Equações de estimação regularizadas

Vinicius Ricardo Riffel
Orientador: Prof. Dr. Wagner Hugo Bonat

Universidade Federal do Paraná
Departamento de Estatística
Laboratório de Estatística e Geoinformação
viniciusriffel@ufpr.br

Set. 14, 2022

Sumário

Sobre a apresentação

- TCC sob orientação do Prof. Dr. Wagner Hugo Bonat.
- Objetivo: propor um algoirtmo de estimação baseado em equações de estimação regularizadas.
- Início em $\approx$ dez/2021 e ainda não foi finalizado.
- Provisoriamente, a implementação está hospeada em `https://github.com/vriffel/mcglm`

Sumário

Introdução

- Modelos de regressão são técnicas que permitem descrever problemas do mundo real de forma probabilística.
- Suposições são feitas na aplicação de tais técnicas.
- Há diferentes modelos disponíveis:
  - Modelo linear (LM)
  - Modelo linear generalizado (GLM)
  - Equações de estimação generalizadas (GEE)
  - Modelos multivariados de covariância linear generalizados (McGLM)

LM

- Especificação de um LM:

$$E(\mathrm{Y}) = \boldsymbol{\mu} = \mathrm{X}\boldsymbol{\beta}$$

$$Var(\mathrm{Y}) = \tau I$$

- Suposições: normalidade, independência, variância constante.

GLM

- Especificação de um GLM:

$$E(\mathrm{Y}) = g^{-1}(\boldsymbol{\mu}) = g^{-1}(\mathrm{X}\boldsymbol{\beta})$$

$$Var(\mathrm{Y}) = \tau V(\boldsymbol{\mu})$$

- Suposições: família exponencial natural e independência.
- Ao utilizar quase-verossimilhança, não haverá a limitação da distribuição assumida.

GEE

- Especificação:

$$E(\mathrm{Y}) = g^{-1}(\boldsymbol{\mu}) = g^{-1}(\mathrm{X}\boldsymbol{\beta})$$

$$Var(\mathrm{Y}) = \tau V(\boldsymbol{\mu})^{\frac{1}{2}} \boldsymbol{\Omega}(\boldsymbol{\rho}) V(\boldsymbol{\mu})^{\frac{1}{2}}$$

- Limitações: dados multivariados e estrutura de covariância.
- Estimado via equações de estimação.

McGLM (Bonat e Jørgensen, 2016).

- Especificação:

$$E\left(\mathrm{Y}\right) = \left(g_1^{-1}\left(\mathrm{X}_1\beta_1\right), g_2^{-1}\left(\mathrm{X}_2\beta_2\right), \ldots, g_R^{-1}\left(\mathrm{X}_R\beta_R\right)\right)$$

$$Var\left(\mathrm{Y}\right) = \Sigma_R \overset{G}{\otimes} \Sigma_b$$

$$\Sigma_R \overset{G}{\otimes} \Sigma_b = \mathrm{Bdiag}\left(\tilde{\Sigma}_1, \tilde{\Sigma}_2, \ldots, \tilde{\Sigma}_R\right)\left(\Sigma_b \otimes \mathrm{I}\right)\mathrm{Bdiag}\left(\tilde{\Sigma}_1^T, \tilde{\Sigma}_2^T, \ldots, \tilde{\Sigma}_R^T\right)$$

$$\Sigma_r = V(\mu_r; p_r)^{\frac{1}{2}}(\Omega(\tau_r))V(\mu_r; p_r)^{\frac{1}{2}}$$

$$h\left(\Omega(\tau_r)\right) = \tau_{r0}Z_{r0} + \tau_{r1}Z_{r1} + \ldots + \tau_{rD}Z_{rD}$$

Sumário

Equações de estimação

- As equações de estimação são métodos de estimação que se destacam por sua generalidade e propriedades.
- Diversos métodos de estimação são casos particulares:
  - mínimos quadrados;
  - método da máxima verossimilhança;
  - quase-verossimilhança.

Definição equação de estimação

Uma equação de estimação de estimação é definida como:

$$g(\mathrm{Y}; \boldsymbol{\beta}) = 0$$

Estimando um McGLM

- Os McGLM são estimados utilizando o método *modified chaser* (Jørgensen e Knudsen, 2004).

Funções de estimação de um McGLM

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \boldsymbol{D}^T \mathsf{C}^{-1} (\mathcal{Y} - \mathcal{M})$$

$$\psi_{\boldsymbol{\lambda}_i}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathrm{tr}\left(W_{\boldsymbol{\lambda}_i} (\mathsf{r}^T \mathsf{r}) - \mathsf{C} \right)$$

Algoritmo de estimação de um McGLM

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - S_{\boldsymbol{\beta}}^{-1} \psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)})$$

$$\boldsymbol{\lambda}^{(i+1)} = \boldsymbol{\lambda}^{(i)} - S_{\boldsymbol{\lambda}}^{-1} \psi_{\boldsymbol{\lambda}_i}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)})$$

Equações de estimação regularizadas

- A regularização é uma restrição imposta no processo de estimação dos modelos.
- Torna possível lidar com *high dimensional data*, diminuir o erro preditivo e fazer seleção de covariáveis.
- Um exemplo de regularização em equações de estimação é (Fu, 2003):

$$g^*(\mathrm{Y};\boldsymbol{\beta}) = g(\mathrm{Y};\boldsymbol{\beta}) - \gamma \boldsymbol{p}'(|\beta|) = g(\mathrm{Y};\boldsymbol{\beta}) - \gamma \boldsymbol{p}'(\beta)\frac{\beta}{|\beta|}$$

- γ pode ser escolhido via validação cruzada.

Estimando um McGLM regularizado

Equações de estimação do McGLM regularizado

$$\psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \boldsymbol{D}^T \mathsf{C}^{-1} (\mathcal{Y} - \mathcal{M}) - \boldsymbol{\gamma} \odot \Gamma(p_1'(|\boldsymbol{\beta}|))$$

$$\psi^*_{\boldsymbol{\lambda}_i}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathrm{tr}\left(W_{\boldsymbol{\lambda}_i}(\mathsf{r}^T\mathsf{r}) - \mathsf{C}\right) - \boldsymbol{\gamma}_i \odot \Gamma(p_2'(|\boldsymbol{\lambda}_i|))$$

Algoritmo de estimação

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - S^{*-1}_{\boldsymbol{\beta}} \psi^*_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)})$$

$$\boldsymbol{\lambda}^{(i+1)} = \boldsymbol{\lambda}^{(i)} - S^{*-1}_{\boldsymbol{\lambda}} \psi^*_{\boldsymbol{\lambda}_i}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)})$$

Por que regularizar?

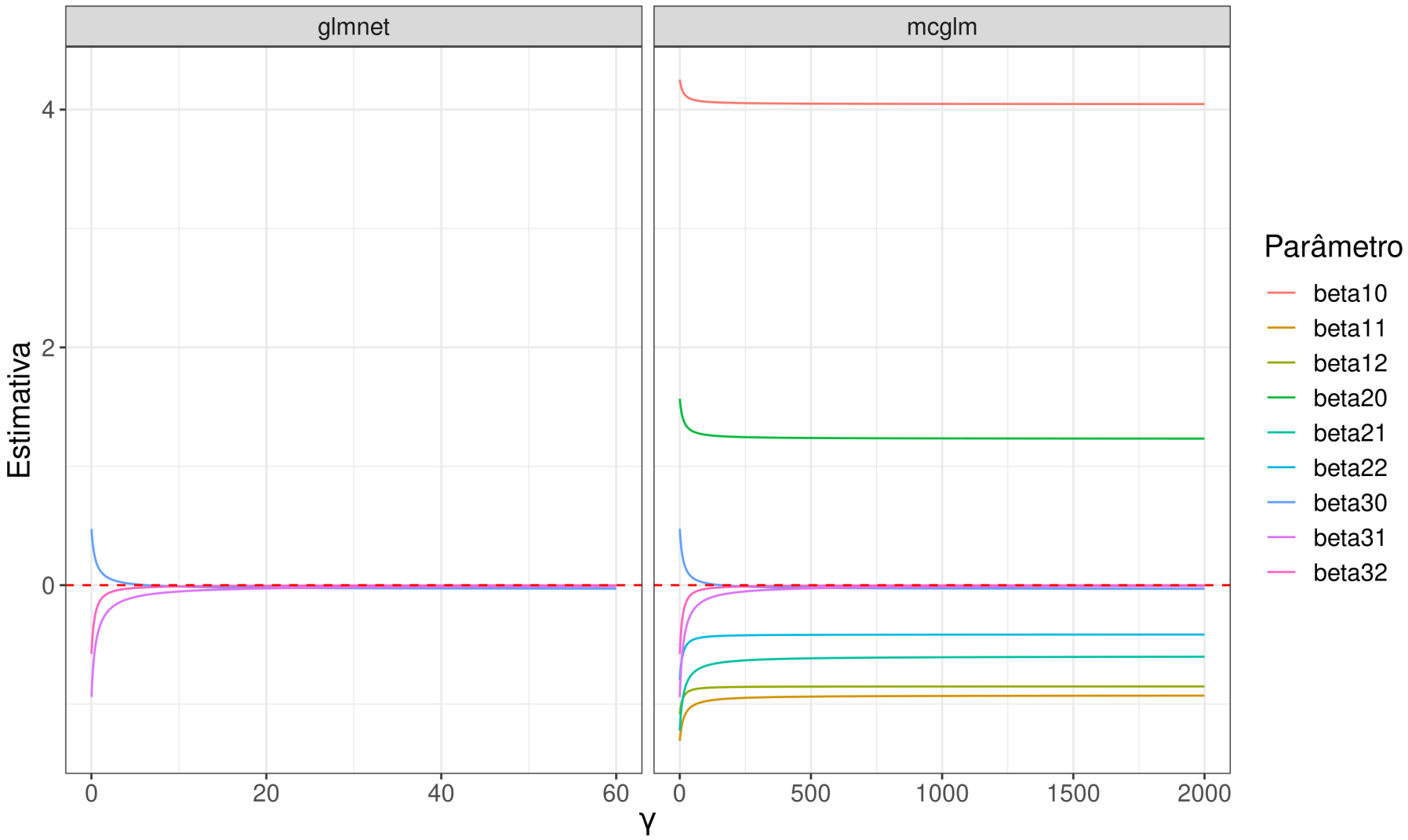
- Torna possível estimar um McGLM em *high dimensional data*.
- Melhora capacidade preditiva.
- Permite diminuir o número de parâmetros e construir redes de associação.
- Pode haver viés nas estimativas penalizadas.

Sumário

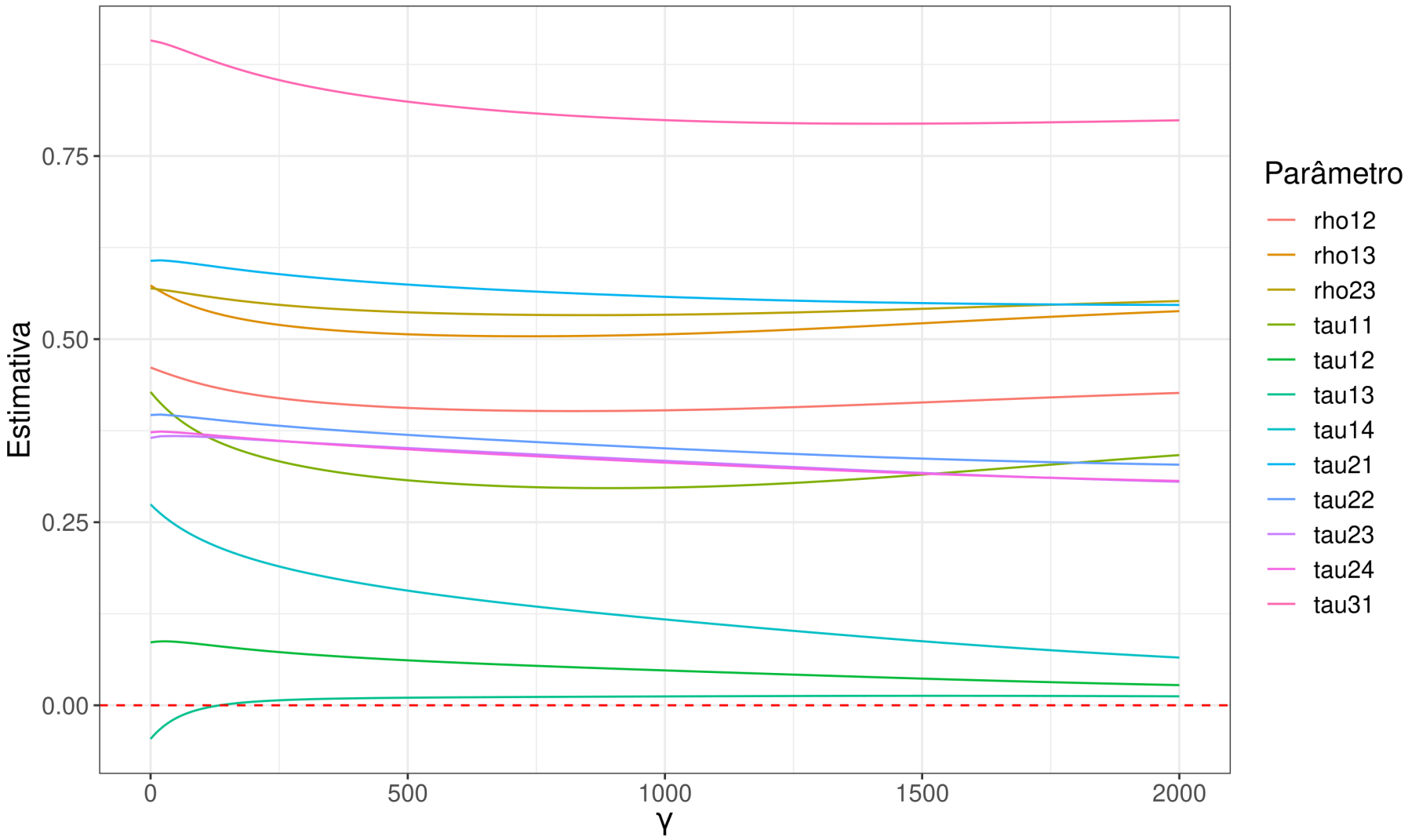
Estudo de simulação

- Conjunto de dados teca
  - Variáveis resposta: níveis de 3 elementos químicos no solo;
  - Covariável: Nível da medição no solo;
  - Disponível no pacote EACS (Zeviani, 2019).
- Objetivo: verificar se as estimativas penalizadas se comportam de maneira esperada.
- Penalização *Ridge*.
- Estrutura de covariância parecida com a não estruturada, mas com os interceptos.
- Estrutura de média linear.

Conjunto de dados teca



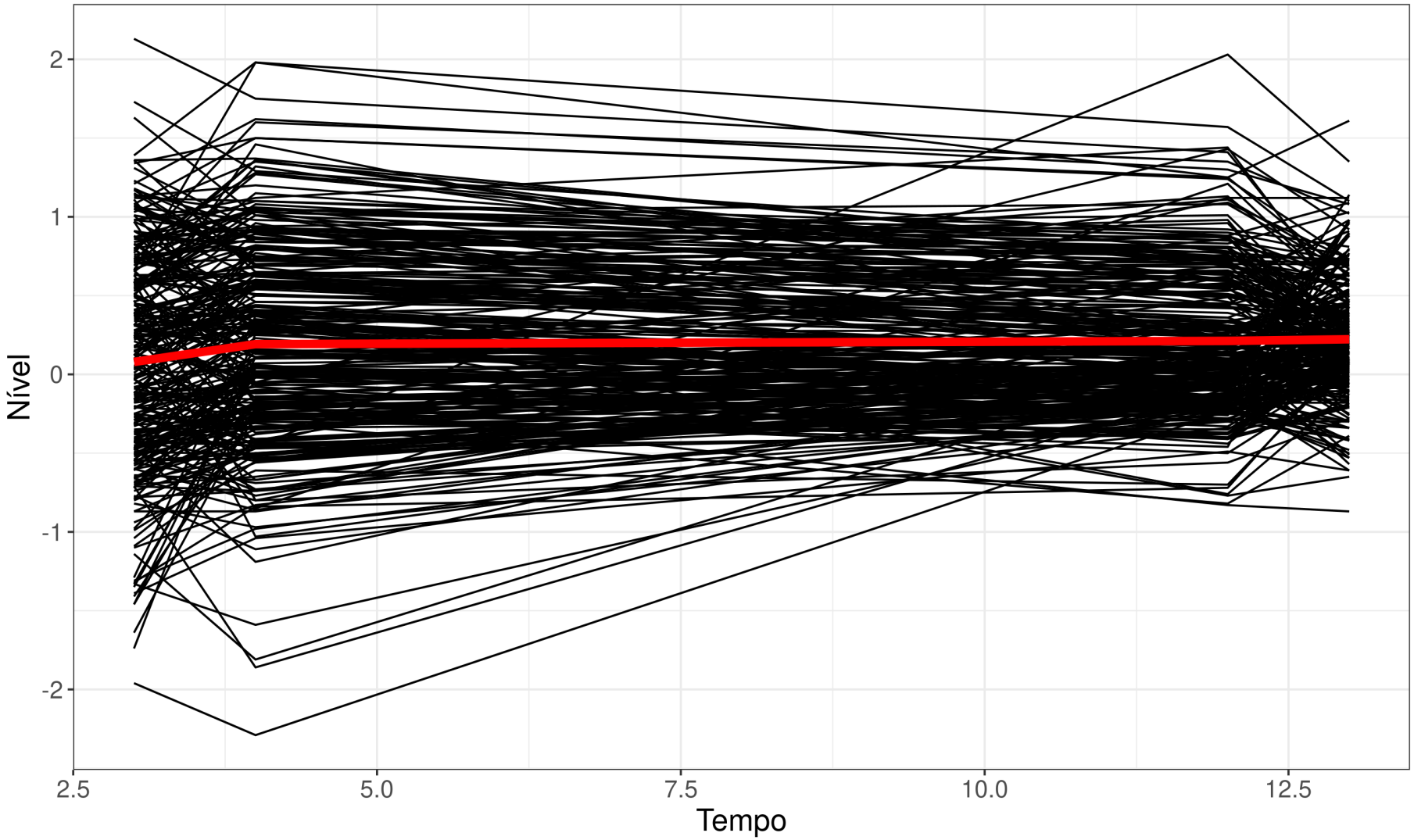
Conjunto de dados teca



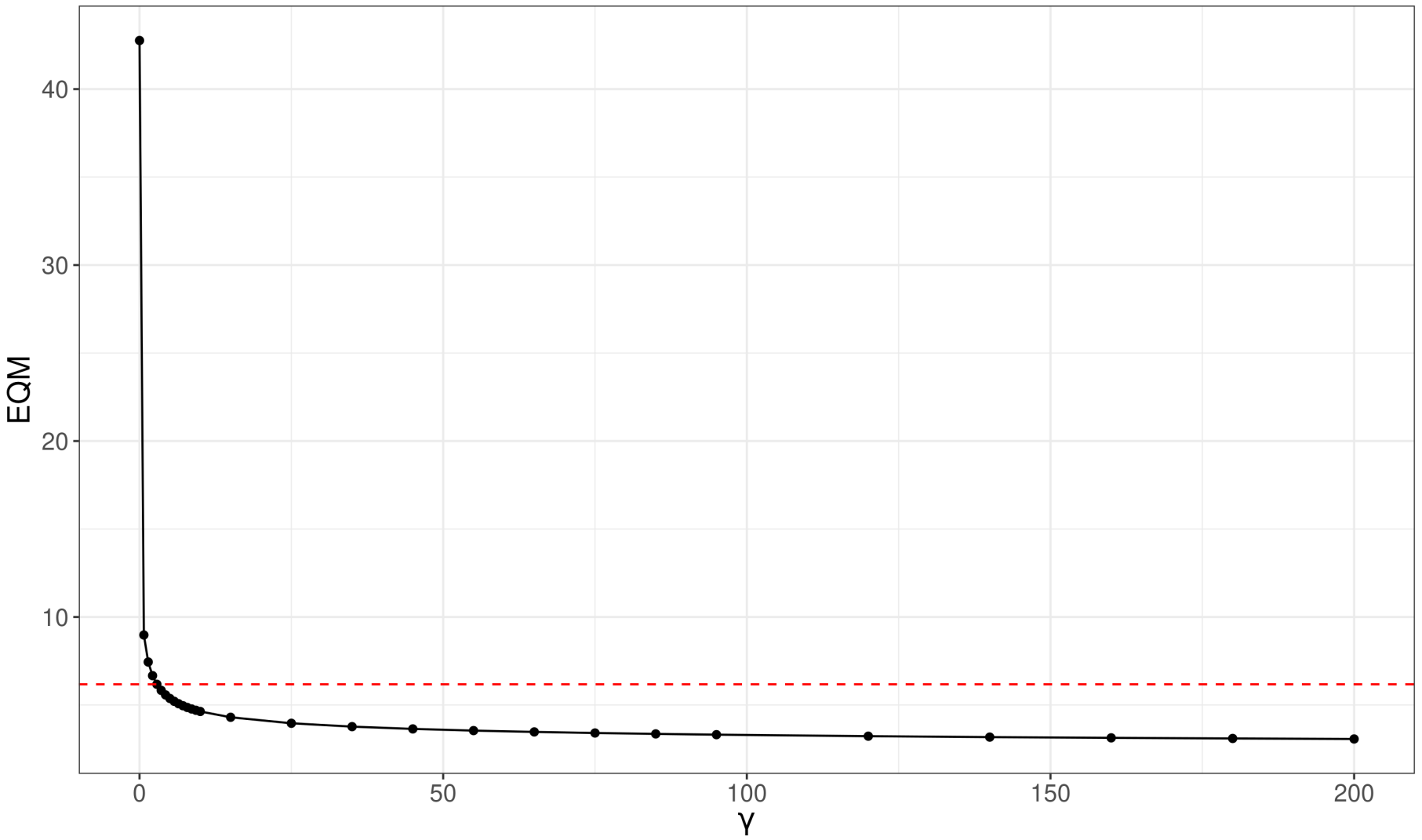
Conjunto de dados yeastG1

- Conjunto de dados yeastG1 (Spellman et al, 1998):
  - Variável resposta: nível de expressão genética em 4 tempo distintos;
  - 96 covariáveis;
  - 263 indivíduos;
  - Disponível no pacote PGEE (Inan et al, 2017).
- Objetivo: regularizar a estrutura de média para melhorar o erro preditivo.
- Penalização *Ridge*.
- Estrutura de covariância parecida com a não estruturada, mas com os interceptos.
- Estrutura de média linear.
- Seleção de γ via validação cruzada com 10 *folds*, utilizamos EQM.
- A estrutura de dependência foi preservada removendo os indivíduos para validação.

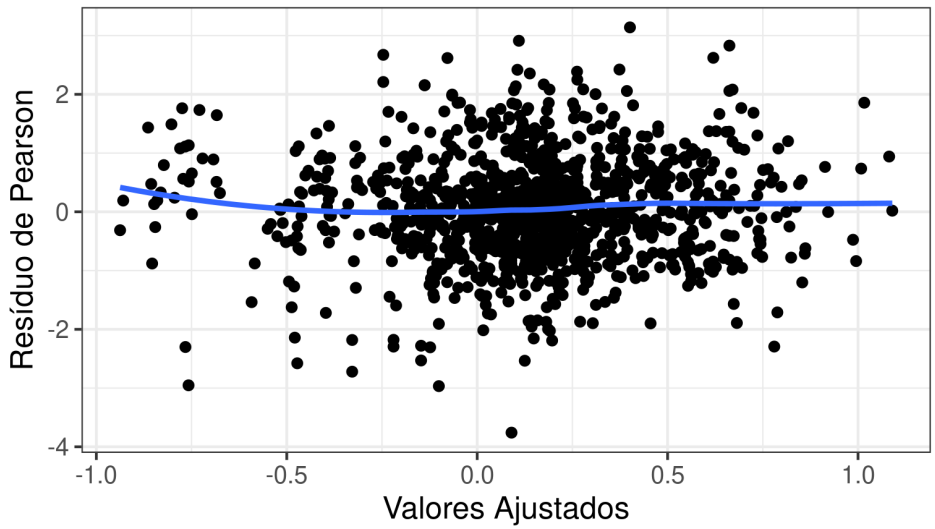
Conjunto de dados yeastG1

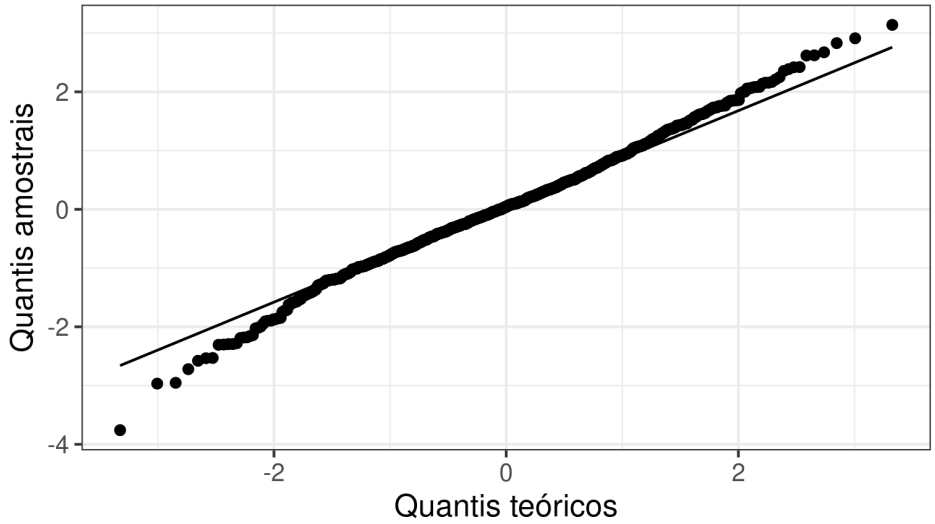


Conjunto de dados yeastG1



Conjunto de dados yeastG1





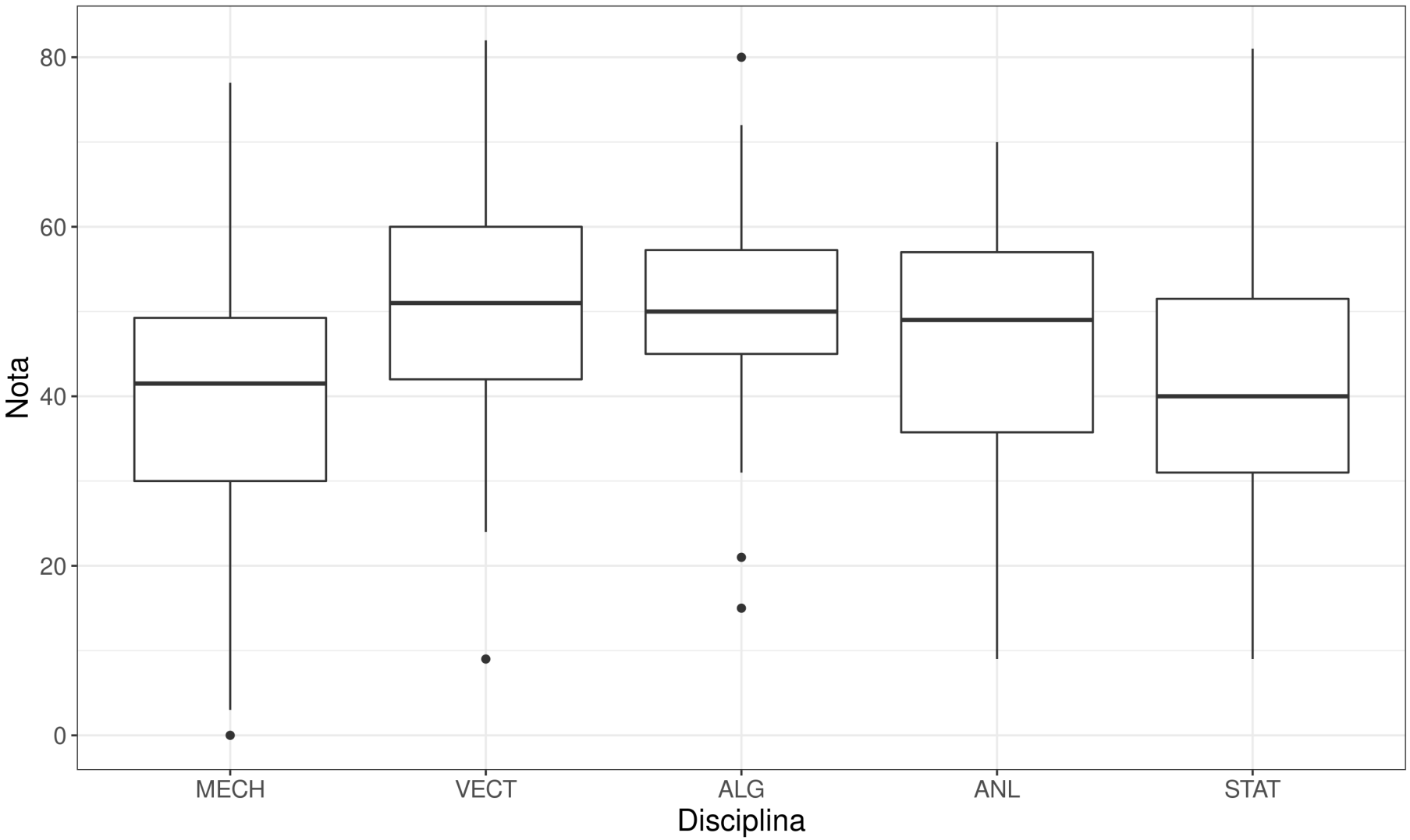
Conjunto de dados marks

- Conjunto de dados `marks` (Mardia et al, 1979):
  - Variável resposta: nota de 88 alunos em 6 disciplinas (Álgebra, Análise, Estatística, Mecânica, Vetores);
  - Disponível no pacote `bnlearn` (Scutari, 2010).
- Objetivo: construir uma rede de associações entre as disciplinas.
- Penalização *Ridge*.
- Estrutura de covariância parecida com a não estruturada, mas com os interceptos.
- Modelamos a precisão (função de ligação inversa).
- Estrutura de média linear.

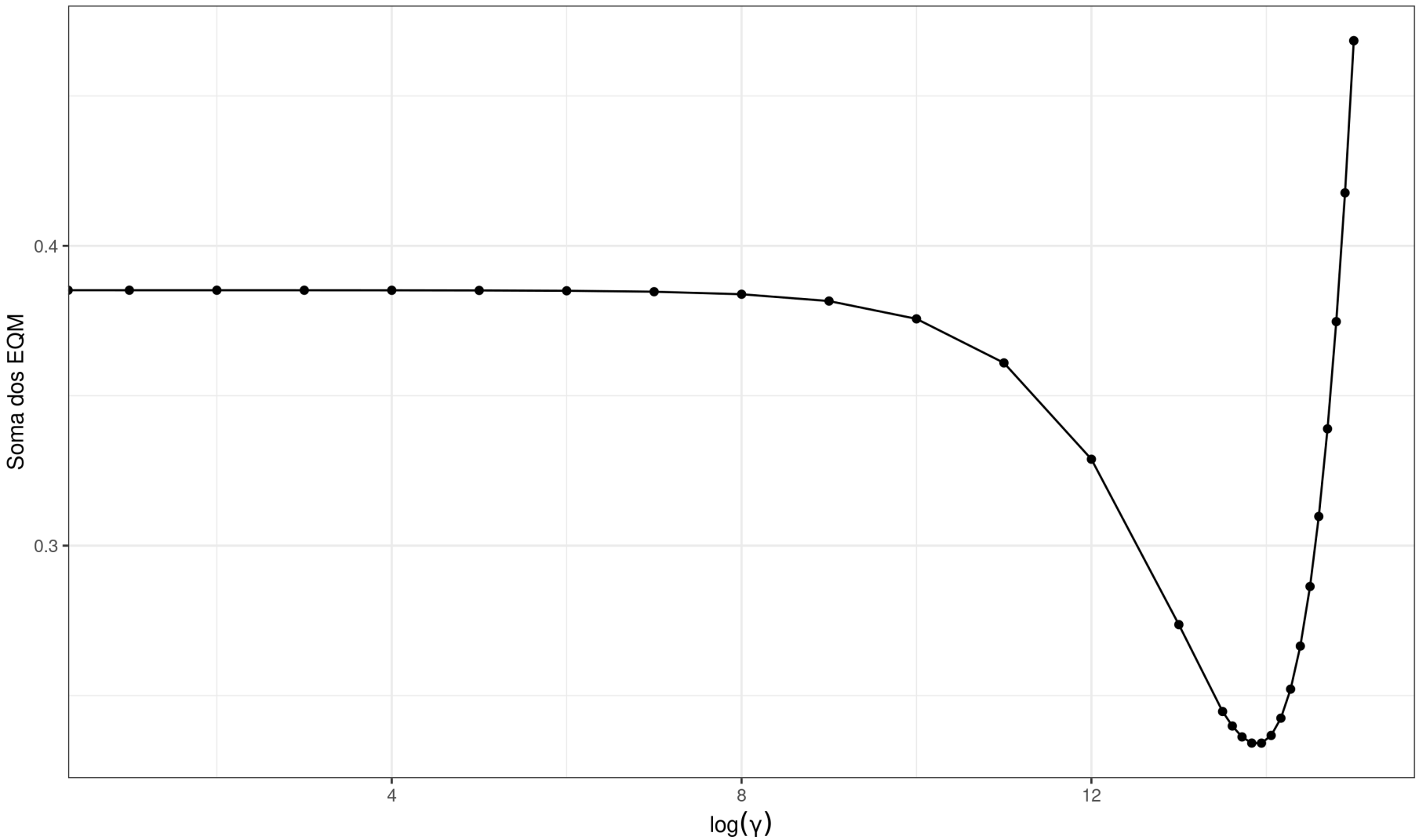
Conjunto de dados marks

- Seleção de γ via validação cruzada com uma fold.
- EQM do valor predito não foi possível.
- Utilizamos a soma dos EQMs entre a matriz de correlação da base de teste e as matrizes estimadas.

Conjunto de dados marks



Conjunto de dados marks

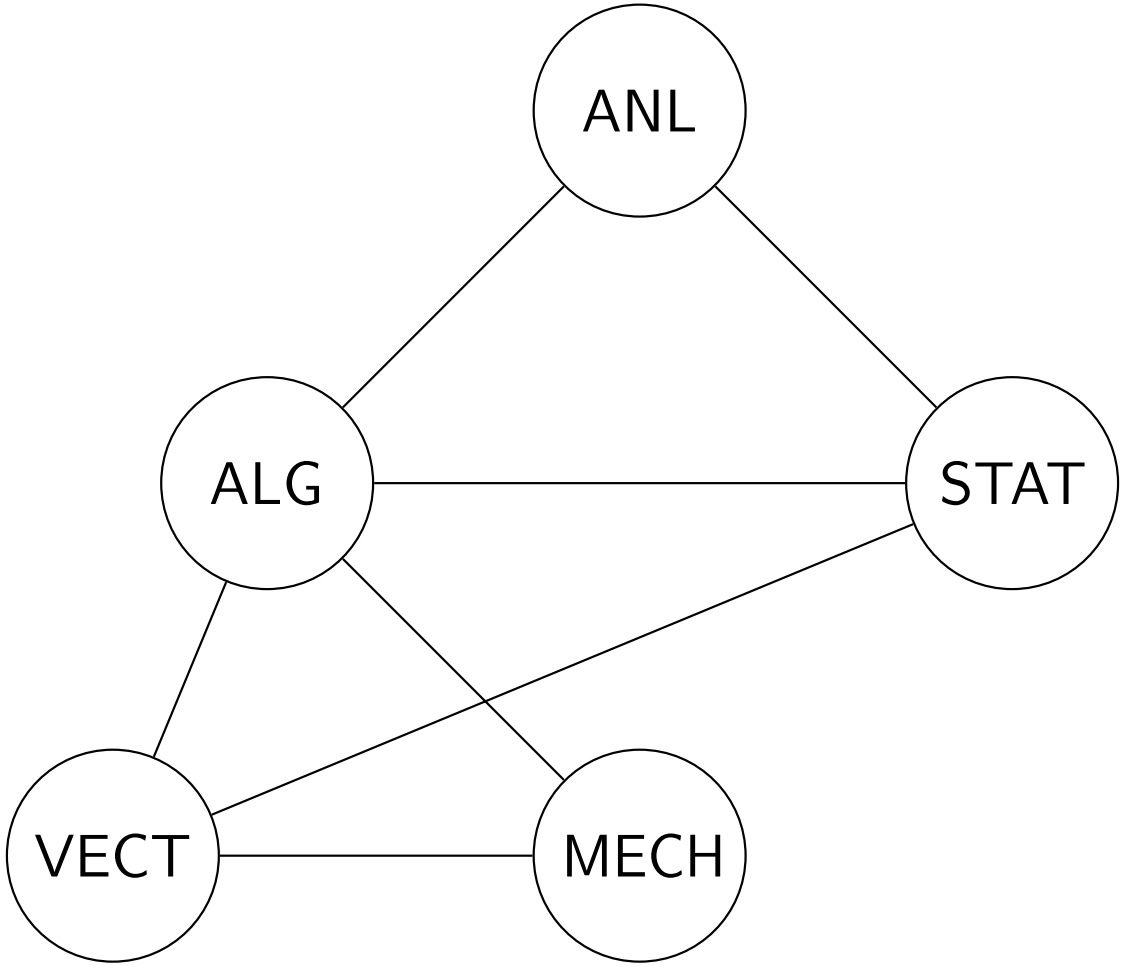


Conjunto de dados marks

- A conclusão pela existência de associação significativa é feita pelas estimativas dos coeficientes de dispersão.

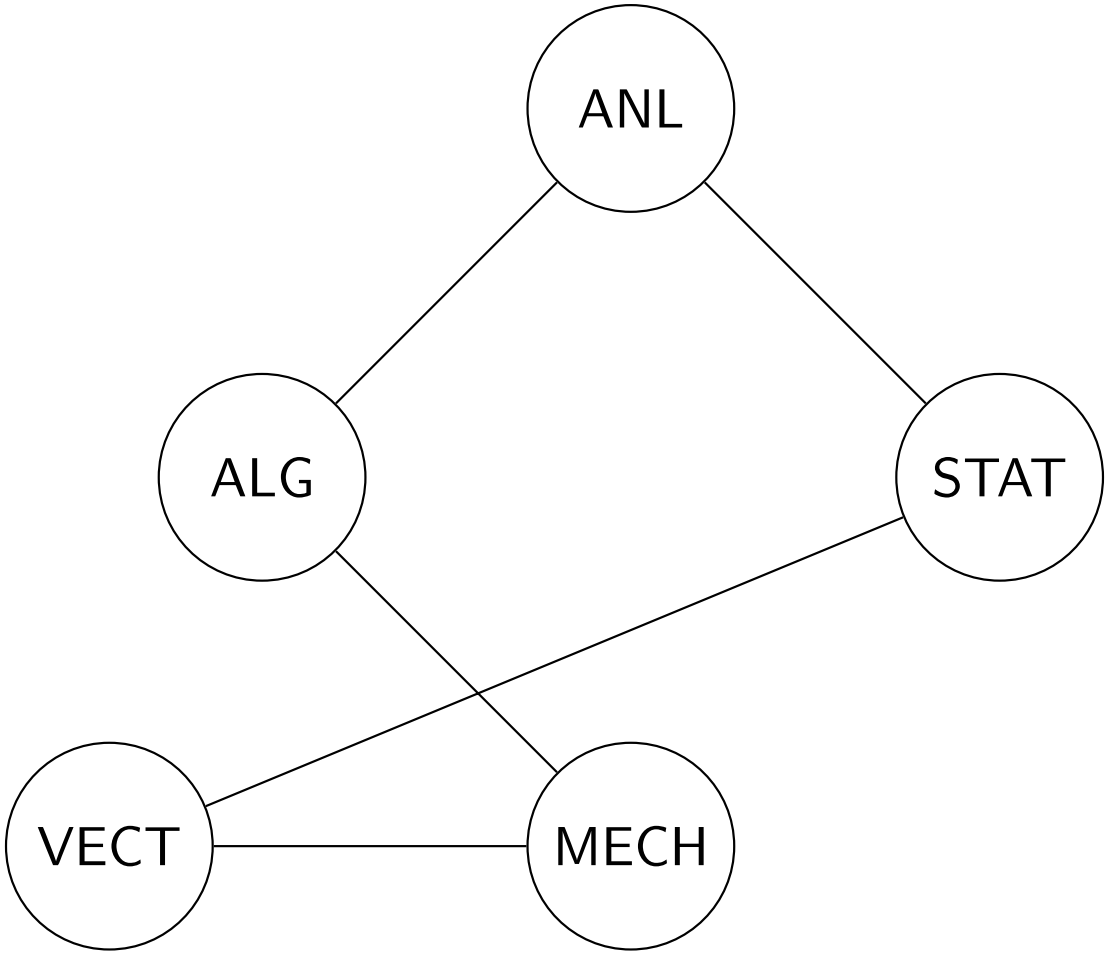
Coeficiente	Estimativa	Erro padrão	Estatística Z	Relação
$\hat{\tau}_{10}$	0.00644	0.00032	20.28	
$\hat{\tau}_{11}$	-0.00215	0.00040	−5.34	MECH–VECT
$\hat{\tau}_{12}$	-0.00124	0.00041	−3.03	MECH–ALG
$\hat{\tau}_{13}$	-0.00027	0.00041	-0.66	MECH–ANL
$\hat{\tau}_{14}$	-0.00055	0.00042	-1.33	MECH–STAT
$\hat{\tau}_{15}$	-0.00111	0.00041	−2.73	VECT–ALG
$\hat{\tau}_{16}$	-0.00068	0.00041	-1.67	VECT–ANL
$\hat{\tau}_{17}$	-0.00123	0.00041	−2.99	VECT–STAT
$\hat{\tau}_{18}$	-0.00201	0.00040	−5.02	ALG–ANL
$\hat{\tau}_{19}$	-0.00097	0.00041	−2.35	ALG–STAT
$\hat{\tau}_{110}$	-0.00164	0.00041	−4.01	STAT–ANL

Conjunto de dados marks



Conjunto de dados marks

- Aplicando a correção de Bonferroni a rede fica:



Sumário

Considerações e trabalho futuros

- Foi apresentado um modelo baseado em equações de estimação regularizadas para lidar com os mais variados conjuntos de dados.
- Consistência dos estimadores.
- Outros tipos de penalização.
- Escolha dos valores de γ.

Sumário

Referências

BONAT, W. H.; JØRGENSEN, B. Multivariate covariance generalized linear models. Journal of the Royal Statistical Society: Series C (Applied Statistics), v. 65, n. 5, p. 649–675, 2016.

FU, W. J. Penalized estimating equations. Biometrics, v. 59, n. 1, p. 126–132, 2003.

JØRGENSEN, B.; KNUDSEN, S. J. Parameter orthogonality and bias adjustment for estimating functions. Scandinavian Journal of Statistics, v. 31, n. 1, p. 93–114, 2004.

INAN, G.; ZHOU, J.; WANG, L. PGEE: Penalized Generalized Estimating Equations in High-Dimension. [S.l.], 2017. R package version 1.5.

MARDIA, K.; KENT, J.; BIBBY, J. Multivariate analysis. London [u.a.]: Acad. Press, 1979. (Probability and mathematical statistics). ISBN 0124712509.

SCUTARI, M. Learning bayesian networks with the bnlearn R package. Journal of Statistical Software, v. 35, n. 3, p. 1–22, 2010.

SPELLMAN, P. T. et al. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell, v. 9, n. 12, p. 3273–3297, 1998. PMID: 9843569.

ZEVIANI, W. M. EACS: Estatística Aplicada à Ciência do Solo. [S.l.], 2019.