtida analiticamente nos modelos HRR, e é essa a saída oferecida.

Em cada iteração t do MCMC também é atualizado cada vetor binário latente $\gamma_k (k = 1, ..., s)$ via Metropolis-Hastings, propondo conjuntamente uma atualização para o correspondente β_k. Após L iterações, usando as matrizes binárias $\boldsymbol{\Gamma}^{(1)}, ..., \boldsymbol{\Gamma}^{(L)}$, as probabilidades de inclusão marginal a *posteriori* são estimadas por:

$$\hat{\boldsymbol{\Gamma}} = \frac{1}{L - b}\sum_{t=b+1}^{L} \boldsymbol{\Gamma}^{(t)}$$

Outro parâmetro importante dos modelos SSUR é $\mathcal{G}$ na *priori* Wisahrt hiper-inversa para a matriz de covariância C. Ela é atualizada via *junction tree sampler* conjuntamente com a proposta correspondente para σ^2_{qt} e $\boldsymbol{\rho}_{qt}|\sigma^2_{qt}$ em (5).

A cada iteração do MCMC é extraída a matriz de adjacência $\mathcal{G}^{(t)}(t = 1, ..., L)$, do qual são derivadas as estimativas da média a *posteriori* das probabilidades de inclusão das areastas como:

$$\hat{\mathcal{G}} = \frac{1}{L-b} \sum_{t=b+1}^{L} \mathcal{G}^{(t)}$$

Mesmo que a *priori* o grafo $\mathcal{G}$ seja decomposto, a média estimada posteriormente $\hat{\mathcal{G}}$ pode estar pode do espaço de modelos decompostos.

O hiperparâmetro τ da Wishart hiper-inversa é atualizado através de um passeio aleatório do amostrador Metropolis-Hastings. Já η e a variância w na priori pico-e-tapa são amostrados das condicionais posteriores.

3 Conjunto de Dados

Os autores simularam dados de polimorfismo de nucleotídeo único (SNP) dentro de um modelo verdadeiro conhecido para demonstrar a performance de recuperação dos modelos introduzidos anteriormente. O algoritmo completo pode ser encontrado em (Zhao et al. 2021).

Para construir variáveis resposta múltiplas $\boldsymbol{Y}$ (com $s = 10$) com uma relação estruturada, os autores fixam uma variável indicadora esparsa $\boldsymbol{\Gamma}$ e desenham um grafo decomposto para as respostas, para construir padrões de associação dentre os múltiplos regressores e variáveis resposta.

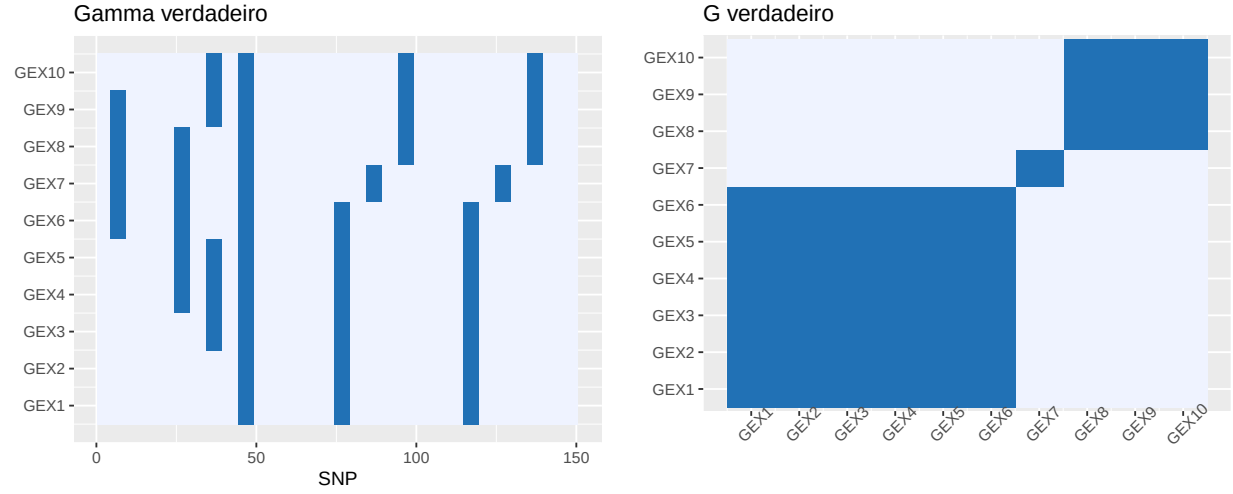

Figure 1: Parâmetros verdadeiros dos dados gerados no conjunto de dados de exemplo. Valores em branco representam 0 e, em azul, 1

As respostas em $\boldsymbol{Y}$ são representadas por uma matriz 10×10, já os preditores $\boldsymbol{X}$ são representados por uma matriz com 6 linhas, que é o tamanho de nossa amostra, e 150 colunas, representando os SNPs. Temos também uma `blockList` que especifica os índices de $\boldsymbol{Y}$ e $\boldsymbol{X}$ em `data`.

O terceiro componente é a verdadeira matriz indicadora $\boldsymbol{\Gamma}$ dos coeficientes de regressão. O quarto componente é o grafo verdadeiro $\mathcal{G}$ entre as variáveis resposta. A Figura 1 mostra os verdadeiros $\boldsymbol{\Gamma}$ e $\mathcal{G}$ utilizados para simular ao exemplo. As matrizes de associação são exibidas na forma de mapas de calor.

4 Análise dos Dados

Os autores utilizam o exemplo para ajustar um modelo SSUR com uma *priori hotspot* para as variáveis indicadoras $\mathbf{\Gamma}$ e a *priori* indutora de esparsidão Wishart hiper-inversa utilizando o pacote `BayesSUR`. No exemplo são 200000 iterações do MCMC, com burn-in de 100000, em três cadeias paralelamente. Para manter a reprodutibilidade, o código foi executado em um único núcleo.

As Figuras 2 e 3 resumem os resultados da inferência a posteriori dos estimadores $\hat{\boldsymbol{B}}, \hat{\mathbf{\Gamma}}$, e $\hat{\mathcal{G}}$.

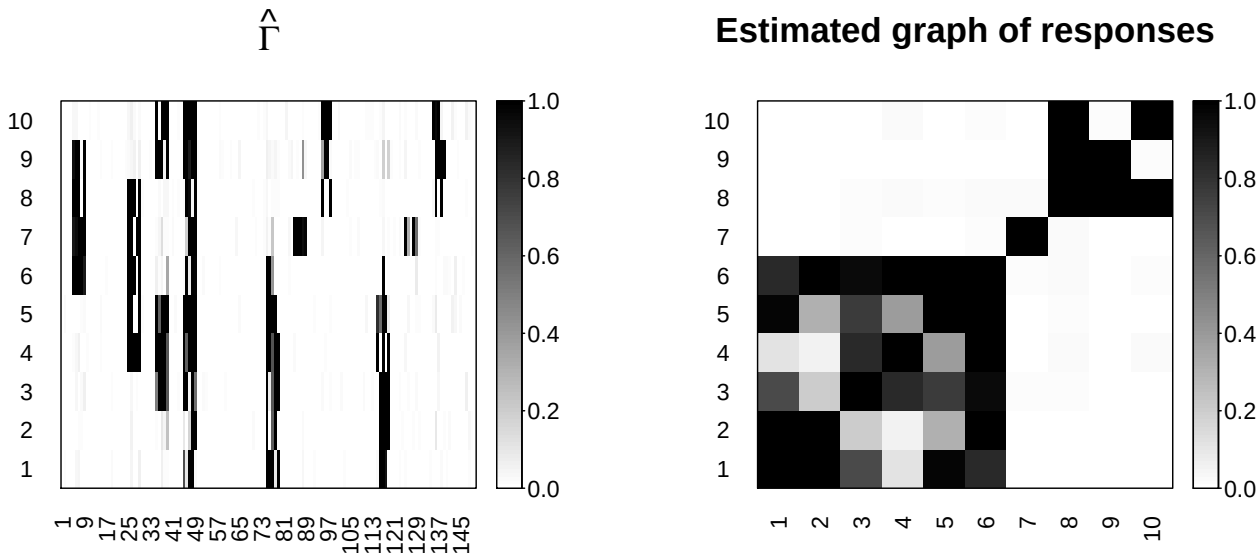

Figure 2: Coeficientes de seleção de variáveis estimados da matriz indicadora latente $\hat{\mathbf{\Gamma}}$ (esquerda) e Esturutura aprendida de $\hat{\mathcal{G}}$ pelo modelo SSUR com priori hotspot e covariância esparsa (direita)

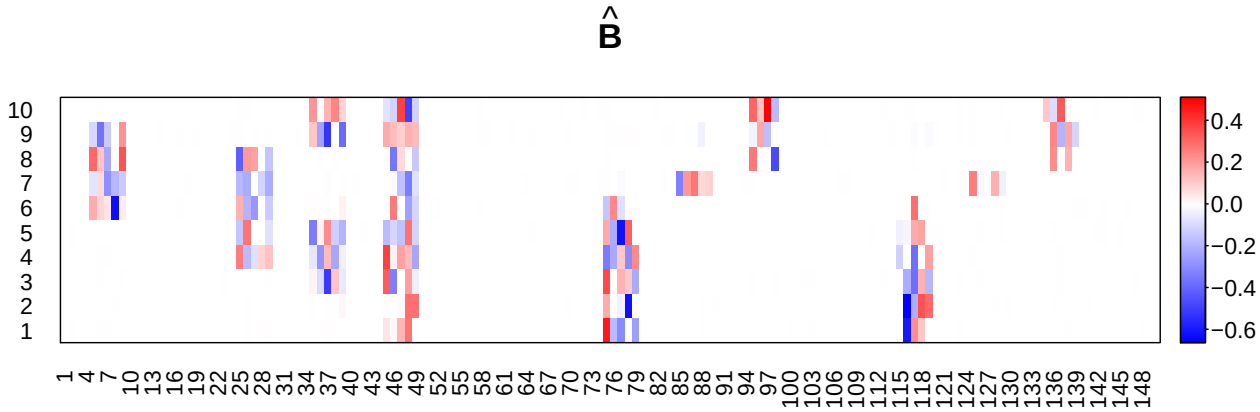

Figure 3: Coeficientes de regressão linear de cada SNP estimados da matriz $\hat{\boldsymbol{B}}$

Vemos que o modelo SSUR possui uma boa recuperação do verdadeira matriz indicadora latente $\mathbf{\Gamma}$ e da estrutura das respostas representada por $\mathcal{G}$. Podemos comparar a verdadeira estrutura com a estimada quando limitamos a probabilidade de seleção a *posteriori* para $\mathbf{\Gamma}$ e $\mathcal{G}$ a 0.5, criando o grafo exibido à direita na Figura 1.

Os plots estilo Manhattan na Figura 5 mostram as probabilidades de inclusão marginal (mPIP) dos SNPs (painel superior) e o número de expressões gênicas da resposta associado com cada SNP (painel inferior). O número de respostas é baseado em $\hat{\mathbf{\Gamma}}$ limitado a 0.5.

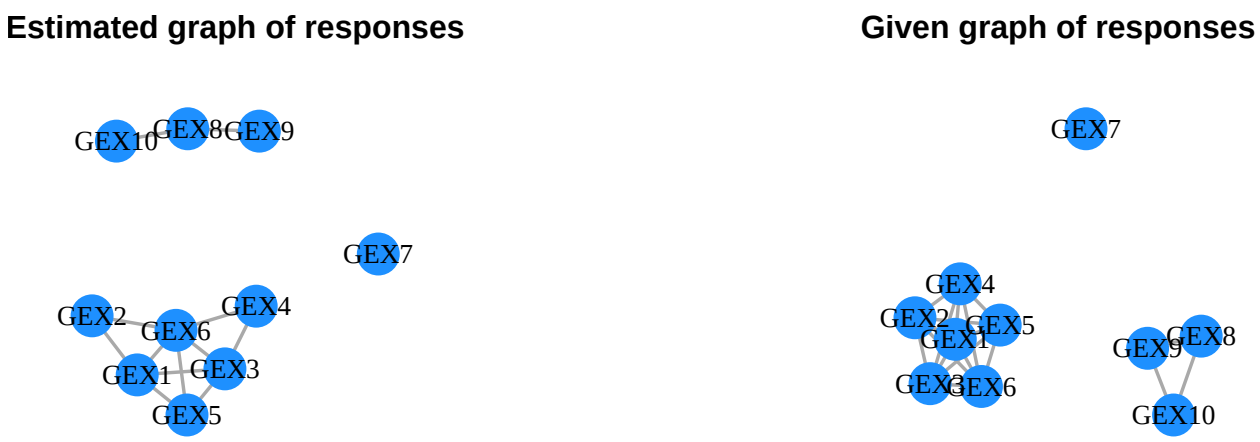

Figure 4: Estrutura estimada das dez variáveis resposta com $\hat{\mathcal{G}}$ limitado a 0.5 (esquerda). Estrutura verdadeira de $\mathcal{G}$ representada pela matriz de adjacência `Gy` (direita).

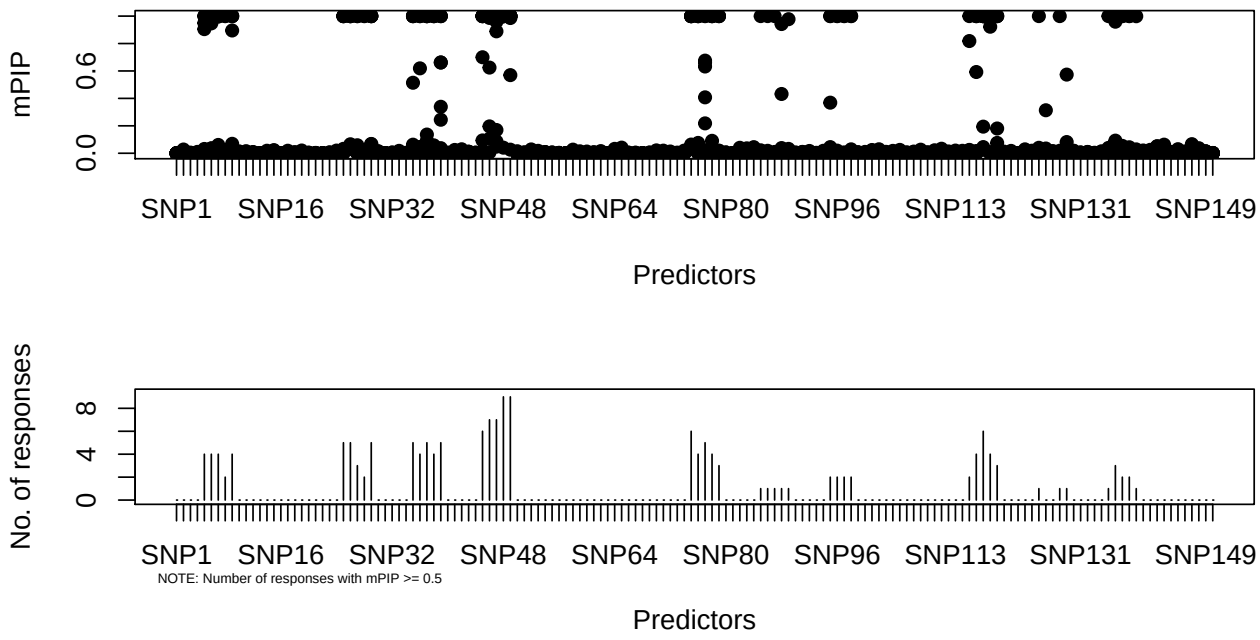

Figure 5: Plots estilo Manhattan, mostrando as probabilidades de inclusão marginal posterior (acima) e o número de expressões gênicas da resposta associada a cada SNP (abaixo).

Para investigarmos o comportamento do nosso amostrador MCMC, podemos utilizar os plots de diagnóstico apresentados na Figura 6. Observa-se que as cadeias de Markov utilizadas aparentemente começam a amostrar da distribuição correta após aproximadamente 50.000 iterações.

Os painéis inferiores da Figura 6 indicam que o logarítmo da distribuição a posteriori da variável indicadora latente $\mathbf{\Gamma}$ é estável para a última metade das cadeias, após subtraído o burn-in.

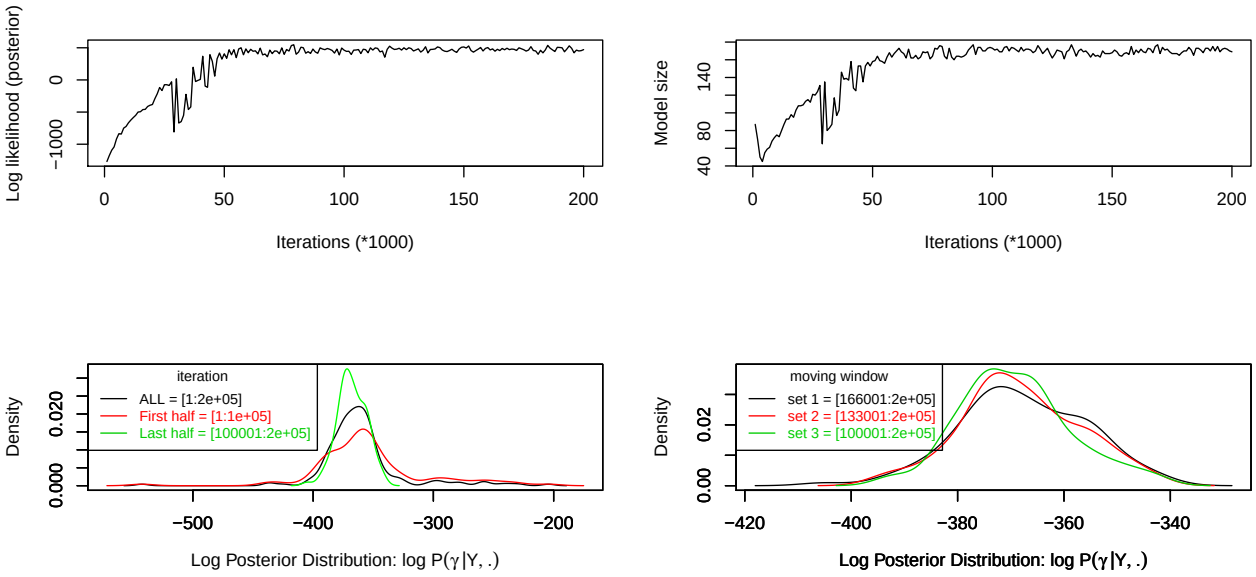

Figure 6: Plots de diagnóstico do amostrador MCMC.

Embora não tenhamos acesso direto às cadeias simuladas, podemos também visualizar a verossimilhança das outras matrizes estimadas pelo modelo conforme mostra a Figura 7. Aparentemente, não há nenhuma mudança brusca de padrão conforme o aumento de iterações.

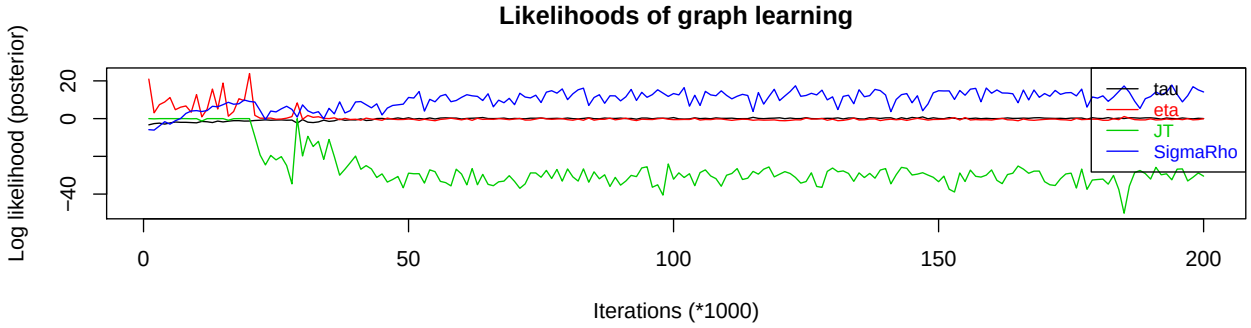

Figure 7: Verossimilhança dos estimadores pelo número de iterações do MCMC.

É também possível utilizar o CPO para encontrar observações destoantes, como uma forma de validação cruzada do modelo com a amostra obtida, conforme mostra a Figura 8. Temos somente duas observações abaixo do limiar.

Utilizando o mesmo limiar, de $\hat{\mathcal{G}}$ limitado a 0.5, e $\hat{\mathbf{\Gamma}}$ limitado a 0.5, apresentamos a rede completa com as dez expressões gênicas e os 150 SNPs na Figura 9.

Finalizamos nossa análise com um sumário do modelo obtido. Após isso, desatachamos o conjunto de dados utilizado.

##

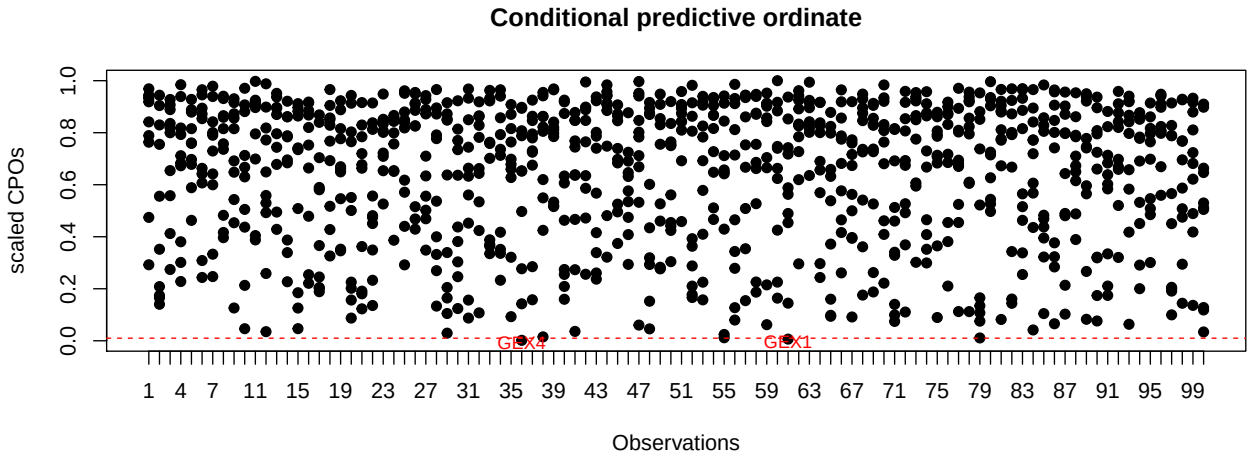

Figure 8: CPOs escalonados para o modelo de regressão ajustado

```
## Call:
##   BayesSUR(data = data, Y = blockList[[1]], X = blockList[[2]], ...)
##
## CPOs:
##         Min.      1st Qu.       Median      3rd Qu.         Max.
## 0.0006981173 0.1984357123 0.3125505299 0.3715488386 0.4266639028
##
## Number of selected predictors (mPIP > 0.5): 165 of 10x150
##
## Top 10 predictors on average mPIP across all responses:
##     SNP48     SNP49     SNP47     SNP46    SNP117     SNP75     SNP45     SNP37
## 0.9040660 0.8577530 0.7152859 0.6919080 0.6294290 0.6180328 0.5822639 0.5345648
##     SNP26     SNP29
## 0.5224508 0.5169448
##
## Top 10 responses on average mPIP across all predictors:
##       GEX4       GEX9       GEX7       GEX5      GEX10       GEX6       GEX3
## 0.14889411 0.14059070 0.13917972 0.13790778 0.11259638 0.10842810 0.10746704
##       GEX8       GEX1       GEX2
## 0.10278926 0.06938224 0.06500430
##
## Expected log pointwise predictive density (elpd) estimates:
##   elpd.LOO = -1435.736,  elpd.WAIC = -1439.831
##
## MCMC specification:
##   iterations = 2e+05,  burn-in = 1e+05,  chains = 3
##   gamma local move sampler: bandit
##   gamma initialisation: R
##
## Model specification:
##   covariance prior: HIW
##   gamma prior: hotspot
##
## Hyper-parameters:
##   a_w   b_w   a_o   b_o  a_pi  b_pi    nu a_tau b_tau a_eta b_eta
##   2.0   5.0   2.0 148.0   2.0   1.0  12.0   0.1  10.0   0.1   1.0
```

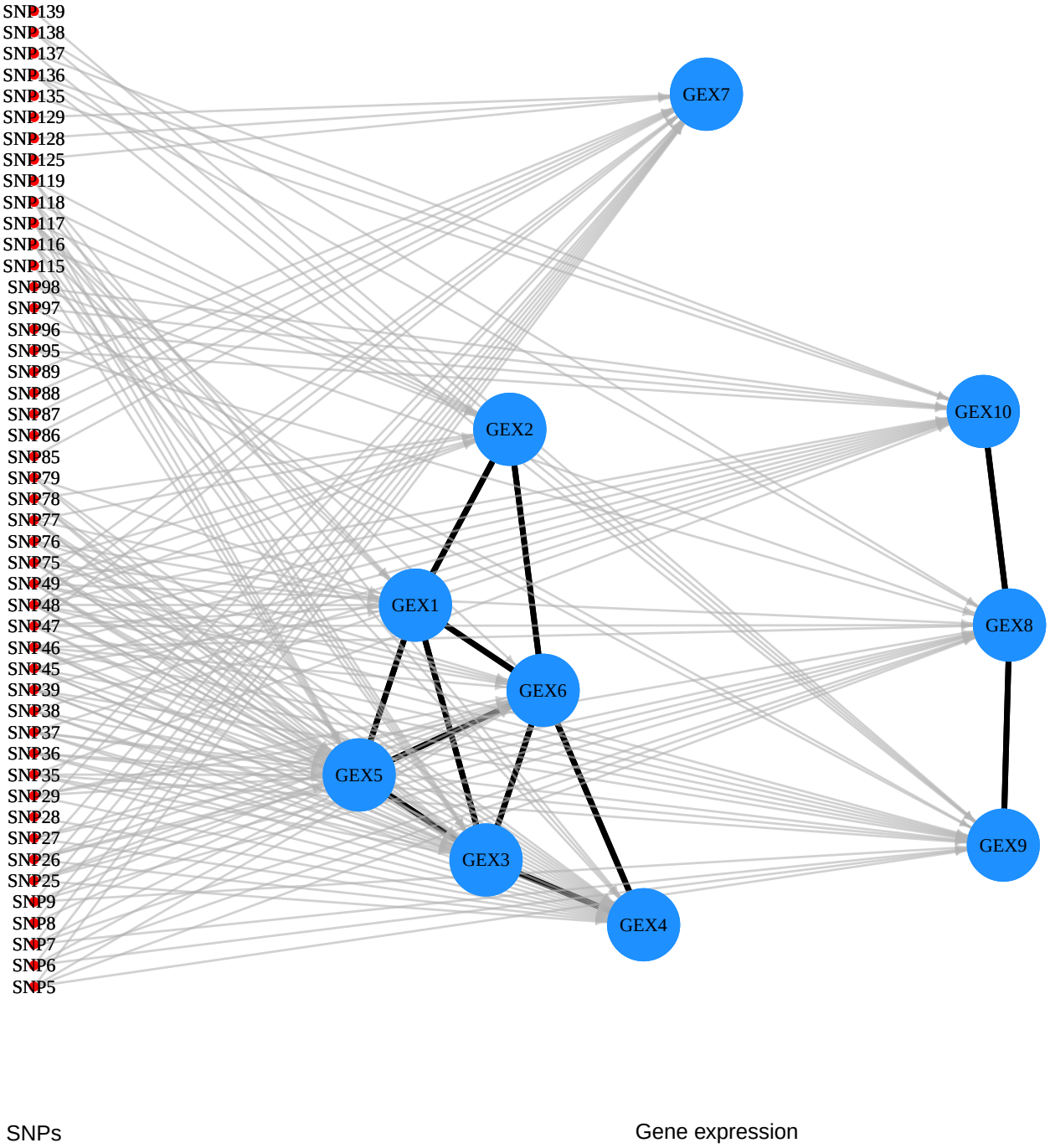

Figure 9: Representação da rede entre as expressões gênicas ($\boldsymbol{Y}$), com $\hat{\mathcal{G}}$ limitado a 0.5, e os SNPs ($\boldsymbol{X}$), com $\hat{\boldsymbol{\Gamma}}$ limitado a 0.5.

5 Conclusão

Neste relatório, procuramos apresentar os modelos para regressão e regularização partindo de uma abordagem Bayesiana, utilizamos o bacote `BayesSUR` para efetuar a regressão e seleção de modelos no conjunto de teste de exemplo apresentado por Zhao et al. (2021).

Foram introduzidos brevemente os aspectos da modelagem da seleção de variáveis, assim como a recuperação das estruturas para identificar relacionamentos entre respostas multivariadas em preditores de alta dimensionalidade.

Como ponto negativo, não foi possível realizar a análise convencional das cadeias conforme esperávamos, contudo dada a complexidade do modelo tal ausência é justificada pela estrutura matricial dos preditores e resposta, que se mostrou um grande desafio em nossa análise.

Entretanto, foi uma oportunidade muito valiosa de conhecer modelos estado-da-arte na área de biologia molecular, implementados de forma eficiente e bem documentada pelos autores do artigo que aqui apresentamos.

6 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.

A Apêndice: códigos

```
# pacote principal
library(BayesSUR)
data("exampleEQTL", package = "BayesSUR")
# attach nos dados para um código mais compacto
attach(exampleEQTL)
# bibliotecas para gráficos
library(tidyverse)
library(gridExtra)
library(ggpubr)
library(tictoc)
# funcao que recebe uma matriz e plota um mapa de calor
 plot_heatmap<- function(df){
reshape2::melt(df) %>% ggplot(aes(x=Var1, y=Var2, fill=-value)) +
    geom_raster() + guides(fill="none") +
    scale_fill_fermenter() + xlab(" ") + ylab(" ")
 }
# rotulos da escala
labs <- as_labeller(
          c("1" = "GEX1", "2" =  "GEX2",  "3" =  "GEX3", "4" = "GEX4","5" = "GEX5",
          "6" = "GEX6", "7" = "GEX7", "8" = "GEX8", "9" = "GEX9", "10" = "GEX10"))
# mapa de calor Y versus X
p <- plot_heatmap(gamma) + scale_y_continuous(breaks= 1:10,labels = labs) +
  xlab("SNP") + ggtitle("Gamma verdadeiro")
# mapa de calor Y versus Y
q <- plot_heatmap(Gy)  + scale_x_continuous(breaks= 1:10,labels = labs) +
```

```
  theme(axis.text.x = element_text(angle = 45)) + ggtitle("G verdadeiro")

grid.arrange(p,q, ncol = 2)
set.seed(28173)
tic("Tempo de ajuste do modelo")
# ajuste do modelo
fit <- BayesSUR(data = data, Y = blockList[[1]], X = blockList[[2]],
                outFilePath = "results", nIter = 5000, nChains = 3,
                burnin = 1000, covariancePrior = "HIW",
                gammaPrior = "hotspot",
                output_CPO = TRUE,
                )
toc()
# ler o modelo ja ajustado
load("results-hiw-hotspot/obj_BayesSUR.RData")
fit <- obj_BayesSUR
class(fit) <- "BayesSUR"
# plota estimador gamma
plotEstimator(fit, c("gamma", "Gy"))
# plota estimador beta
plotEstimator(fit, "beta")
# estrutura estimada vs estrutura verdadeira
layout(matrix(1:2, ncol = 2))
plot(fit, estimator = "Gy", type = "graph")
plotGraph(Gy)
plot(fit, estimator = "gamma", type = "Manhattan")
plot(fit, estimator = "logP", type = "diagnostics")
plotMCMCdiag(fit, HIWg = "lik")
# CPO
plotCPO(fit)
# estrutura entre X e Y
plot(fit, estimator = c("gamma", "Gy"), type = "network",
     name.predictors = "SNPs", name.responses = "Gene expression")
# sumario do modelo
summary(fit)
detach(exampleEQTL)
```