Uma Aplicação do Método Bayesiano para Seleção de Modelos via Mistura de Componentes *Spike-and-Slab*

Erick da Conceição Amorim[1] Guilherme Lopes de Oliveira[1,2]

[1]Departamento de Estatística - ICEx - UFMG

[2]Departamento de Computação - CEFET-MG

1 Introdução

Um dos problemas comum em análise de regressão é selecionar, dentro do conjunto de covariáveis preditoras disponíveis, aquelas que são estatisticamente mais relevantes para descrever o comportamento da variável resposta de interesse. Existe uma variedade de critérios para seleção modelos, tais como, o *Deviance Information Criterion* (DIC), *Akaike Information Criterion* (AIC), *Conditional Predictive Ordinate* (CPO) e *Bayesian Information Criterion* (BIC).

No contexto Bayesiano, Chen et al. (2008) consideraram conexões teóricas e computacionais entre seis dos mais populares métodos de seleção de variáveis na classe dos Modelos Lineares Generalizados (MLG) e, com base nas distribuições *a priori* conjugadas abordadas por Chen e Ibrahim (2003), obtiveram uma forma analítica fechada para relações entre o Fator de Bayes, CPO, DIC, AIC e BIC nesta classe de modelos. O desafio de encontrar o melhor modelo no contexto linear geral em que $g(Y_i) = \beta_0 + \beta_1 X^*_{i1} + ... + \beta_q X^*_{iq}$, com $X^*_{i1}, X^*_{i2}, ..., X^*_{iq}$ selecionados de $X_{i1}, X_{i2}, ..., X_{ik}$, $q \leq k$; é que os procedimentos de seleção são baseados na comparação de 2^k possíveis submodelos gerados pela retirada e/ou inclusão de subconjuntos das covariáveis. Neste processo, são comumente usados os métodos conhecidos como *backward elimination*, *forward selection* ou *stepwise*, os quais geralmente se baseiam no AIC ou DIC. A principal dificuldade aparece quando k é muito grande, pois o custo computacional envolvido no uso destes critérios de seleção pode ficar demasiadamente elevado.

George e McCulloch (1993) apresentam um procedimento Bayesiano para selecionar subconjuntos de variáveis preditoras que são mais relevantes para o ajuste de um modelo de regressão múltipla. Esse procedimento é baseado na especificação de uma distribuição *a priori* em forma de mistura para os coeficientes do modelo de regressão. Tal proposta foi originalmente apresentado por Leamer (1978) e Mitchell e Beauchamp (1998), mas foi popularmente difundida por George e McCulloch (1993). Em geral, considera-se uma mistura de duas componentes com distribuição Gaussiana (ou Normal), ambas centradas em zero, uma delas com variância pequena (chamada de componente *spike*) e outra com variância grande (chamada de componente *slab*). Propostas de misturas como essas são chamadas de distribuições *a priori* esparsas e um modelo que usa essa especificação é chamado de modelo esparso. Opcionalmente, Geweke (1996) propôs uma distribuição *a priori* para cada coeficiente de regressão representada por uma mistura com uma componente degenerada em zero e uma distribuição Gaussiana possivelmente truncada.

No contexto Bayesiano a ideia principal é deixar que os dados determinem as probabilidades *a posteriori* associada a cada modelo. Nesse contexto, quando o número de covariáveis k é muito grande, fazer uma exploração completa no espaço dos modelos (2^k submodelos) fica inviável computacionalmente. A vantagem dos procedimentos Bayesianos é que métodos Monte Carlo via Cadeias de Markov (MCMC) apropriados podem ser usados para fazer uma busca estocástica de forma rápida e eficiente, explorando a distribuição *a posteriori* em busca do melhor modelo, ou seja, aquele com a maior probabilidade *a posteriori*.

Neste trabalho iremos utilizar distribuições *a priori* na forma de misturas, as quais são conhecidas na literatura como *priori "spike and slab"* no sentido discutido previamente. Apresentaremos o modelo de regressão esparso que incorpora os elementos essenciais para a seleção de covariáveis. Em seguida é feito uma aplicação em um conjunto de dados reais sobre avaliação da qualidade de vinhos. O resultado obtido pelos métodos Bayesiano e o método *stepwise* convencional serão comparados quanto ao subconjunto de covariáveis mantidas na análise.

2 Metodologia

Considere o modelo linear Gaussiano tal que $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik} + \epsilon_i$, com $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, e $i = 1, ..., n$. Utilizaremos como notação: $Y = (Y_1, ..., Y_n)'$, $\beta = (\beta_0, \beta_1, ..., \beta_k)'$, $\beta_{-j} = (\beta_0, \beta_1, ..., \beta_{j-1}, \beta_{j+1}, ..., \beta_k)'$, para $j = 1, ..., k$ e n sendo o tamanho da amostra.

No contexto de seleção de variáveis Bayesiano fazendo uso de distribuições *a priori spike and slab*, supondo a existência de um conjunto de k covariáveis, teremos o seguinte modelo hierárquico:

$$\begin{aligned} (\beta_j \mid q_j, V_{\beta_j}) &\sim (1 - q_j)\delta_0(\beta_j) + q_j N(0, V_{\beta_j}) \\ q_j &\sim \text{Beta}(\gamma_{1j}, \gamma_{2j}), \end{aligned}$$

para $j = 1, ..., k$, sendo $\delta_0(\beta_j)$ uma distribuição com ponto de massa em zero, ou seja, β_j é igual a zero com probabilidade 1.

A fim de facilitar o processo de amostragem das distribuições *a posteriori* de interesse, é comum fazer uso de um modelo aumentado por meio de variáveis latentes z_j, $j = 1, ..., k$, de forma que

$$\begin{aligned} (\beta_j \mid z_j, V_{\beta_j}) &\sim (1 - z_j)\delta_0(\beta_j) + z_j N(0, V_{\beta_j}) \\ (z_j \mid q_j) &\sim \text{Bernoulli}(q_j) \\ q_j &\sim \text{Beta}(\gamma_{1j}, \gamma_{2j}). \end{aligned}$$

A especificação do modelo é completada pela elicitação de uma distribuição *a priori* $N(0, V_{\beta_0})$ para o parâmetro β_0 e para o termo de variância do erro aleatório incluído no preditor linear, σ^2, atribuímos uma distribuição *a priori* Gama Inversa com parâmetros a e b. As quantidades V_{β_j}, $\forall\ j$, V_{β_0}, a e b são fixadas. Com o modelo completamente especificado partiremos para o procedimento de inferência *a posteriori*. A função de verossimilhança associada ao problema é dada por:

$$\begin{aligned} \pi(Y|\beta, \sigma^2, X) &= \prod_{i=1}^{n} \left[(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (Y_i - \mathbf{X}_i\beta) \right\} \right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2} - ... - \beta_k X_{ik}) \right\}. \end{aligned}$$

A partir da regra de Bayes podemos obter a distribuição *a posteriori*, assim como as distribuições condicionais completas *a posteriori*, que serão proporcionais a distribuição conjunta:

$$\pi(\beta, \sigma^2, z, q, Y \mid X) = \pi(Y \mid \beta, \sigma^2, X)\pi(\beta \mid z)\pi(z \mid q)\pi(q)\pi(\sigma^2).$$

Para amostrar da distribuição *a posteriori* e obter estimativas dos parâmetros do modelo, utiliza-se o algoritmo *Gibbs sampling* que é um esquema iterativo de amostragem de uma cadeia de Markov cujo núcleo de transição é formado pelas distribuições condicionais completas (expressões omitidas). Abaixo é fornecido o algoritmo utilizado para a geração das amostras da distribuição *a posteriori*:

Passo 1 Faça $t = 0$ e escolha os valores iniciais $\beta_0^{(t)}, ..., \beta_k^{(t)}$; $\sigma^{2(t)}$; $q_1^{(t)}, ..., q_k^{(t)}$.

Passo 2 Calcule $q_1^{*(t)}, ..., q_k^{*(t)}$ condicionado em $\beta_0^{(t)}, ..., \beta_k^{(t)}$; $\sigma^{2(t)}$; $q_1^{(t)}, ..., q_k^{(t)}$.

Passo 3 Gere $z_j^{(t)}$ de uma distribuição Bernoulli($q_j^{*(t)}$) para $j = 1, ..., k$.

Passo 4 Gere o $\beta_0^{(t+1)}, ..., \beta_k^{(t+1)}$ e $\sigma^{2(t+1)}$:

- $\beta_0^{(t+1)}$ de $(\beta_0 \mid \beta_1^{(t)}, ..., \beta_k^{(t)}, \sigma^{2(t)}, Y, X)$,
- $\beta_1^{(t+1)}$ de $(\beta_1 \mid \beta_0^{(t+1)}, \beta_2^{(t)}, ..., \beta_k^{(t)}, \sigma^{2(t)}, z_1^{(t)} = r, Y, X)$,

  $\vdots$
- $\beta_k^{(t+1)}$ de $(\beta_k \mid \beta_0^{(t+1)}, ..., \beta_{k-1}^{(t+1)}, \sigma^{2(t)}, z_k^{(t)} = r, Y, X)$, $r = 0$ ou 1,
- $\sigma^{2(t+1)}$ de $(\sigma^2 \mid \beta_0^{(t+1)}, ..., \beta_k^{(t+1)}, Y, X)$.

Passo 5 Gere $q_j^{(t+1)}$ de $(q_j \mid z_1^{(t)}, ..., z_k^{(t)})$.

Passo 6 Faça $t = t + 1$ e volte ao Passo 2 até a convergência das cadeias.

Na próxima Seção, este algoritmo é empregado para analisar um conjunto de dados reais de interesse, envolvendo a avaliaçao de 38 marcas de vinhos.

3 Aplicação

A variável resposta, Y, de interesse representa a qualidade de 38 marcas de vinhos, a qual será linearmente relacionada com as seguintes covariáveis: claridade do vinho (X_1), aroma (X_2), corpo (X_3), sabor (X_4) e afinação (X_5). A partir da Figura 1 pode-se perceber que há uma relação linear mais marcada entre a variável Y e as variáveis aroma (X_2) e sabor (X_4) do que com as demais.

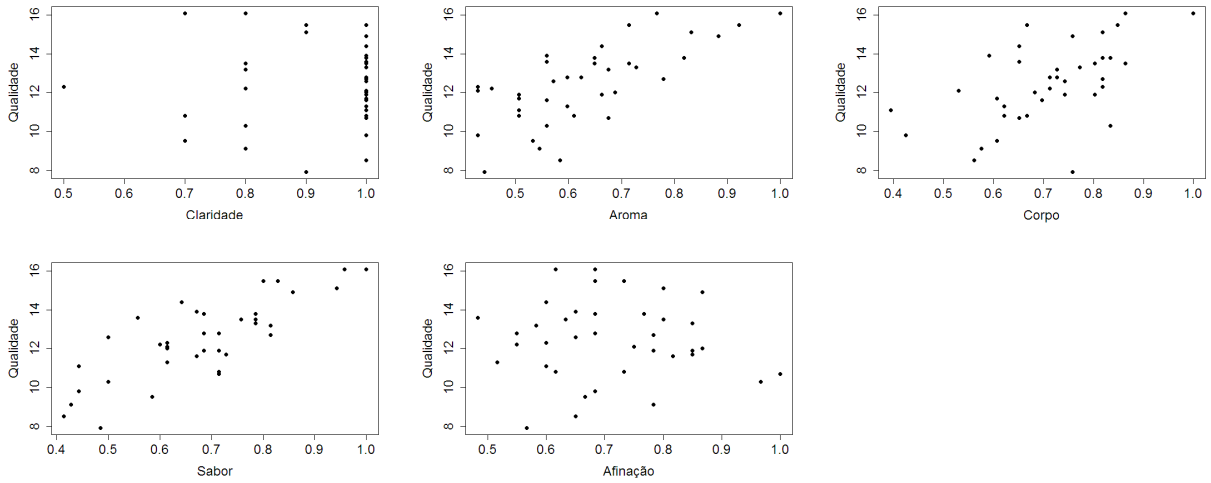

Figura 1: *Gráfico de dispersão da qualidade do vinho, Y, com as variáveis preditoras X_1-X_5, respectivamente, da esquerda pra direita em sentido horário.*

Inicialmente, por meio do pacote *leaps* do *software* `R`, consideramos os métodos usuais de seleção de variáveis como o coeficiente de determinação e o *stepwise*, a fim de identificar o melhor modelo para os dados em estudo. Diferentes modelos foram comparados a partir do coeficiente de determinação $R^2_{ajustado}$, sendo o melhor deles aquele que incluiu as covariáveis X_1, X_2, X_4 e X_5. Já o modelo com o segundo maior $R^2_{ajustado}$ inclui as covariáveis X_2, X_4 e X_5. Ao utilizarmos o método *stepwise* também obtivemos o modelo com as covariáveis X_2, X_4 e X_5. Na Tabela 1 apresentamos os cinco modelos com os maiores $R^2_{ajustado}$. Na coluna "Modelos" tem-se uma sequência de zeros e uns, em que 1 representa a inclusão da covariável e 0 indica que a variável não foi incluída. Por exemplo, (1,1,0,1,1) representa o modelo com o maior $R^2_{ajustado}$, o qual não inclui apenas a covariável X_3. Tais modelos podem ainda ser avaliados fazendo uma análise dos resíduos com o objetivo de verificar se cada um deles satisfazem as suposições básicas de normalidade e homocedasticidade do termo de erro. Uma breve análise dos gráficos dos resíduos foi feita e não revelaram nenhum grande problema para os dois modelos apontados com os maiores $R^2_{ajustado}$.

Tabela 1: Modelos ajustados para as variáveis preditoras

Modelos	$R^2_{ajustado}$
(1,1,0,1,1)	0.68012
(0,1,0,1,1)	0.67763
(1,1,1,1,1)	0.67694
(0,1,1,1,1)	0.66893
(1,0,1,1,1)	0.65602

Na sequência, a metodologia Bayesiana de seleção de modelos discutida na Seção 2 será empregada ao banco de dados sobre qualidade de vinhos. Foi feita uma transformação nas covariáveis dividindo-as pelo seu valor máximo antes de executar o algoritmo MCMC previamente apresentado. A execução foi feita com 30000 iterações, sendo que as 20000 primeiras foram consideradas como período de aquecimento das cadeias (*burn-in*) e retiradas da análise. A Figura 2 apresenta os gráficos das cadeias de β_j, $j = 0, ..., 5$ e σ^2 após o *burn-in*. Pode-se observar a convergência da cadeia de σ^2. Para os coeficientes β_1, β_2, β_3 e β_5, vê-se que as cadeias oscilam entre o valor zero e algum outro valor gerado da distribuição *a posteriori*, o que é explicado pelo processo de seleção de variáveis utilizado. A significância da covariável X_4 fica evidente pelo fato dos valores gerados para o parâmetro β_4 serem sempre diferentes de zero. Note que o comportamento das cadeias de β_1 e β_3 apresentam uma quantidade de valores zeros sendo maior em relação aos demais $\beta's$. Isto pode sugerir que as covariáveis associadas aos parâmetros β_1 e β_3 não sejam significativas para explicar a qualidade do vinho.

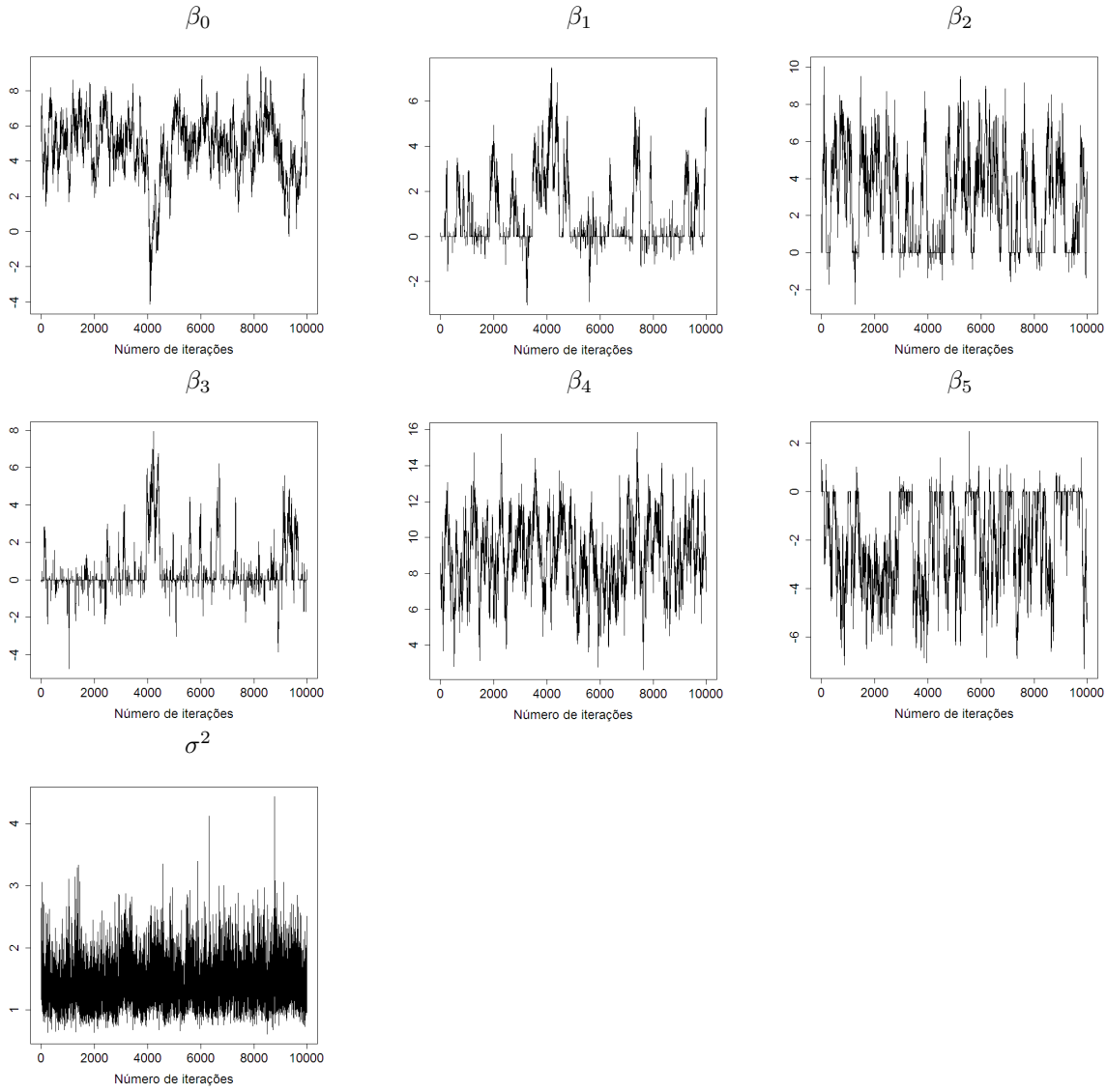

Figura 2: *Gráfico das cadeias de de* β_0, β_1 - β_5 *e* σ^2 *após o burn-in.*

A significância dos parâmetros pode ainda ser avaliada pelas estimativas médias *a posteriori* das probabilidade q_j^*, $j = 1, ..., 5$. Na Figura 3 podem ser observadas as cadeias e densidades das amostras de q_j^*. Os segmentos de retas apresentados nas verticais interceptam os valores 0.5 (reta pontilhada) e a média *a posteriori* (reta contínua). Veja que as densidades de q_2^*, q_4^* e q_5^* apresentam maior massa probabilística à direita e também médias *a posteriori* acima de 0.5. Por outro lado, as densidades de q_1^* e q_3^* apresentam maior massa probabilística à esquerda e médias *a posteriori* abaixo de 0.5. Isso nos dá indícios da significância das covariáveis X_2, X_4 e X_5.

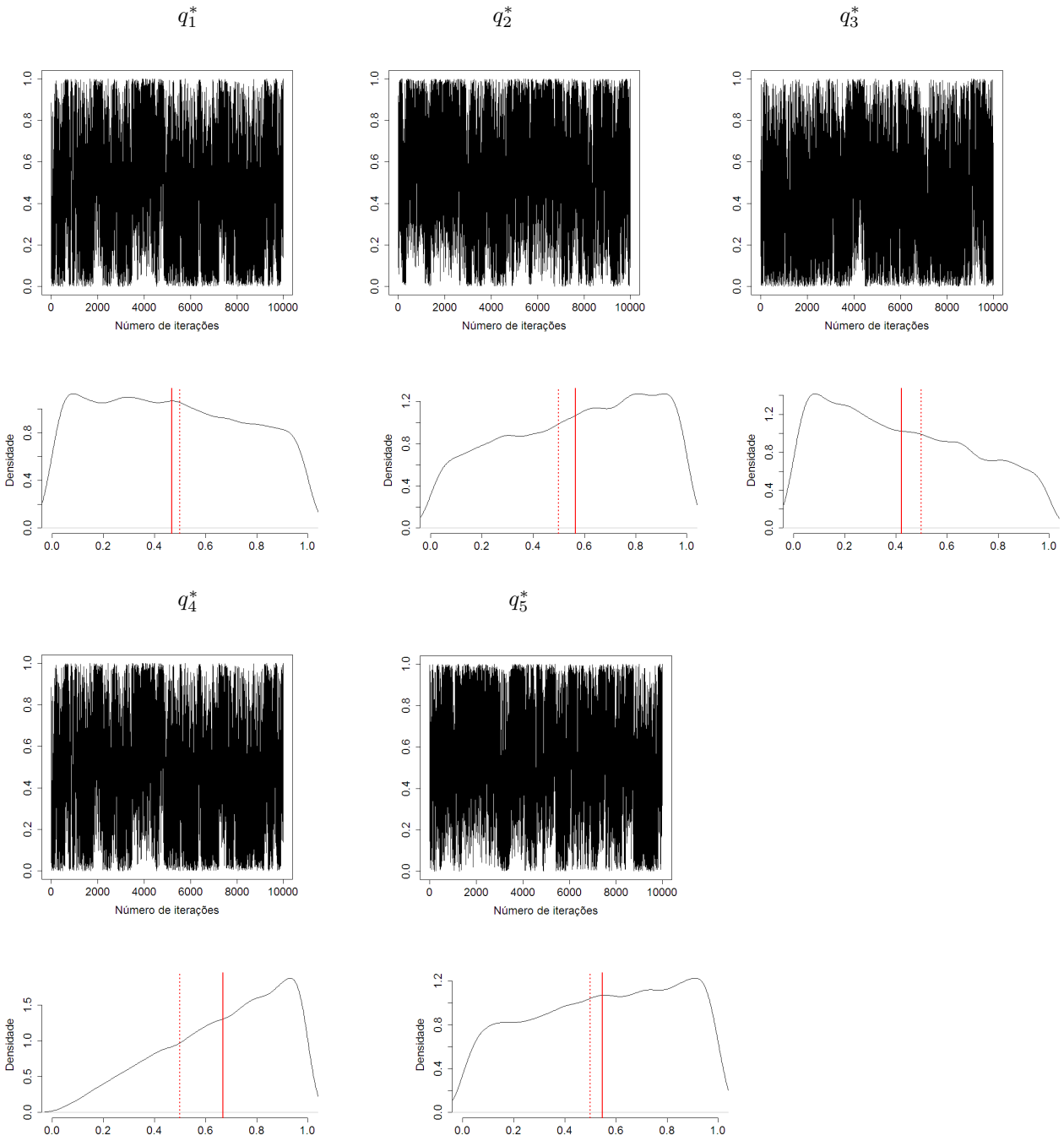

Figura 3: *Gráfico das cadeias e densidades de* q_1^* - q_5^* *após o burn in* .

A Tabela 2 apresenta algumas estatísticas *a posteriori* das mostras de q_j^* para $j = 1, ...5$ e também para os demais parâmetros. Nela, pode-se observar que os intervalos de credibilidade dos $q_{j's}^*$ são bastan-

tes amplos. Além disso, as estimativas pontuais da média *a posteriori* de q_2^*, q_4^* e q_5^* são maiores que 0.5, dando indícios de que as covariáveis X_2, X_4 e X_5 são de fato significativas para explicar a variabilidade da variável Y. Veja que este método de seleção de modelos corrobora os resultados obtidos com o método *stepwise* e também concorda com o segundo melhor modelo quando utilizamos o $R^2_{ajustado}$ como critério de seleção.

Tabela 2: Estatísticas *a posteriori*

Parâmetro	Média	Variância	HPD^*
q_1^*	0.4682	0.0702	[0.0004; 0.9407]
q_2^*	0.5651	0.0693	[0.0793; 1.0000]
q_3^*	0.4221	0.0655	[0.0001; 0.9199]
q_4^*	0.6695	0.0651	[0.2312; 1.0000]
q_5^*	0.5479	0.0686	[0.0726; 0.9996]
β_0	4.8054	3.1869	[1.3064; 8.3043]
β_1	2.2030	2.6981	[-1.0165; 5.4225]
β_2	0.5351	0.0645	[0.0374; 1.0328]
β_3	0.2507	0.0994	[-0.3672; 0.8687]
β_4	1.2866	0.0831	[0.7215; 1.8517]
β_5	-0.5224	0.0578	[-0.9935; -0.0512]
σ^2	1.4102	0.1271	[0.7952; 2.104265]

*HPD: *Highest Posterior Density interval*

4 Conclusão

Neste trabalho apresentamos o desenvolvimento e uma aplicação de um modelo de regressão esparso que incorpora os elementos para seleção de variáveis no contexto Bayesiano. A importância de usarmos distribuições *a priori* no formato de misturas com uma das componentes degeneradas em zero para os coeficientes de regressão, está no fato delas removerem o efeito de covariáveis estatisticamente insignificantes para explicar o fenômeno de interesse. Além do mais, esse procedimento Bayesiano para selecionar subconjuntos de covariáveis é muito elegante ao ponto de deixar que os dados determinem as probabilidades de inclusão *a posteriori* de cada coeficiente, indicando assim os modelos mais promissores. Em nosso trabalho sobre o ajuste da qualidade (Y) de 38 marcas de vinho, mostramos o uso dessa poderosa ferramenta que indicou um modelo formado pelas covariáveis aroma (X_2), sabor (X_4) e afinação (X_5). Este também foi o modelo indicado pelo método de seleção *stepwise* e o segundo melhor modelo sugerido pelo critério do $R^2_{ajustado}$.

Referências Bibliográficas

Chen, M. H. e Ibrahim, J. G. (2003), "Conjugate Priors for Generalized Linear Models," *Statistica Sinica*, pp. 461 – 476.

Chen, M. H., Huang, L., Ibrahim, J. G., e Kim, S. (2008), "Bayesian Variable Selection and Computation for Generalized Linear Model with Conjugate Priors," Tech. rep.

George, E. I. e McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Geweke, J. (1996), "Variable selection and model comparison in regression, in A. D. J.M. Bernardo, J.O. Berger and A. Smith, eds," in *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*, pp. 609–620.

Leamer, E. (1978), "Regression selection strategies and revealed priors," *Journal of the American Statistical Association*, pp. 580–587.

Mitchell, T. J. e Beauchamp, J. J. (1998), "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*.