Priors for sparse regression modelling

Paloma Vaissman Uribe

Instituto de Matemática e Estatística - USP

paloma.uribe@gmail.com

19 de fevereiro de 2016

Inferência com a priori normal-gamma em problemas de regressão

- O modelo padrão de regressão múltipla assume que o vetor de respostas $y = (y_1, y_2, ..., y_n)$pode ser representado por

$$y = \alpha 1 + X\beta + \epsilon, \tag{1}$$

em que $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ são independentes, $p(\epsilon_i) = N(\epsilon_i|0, \sigma^2)$ e X é uma matriz $(n \times p)$ de variáveis explicativas. O escalar α é o intercepto e 1 é um vetor unitário $(n \times 1)$.

- [Griffin et al., 2010] está preocupado com a escolha da distribuição a priori do vetor β, de dimensão $(p \times 1)$ de coeficientes da regressão múltipla de forma a realizar **seleção de variáveis**.

A abordagem padrão: priori "spike-and-slab"[Mitchell and Beauchamp, 1988]

- A variável indicadora z_i é introduzida para identificar se a i-ésima variável deve ser incluída no modelo ($z_i = 1$) ou excluída ($z_i = 0$). A priori para β_i pode ser expressa por

$$\pi(\beta_i) = z_i N(\beta_i|0, {\sigma_\beta}^2) + (1 - z_i)\delta_{\beta_i=0},\ p(z_i = 1) = \omega, \qquad (2)$$

em que $\delta_{\beta_i=0}$ é a medida delta de Dirac que coloca toda a massa em zero.

- As variáveis Bernouilli independentes z_i tem média ω, e portanto o hiperparâmetro ω pode ser interpretado como sendo a proporção a priori de regressores não nulos. Alternativamente, uma distribuição a priori pode ser utilizada para ω e seu valor inferido através dos dados.
- A escala ${\sigma_\beta}^2$ controla a variância da priori.

A esperança a posteriori dos coeficientes da regressão linear

Proposição. Suponha que temos um modelo de regressão linear dado por (1) em que a variância do erro da regressão σ^2 é conhecida, a matriz de design X, de dimensão $n \geq p+1$, é não singular e suas colunas foram centradas, e o intercepto α é independente de β a priori. Seja $\hat{\beta}$ o estimador de mínimos quadrados de β, e $h(\hat{\beta}) = \int N(\hat{\beta}|\beta, \sigma^2(X^TX)^{-1})\pi(\beta)d\beta$, em que $\pi(\beta)$ é a distribuição a priori de β, então

$$\begin{aligned} E(\beta|\hat{\beta}) &= (I - S(\hat{\beta}))\hat{\beta} \\ V(\beta|\hat{\beta}) &= \sigma^2(X^TX)^{-1} - \sigma^4(X^TX)^{-1}W(\hat{\beta})(X^TX)^{-1}, \end{aligned} \tag{3}$$

onde $S(\hat{\beta}) = \sigma^2(X^TX)^{-1}R(\hat{\beta})$, $R(x)$ é uma matriz diagonal com elementos $R_{ii}(x) = -\frac{1}{x_i}\frac{\partial}{\partial x_i}\log h(x)$, e $W(x) = -\frac{\partial}{\partial x}\frac{\partial}{\partial x^T}\log h(x)$.

Implicações da Proposição

- A densidade amostral de $\hat{\beta}$ é $N(\beta, \sigma^2(X^T X)^{-1})$ e assim $h(\hat{\beta})$ é a **densidade preditiva a priori** de $\hat{\beta}$.
- A esperança a posteriori é sempre **uma versão matricial encolhida do estimador de mínimos quadrados**. A magnitude do encolhimento ou compressão é controlada pela forma de h (a derivada do log da distribuição preditiva) e pelo erro padrão de $\hat{\beta}$.
- Em contraste, o estimador de máxima verossimilhança penalizado também é um estimador encolhido, sendo que a magnitude da compressão controlada pela derivada da função de penalização.
- O resultado pode ser estendido para X singular, em que $p > n + 1$, utilizando a decomposição do valor singular de X e explorando propriedades das distribuições de misturas de escala normal.

A priori Normal-Gamma

- Uma classe natural e bastante extensa de densidades a priori dos coeficientes de regressão é conhecida como distribuições de mistura de escala normal, cuja densidade pode ser expressa por

$$\begin{aligned} \pi(\beta_i) &= \int N(\beta_i|0,\Psi_i)dG(\Psi_i) \\ \beta_i|\Psi_i &\sim N(0,\Psi_i),\ \Psi_i \sim G, \end{aligned} \tag{4}$$

  onde G é uma distribuição de mistura.

- A distribuição normal-gamma surge assumindo que a distribuição de mistura G é $g(x) = Gamma(x|\lambda, 1/(2\gamma^2))$. A função de densidade é

$$\pi(\beta_i) = \frac{1}{\sqrt{\pi}2^{\lambda-1/2}\gamma^{\lambda+1/2}\Gamma(\lambda)}|\beta_i|^{\lambda-1/2}K_{\lambda-1/2}(|\beta_i|/\gamma), \tag{5}$$

  em que K é a função Bessel do terceiro tipo. A variância de β_i é $2\lambda\gamma^2$ e o excesso de curtose $3/\lambda$.

A priori Normal-Gamma

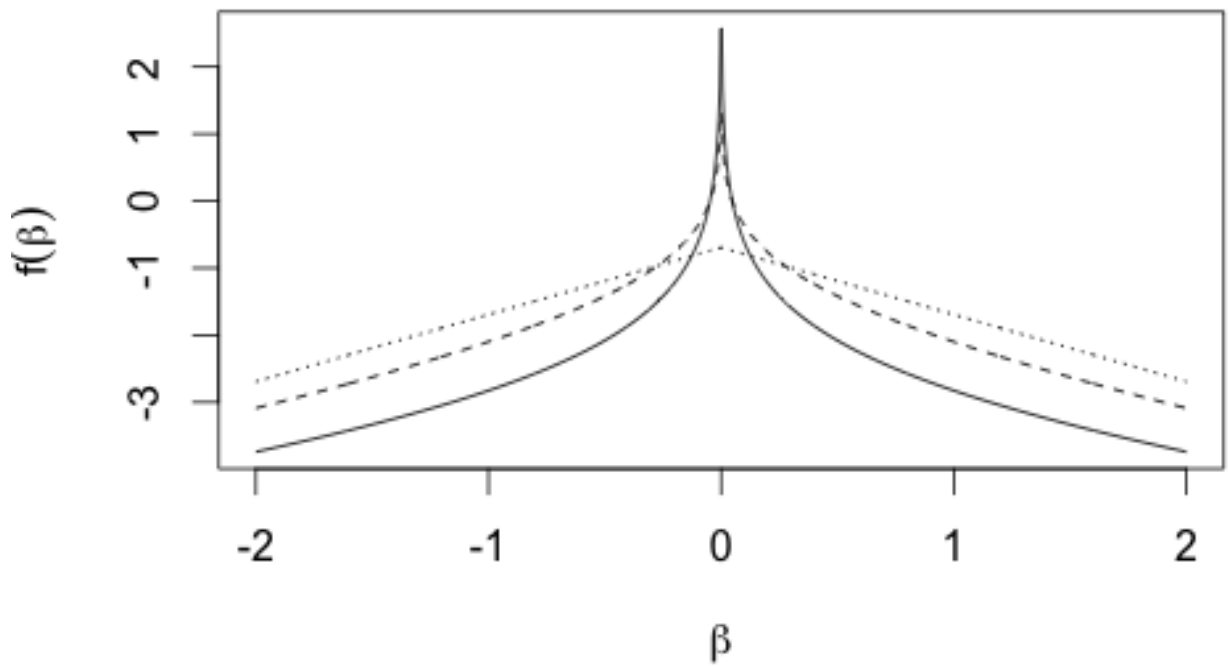

Figura 1: Log da densidade da priori normal-gamma prior com variância 2 e diferentes valores para λ. $\lambda = 0,1$ (linha cheia), $\lambda = 0,333$ (linha tracejada) and $\lambda = 1$ (linha pontilhada).

Outras misturas de escala normal

- A priori "spike-and-slab"também é uma mistura de escala normal em que a distribuição de mistura G é

$$G(\Psi_i) = z_i \delta_{\Psi_i = \sigma_\beta{}^2} + (1 - z_i)\delta_{\Psi_i = 0}. \tag{6}$$

- A priori exponencial dupla do Lasso Bayesiano também pertence à classe, sendo G uma distribuição exponencial, o que é o mesmo que substituir $\lambda = 1$ na densidade da Normal-Gamma.

O efeito de λ na priori Normal-Gamma

- Segue da definição do modelo de regressão linear em (1) e das distribuições de misturas de escala normal em (4) que

$$V[y_i|\Psi, \sigma^2] = V[\alpha] + \sum_{j=1}^{p} \Psi_j + \sigma^2 \tag{7}$$

  se os regressores foram padronizados tal que a média amostral seja 0 e a variância 1.
- Assim, $\zeta_j = \frac{\Psi_j}{\sum_{k=1}^{p} \Psi_k}$ pode ser interpretado como sendo a proporção da variabilidade total explicada pelo j-ésimo regressor. Se $G \sim Gamma$, então $\zeta \sim Di(\lambda, ..., \lambda)$.
- Aumentar λ faz com que $\zeta_1, ..., \zeta_p$ sejam mais próximos entre si.
- Pequenos valores para λ estão associados a maiores diferenças entre as proporções => **maior encolhimento**.

O fator de encolhimento $S(\hat{\beta})$

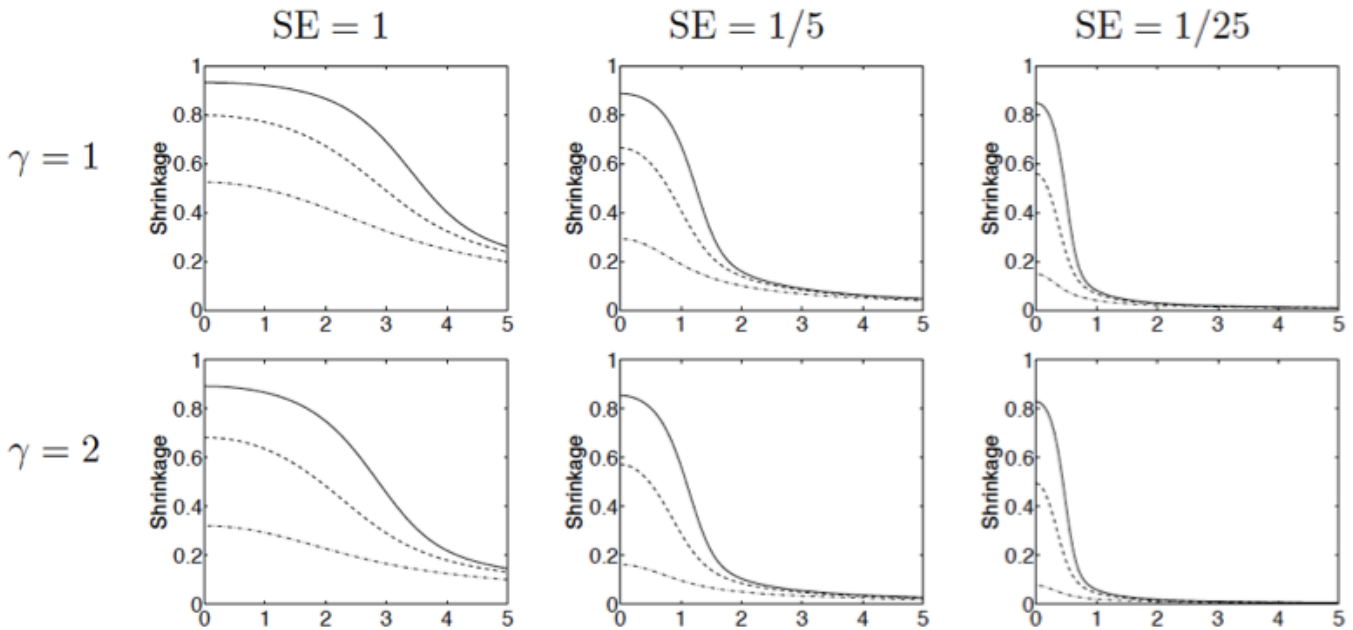

Figura 2: As propriedades a posteriori dos coeficientes da regressão podem ser estudadas via Proposição. O fator de encolhimento, $S(\hat{\beta})$, para a esperança a posteriori de um dos coeficientes da regressão múltipla é representado graficamente para diferentes valores de erro padrão de $\hat{\beta}$ e de λ: $\lambda = 0,1$ (linha cheia), $\lambda = 0,333$ (linha tracejada) e $\lambda = 1$ (linha pontilhada).

Hiperparâmetros da priori Normal-Gamma

- Os hiperparâmetros da priori normal-gamma podem ser escolhidos de forma a aproximar-se da priori "spike-and- slab". Contudo, também pode-se estabelecer distribuições a priori para γ e λ:

$$\lambda \sim Exp(1), \tag{8}$$

que oferece variabilidade ao redor da priori do Lasso Bayesiano ($\lambda = 1$), e

$$v_\beta = 2\lambda\gamma^2 \sim IG(2, M), \tag{9}$$

em que IG é a distribuição gamma invertida,tal que $IG(2, M)$ tem média M. Quando X é não singular $M = \frac{1}{p}\sum_{i=1}^{p} \hat{\beta}_i^2$ em que $\hat{\beta}$ é a estimativa de mínimos quadrados. Quando X é singular, como quando $p > n + 1$, $M = \frac{1}{n}\sum_{i=1}^{p} \hat{\beta}_i^2$, em que $\hat{\beta}$ é a estimativa de mínima distância.

A distribuição Normal-Gamma Multivariada

- Suponha $\beta = C\phi$, em que $C = (C_{ik})$ é uma matriz de dimensão $(p \times q)$ com elementos reais e ϕ é um vetor q-dimensional de variáveis independentes $\phi_k \sim NG(\lambda_k, 1/2)$. Diz-se então que β tem distribuição normal-gamma correlacionada p-variada, expressa por:

$$\beta \sim CNG(\lambda, C), \tag{10}$$

  em que $\lambda = (\lambda_1, ..., \lambda_q)$, $\lambda_k \geq 0$, $(k = 1, ..., q)$.
- Assume-se que a matriz de covariância $(p \times p)$-dimensional de β, $Cdiag(\lambda)C^T$, tem posto pleno p.

A distribuição Normal-Gamma Multivariada

- Seja $S_i(C)$ o subconjunto de $\{1, ..., q\}$ tal que C_{ik} é não nulo $(k = 1, ..., q)$. A densidade marginal de β_i pode ser expressa por uma distribuição de misturas de escala normal

$$p(\beta_i) = N(\beta_i|0, \Psi_i)g(\Psi_i)d\Psi_i, \tag{11}$$

em que $\Psi_i = \sum_{k \in S_i(C)} \zeta_{ik}$ e $\zeta_{ik} \sim Gamma(\lambda_k, 1/(C_{ik}^2))$, tal que Ψ_i é convolução de variáveis gamma independentes com diferentes escalas.

- A densidade de Ψ_i é uma soma infinita dada por [Moschopoulos, 1985]:

$$g(\Psi_i) \propto \Psi_i^{\eta_i - 1} exp(-b_i^* \Psi_i) \sum_{l=0}^{\infty} \delta_l \Psi_i^l, \tag{12}$$

e, portanto, o formato da densidade é do tipo gamma com o parâmetro de forma controlado pelo parâmetro de forma agregado $\eta_i = \sum_{k \in S_i(C)} \lambda_k$, com um pico se $\eta_i \leq 1$.

A distribuição Normal-Gamma Multivariada

- A escolha de $C = \gamma^2 B$, em que γ é um escalar e B é uma matriz $(p \times q)$, sendo que B_{ik} é 0 ou 1, gera a seguinte distribuição marginal para β_i:

$$\beta_i \sim NG(\sum_{k=1}^{q} B_{ik}\lambda_k, 1/(2\gamma^2)), \tag{13}$$

  com $var(\beta_i) = 2\gamma^2 \sum_{k=1}^{q} B_{ik}\lambda_k$.

- Se além disso, β_i e β_j forem identicamente distribuídas, então a correlação entre eles é

$$corr(\beta_i, \beta_j) = \frac{\sum_{k=1}^{q} B_{ik}B_{jk}\lambda_k}{\sum_{k=1}^{q} B_{ik}\lambda_k},\ i,j = 1, ..., p. \tag{14}$$

A distribuição Normal-Gamma Multivariada: exemplo

- Considere $p = 2$, $q = 3$, $B = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$, e $\lambda = \{\rho\lambda^*, (1-\rho)\lambda^*, \lambda^*\}$.
- Essa parametrização implica que as distribuições marginais de β_1 e β_2 são $NG(\lambda^*, \gamma)$ e que $corr(\beta_1, \beta_2) = \rho$, sendo $\gamma = \frac{1}{\sqrt{\lambda^*}}$.
- Nota-se que a massa de probabilidade da densidade condicional de β_1 dado β_2 é aumentada ao redor do valor de β_2. Esse efeito é maior, quanto maior for ρ, e menor for λ^*, tudo mais constante.

A distribuição Normal-Gamma Multivariada: exemplo

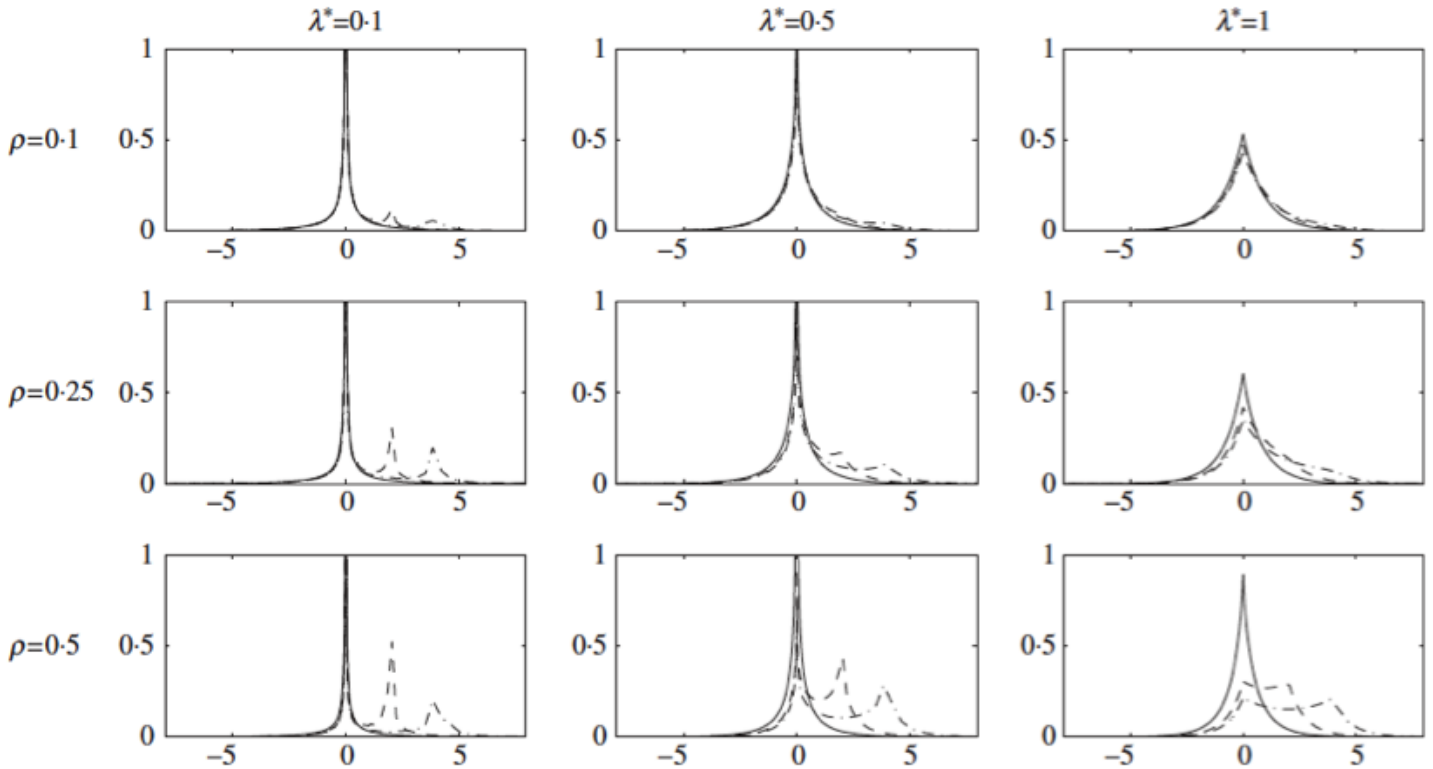

Figura 3: Distribuição condicional de β_1 dado $\beta_2 = 0,0001$ (linha cheia), $\beta_2 = 2$ (linha tracejada), e $\beta_2 = 4$ (linha pontilhada).

Regressão com variáveis categóricas

- Suponha um modelo que os regressores são variáveis categóricas não ordenadas, em que a g-ésima variável tem p_g níveis:

$$y = \alpha 1 + \sum_{g=1}^{G} X_g \beta_g + \epsilon, \tag{15}$$

  em que X_g é uma matriz $n \times (p_g - 1)$ de variáveis *dummy* construídas tomando o primeiro nível como base e definindo:
  - $(X_g)_{ij} = 1$, se a i-ésima observação pertence ao $(j+1)$-ésimo nível,
  - $(X_g)_{ij} = 0$, c.c.
- O coeficiente $\beta_{g,j}$ denota a diferença entre o efeito do $(j+1)$-ésimo nível e do primeiro nível.
- Assume-se que $\beta_1, ..., \beta_G$ são independentes e que $\beta_i \sim CNG(\lambda_i, \gamma^2 B_i)$.

Regressão com variáveis categóricas

- Se a variável categórica é não ordenada, β_g deveria ser invariante às permutações entre os níveis e ao nível base escolhido. Utilizando a Normal-Gamma Correlacionada, isso pode ser atingido assumindo

$$\beta_g \sim CNG(\lambda_g, \gamma^2 B^{(p)}), \tag{16}$$

  em que $\lambda_{g,i} = \lambda^*/2^{p_g - 1}$, e $B^{(p)}$ a matriz $p \times (2^p - 1)$, cujas colunas são as 2^p combinações de 0s e 1s, omitindo-se o vetor nulo.
- Segue que $\beta_{g,i} \sim NG(\lambda^*, 1/(2\gamma^2))$ e também que $\beta_{g,i} - \beta_{g,j} \sim NG(\lambda^*, 1/(2\gamma^2))$, para qualquer que seja o nível base.
- Além disso, $corr(\beta_{g,i}, \beta_{g,j}) = 0,5$, $i, j = 1, ..., p_g$.

Regressão com variáveis categóricas: exemplo

- Regressão da pontuação média na parte verbal do teste SAT em cada estado dos EUA. Regressores: população do estado, percentual dos alunos que concluíram o *high school* que fizeram o SAT, gastos públicos por aluno feitos por cada estado, média de salários dos professores de cada estado, além da variável categórica de região, composta de 9 níveis (East North Central (1), East South Central (2), Mid-Atlantic (3), Mountain (4), New England (5), Pacific (6), South Atlantic (7),West North Central (8) e West South Central (9)).

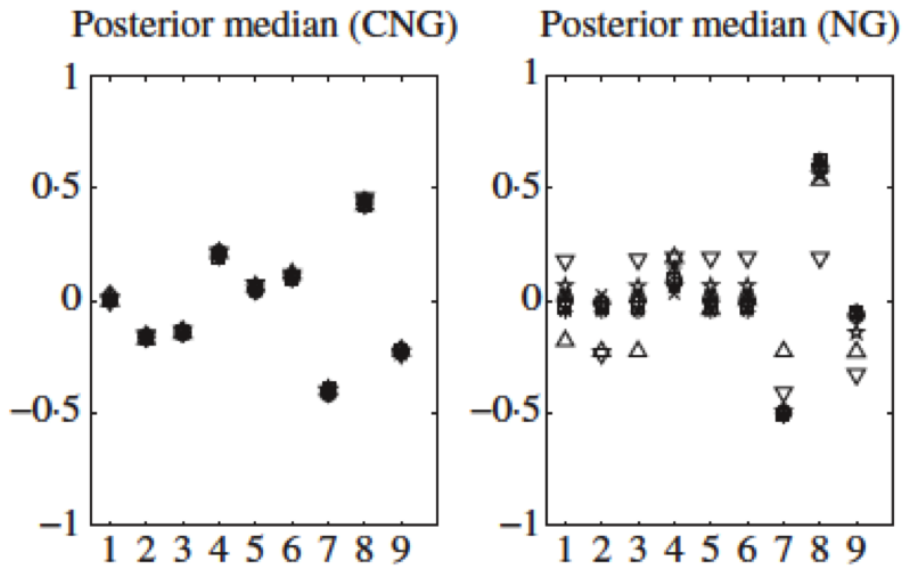

Figura 4: Medianas a posteriori para os coeficientes das regressões para diferentes níveis base, usando a Normal-Gamma Correlacionada (CNG) e prioris Normal-Gamma independentes.

Referências I


Griffin, J. E. and Brown, P. J. (2012).
Structuring shrinkage: some correlated priors for regression.
Biometrika, page asr082.


Griffin, J. E., Brown, P. J., et al. (2010).
Inference with normal-gamma prior distributions in regression problems.
Bayesian Analysis, 5(1):171–188.


Mitchell, T. J. and Beauchamp, J. J. (1988).
Bayesian variable selection in linear regression.
Journal of the American Statistical Association, 83(404):1023–1032.


Moschopoulos, P. G. (1985).
The distribution of the sum of independent gamma random variables.
Annals of the Institute of Statistical Mathematics, 37(1):541–544.