Bayesian Analysis (2011) **6**, Number 1, pp. 1–24

Data Augmentation for Support Vector Machines

Nicholas G. Polson* and Steven L. Scott†

Statistical Learning – Insper

Guaraci Requena

18 de Março de 2016

- **Problema original:** maximizar a margem do hiperplano separador

$$\min_{\beta} \frac{||\beta||^2}{2}$$

sujeito às restrições

$$y_i x_i^T \beta \geq 1, \forall i$$

- $y_i \in \{-1, 1\}$; $x_i^T = (1, x_1, ..., x_{k-1})$

- **Margem soft:** maximizar a margem do hiperplano

$$\min_{\beta} \frac{||\beta||^2}{2} + C \sum_{i=1}^{n} \xi_i$$

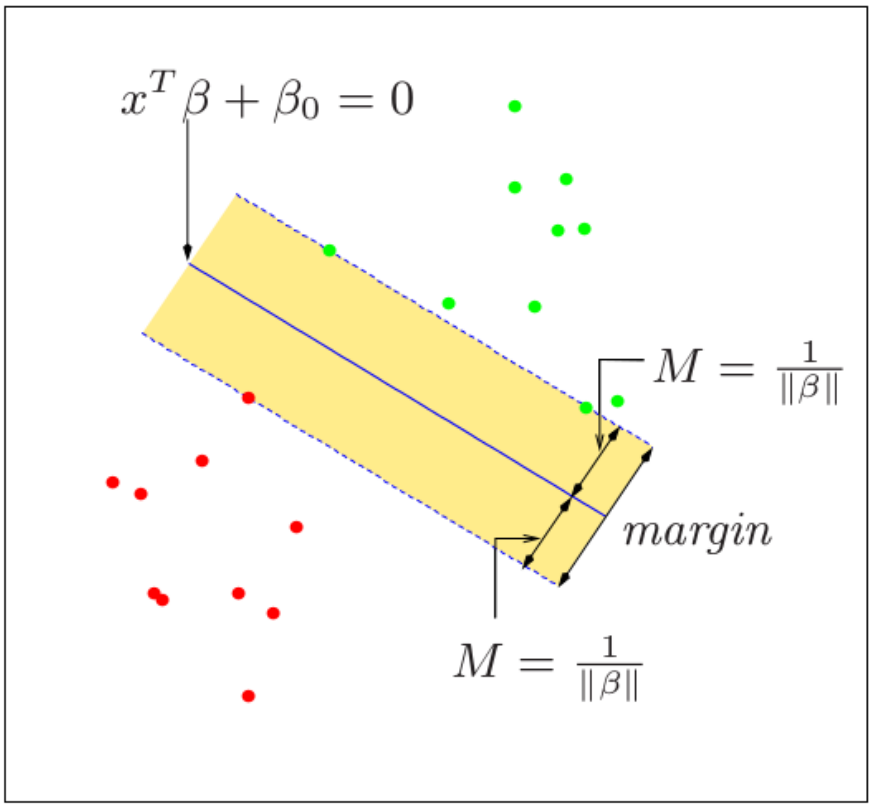
sujeito às restrições

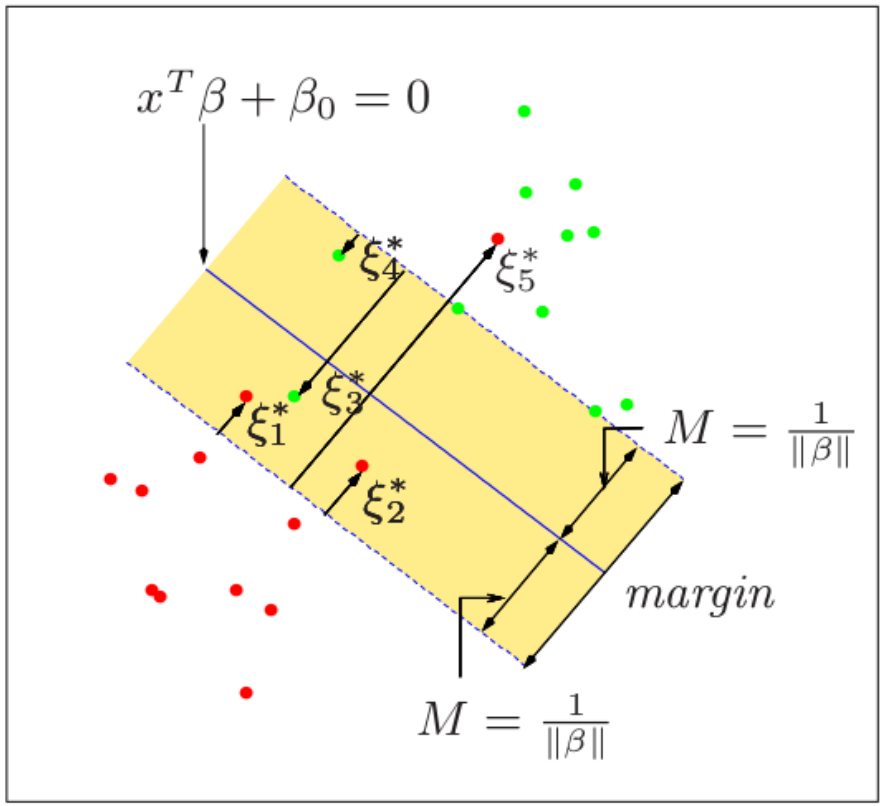
$$y_i x_i^T \beta \geq 1 - \xi_i, \forall i$$

em que $\xi_i \geq 0 \ \forall i$ e $C > 0$ é o parâmetro de custo

- Recuperamos a margem "hard" com $C \to \infty$

- Soft margin (p.418 "*The Elements of Statistical Learning*")





- $\xi_i \geq 1 - y_i x_i^T \beta$ e $\xi_i \geq 0$

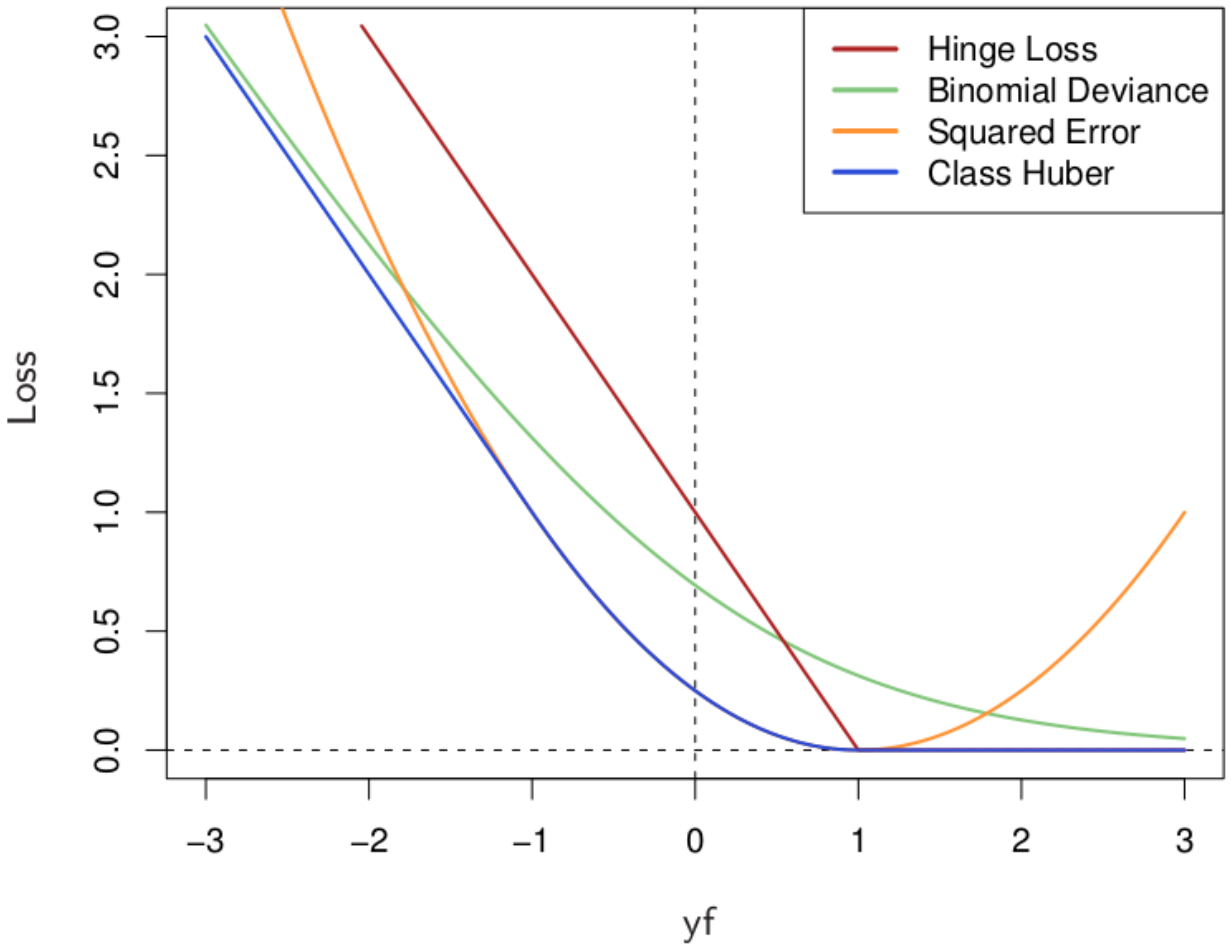
$$\xi_i = \max(0, 1 - y_i x_i^T \beta)$$

- O problema original equivale a minimizar

$$\sum_{i=1}^{n} \max(0, 1 - y_i x_i^T \beta) + \frac{\lambda}{2} ||\beta||^2$$

em que $\lambda = 1/C$

- Hinge loss (p.456 "*The Elements of Statistical Learning*")



- SVM regularizado: pênalti L^α

$$d_\alpha(\beta, \nu) = \sum_{i=1}^{n} \max(1 - y_i x_i^T \beta, 0) + \nu^{-\alpha} \sum_{j=1}^{k} \left| \frac{\beta_j}{\sigma_j} \right|^\alpha$$

- Agnóstico a expansões via kernel, assumindo que x_i já inclui todas as expansões desejadas
- σ_j é o desvio padrão do j-ésimo elemento de x
  - Mitchell and Beauchamp (1988)
  - George and McCulloch (1997)
  - Clyde and George (2004)
  - Fan and Li (2001)
  - Griffin and Brown (2005)
  - Holmes and Held (2006)
- ν é o parâmetro de *tuning*

- Minimizar

$$d_\alpha(\beta, \nu) = \sum_{i=1}^{n} \max(1 - y_i x_i^T \beta, 0) + \nu^{-\alpha} \sum_{j=1}^{k} \left| \frac{\beta_j}{\sigma_j} \right|^\alpha$$

é equivalente a maximizar

$$\begin{aligned} p(\beta|\nu, \alpha, y) &\propto \exp(-d_\alpha(\beta, \nu)) \\ &\propto L(y|\beta)p(\beta|\nu, \alpha) \end{aligned}$$

- $p(\beta|\nu, \alpha, y) \propto L(y|\beta)p(\beta|\nu, \alpha)$
- Pseudo-verossimilhança

$$L(y|\beta) = \prod_{i=1}^{n} L_i(y_i|\beta) = \prod_{i=1}^{n} \exp\left\{-2\max(1 - y_i x_i^T \beta, 0)\right\}$$

- Priori do pênalti (exponencial potência)

$$p(\beta|\nu, \alpha) = \prod_{j=1}^{k} p(\beta_j|\nu, \alpha) = \left(\frac{\alpha}{\nu\Gamma(1+\alpha^{-1})}\right)^k \exp\left(-\sum_{j=1}^{k}\left|\frac{\beta_j}{\nu\sigma_j}\right|^\alpha\right)$$

  - caso mais geral: $\alpha \in (0, 2]$
  - casos particulares: $\alpha = 2$ (ridge) e $\alpha = 1$ (lasso)

- o objetivo é aprender β de $p(\beta|\nu, \alpha, y)$

- Como amostrar de $p(\beta|\nu,\alpha,y) \propto L(y|\beta)p(\beta|\nu,\alpha)$?
- **Teorema 1:**

$$\begin{aligned} L_i(y_i|\beta) &= \exp(-2\max(1-y_i x_i^T\beta, 0)) \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left(-\frac{1}{2}\frac{(1+\lambda_i - y_i x_i^T\beta)^2}{\lambda_i}\right) d\lambda_i \\ &= \int_0^\infty \phi(1-y_i x_i^T\beta| - \lambda_i, \lambda_i) d\lambda_i \end{aligned}$$

ideia da prova:

Andrews & Mallows (1974):

$\int_0^\infty \frac{a}{\sqrt{2\pi\lambda}} e^{-\frac{1}{2}(a^2\lambda + b^2\lambda^{-1})} d\lambda = e^{-|ab|}, \forall a, b > 0$

$a = 1; b = u; \times e^{-u} \Rightarrow$

$\int_0^\infty \frac{a}{\sqrt{2\pi\lambda}} e^{-\frac{u^2}{2\lambda} - u - \frac{1}{2}\lambda} d\lambda = e^{-|u|-u} = e^{-2\max(u,0)}$

- Como amostrar de $p(\beta|\nu,\alpha,y) \propto L(y|\beta)p(\beta|\nu,\alpha)$?
- **Teorema 2:** (Pollard, 1946; West, 1987)

$$p(\beta_j|\nu,\alpha) = \int_0^\infty \phi(\beta_j|0,\nu^2\omega_j\sigma_j^2)p(\omega_j|\alpha)d\omega_j$$

$$p(\omega_j|\alpha) \propto \omega_j^{-\frac{3}{2}} St^+_{\frac{\alpha}{2}}(\omega_j^{-1})$$

- **Corolário 1:** (Andrews and Mallows, 1974) $\alpha = 1$

$$p(\beta_j|\nu,\alpha=1) = \int_0^\infty \phi(\beta_j|0,\nu^2\omega_j\sigma_j^2)\frac{1}{2}e^{-\frac{\omega_j}{2}}d\omega_j$$

$$\omega_j|\alpha = 1 \sim \mathcal{E}(2)$$

- $p(\beta|\nu,\alpha,y) \propto L(y|\beta)p(\beta|\nu,\alpha)$

$$
\begin{aligned}
p(\beta|\nu,\alpha,y) &\propto \prod_{i=1}^{n}\int_0^\infty \frac{1}{\sqrt{2\pi\lambda_i}}\exp\left(-\frac{1}{2}\frac{(1+\lambda_i-y_i x_i^T\beta)^2}{\lambda_i}\right)d\lambda_i \\
&\times \prod_{j=1}^{k}\int_0^\infty \phi(\beta_j|0,\nu^2\omega_j\sigma_j^2)p(\omega_j|\alpha)d\omega_j \\
&\propto \int \prod_{i=1}^{n}\left[\lambda_i^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{(1+\lambda_i-y_i x_i^T\beta)^2}{\lambda_i}\right)\right] \\
&\times \prod_{j=1}^{k}\left[\omega_j^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{\beta_j^2}{\nu^2\sigma_j^2\omega_j}\right)p(\omega_j|\alpha)\right]d\lambda d\omega
\end{aligned}
$$

- $p(\beta|\nu,\alpha,y) = \int p(\beta,\lambda,\omega|\nu,\alpha,y)d\lambda d\omega$

$$p(\beta,\lambda,\omega|\nu,\alpha,y) \propto \prod_{i=1}^{n}\left[\lambda_i^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{(1+\lambda_i-y_i x_i^T\beta)^2}{\lambda_i}\right)\right] \times \prod_{j=1}^{k}\left[\omega_j^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\frac{\beta_j^2}{\nu^2\sigma_j^2\omega_j}\right)p(\omega_j|\alpha)\right]$$

- Condicionais completas

1. $\beta|\lambda,\omega,\nu,y$
2. $\lambda_i|\beta,y_i$
3. $\omega_j|\beta_j,\nu,\alpha$

$$\propto p(\beta,\lambda,\omega|y,\nu,\alpha)$$

1. $\beta|\lambda, \omega, \nu, y$

- $p(\beta|\lambda, \omega, \nu, y) \propto p(\beta, \lambda, \omega|\nu, y)$

$$\beta|\lambda, \omega, \nu, y \sim \mathcal{N}(b, B)$$

$$B^{-1} = \nu^{-2}\Sigma^{-1}\Omega^{-1} + X^T\Lambda^{-1}X$$

$$b = BX^T(\mathbf{1} + \lambda^{-1})$$

- em que

$\Lambda_{n\times n} = diag(\lambda_1, \lambda_2, ..., \lambda_n)$
$\Omega_{k\times k} = diag(\omega_1, \omega_2, ..., \omega_k)$
$\Sigma_{k\times k} = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_k^2)$
$X_{n\times k} = \left(y_i x_i^T\right)$

- $\lambda_i|\beta, y$ e $\omega_j|\beta_j, \nu$ são expressos em termos da distribuição normal inversa
  $X \sim \mathcal{IG}(\mu, \tau) \Rightarrow p(x|\mu, \tau) = \sqrt{\frac{\tau}{2\pi x^3}} \exp\left(-\frac{\tau(x-\mu)^2}{2\mu^2 x}\right)$
- No fundo, tais distribuições são reconhecidas como normais inversas generalizadas
  $X \sim \mathcal{GIG}(\gamma, \psi, \chi) \Rightarrow$
  $p(x|\gamma, \psi, \chi) = C(\gamma, \psi, \chi) x^{\gamma-1} \exp\left(-\frac{1}{2}\left(\chi x^{-1} + \psi x\right)\right)$
- $X \sim \mathcal{GIG}\left(\frac{1}{2}, \tau, \frac{\tau}{\mu^2}\right) \Rightarrow X^{-1} \sim \mathcal{IG}(\mu, \tau)$

2. $\lambda_i|\beta, y_i$

- $p(\lambda|\beta, y) \propto p(\beta, \lambda, \omega|\nu, y)$
- **Corolário 2:** $\left(\lambda_i|\beta, y \sim \mathcal{GIG}\left(\frac{1}{2}, 1, (1 - y_i x_i^T \beta)^2\right)\right)$

$$\lambda_i^{-1}|\beta, y \sim \mathcal{IG}(|1 - y_i x_i^T \beta|^{-1}, 1)$$

3. $\omega_j|\beta_j, \nu, \alpha$

- Em geral, esta é uma distribuição complicada porque a densidade está, geralmente, somente disponível em termos de sua função característica
  - $\alpha = 2 \Rightarrow p(\omega_j|\alpha)$ é degenerada em 1
  - $\alpha = 1 \Rightarrow$
  - **Corolário 3** (do Corolário 1)
    $\left(\omega_j|\beta_j, \nu, \alpha = 1 \sim \mathcal{GIG}\left(\frac{1}{2}, 1, \frac{\beta^2}{\nu^2\sigma_j^2}\right)\right)$

$$\omega_j^{-1}|\beta_j, \nu, \alpha = 1 \sim \mathcal{IG}(\frac{\nu\sigma_j}{|\beta_j|}, 1)$$

Algoritmo MCMC-SVM (caso $\alpha = 1$)

1. Amostrar $\beta^{(g+1)}$ através de
   $\beta|\lambda^{(g)}, \omega^{(g)}, \nu, y \sim \mathcal{N}(b^{(g)}, B^{(g)})$
2. Amostrar $\lambda^{(g+1)}$ através de
   $\lambda_i^{-1}|\beta^{(g+1)}, y \sim \mathcal{IG}(|1 - y_i x_i^T \beta^{(g+1)}|^{-1}, 1)$
3. Amostrar $\omega^{(g+1)}$ através de
   $\omega_j^{-1}|\beta_j^{(g+1)}, \nu, \alpha = 1 \sim \mathcal{IG}\left(\frac{\nu\sigma_j}{\left|\beta_j^{(g+1)}\right|}, 1\right)$

- É possível aprender ν?
- $p(\nu|\beta,\alpha) \propto p(\beta|\nu,\alpha)p(\nu|\alpha)$
- $\nu^{-1} \sim \Gamma(a_\nu, b_\nu)$

Algoritmo MCMC-SVM (caso $\alpha = 1$)

4 Amostrar $\nu^{(g+1)}$ através de
$\nu^{-1}|\beta^{(g+1)} \sim \Gamma\left(a_\nu + k, b_\nu + \sum_{j=1}^{k} |\beta_j^{(g+1)}/\sigma_j|\right)$

- É possível aprender α?
- $p(\alpha|\beta,\nu) \propto p(\beta|\nu,\alpha)p(\alpha|\nu)$
- $p(\alpha|\nu) \propto 1, \alpha \in (0,2]$
- "*A somewhat more radical departure would be to simulate* α *from*"

$$p(\alpha|\beta,\nu) \propto \left(\frac{\alpha}{\Gamma(1+\alpha^{-1})}\right)^k \exp\left(-\sum_{j=1}^{k}\left|\frac{\beta_j}{\nu\sigma_j}\right|^{\alpha}\right)$$

- Pode-se amostrar de $p(\alpha|\beta,\nu)$ utilizando *slice sampler* (Neal, 2003)

- Uma estimativa Rao-Blackwell*izada* para β

$$E(\beta|y) = \frac{1}{G}\sum_{g=1}^{G} b^{(g)}$$

- O algoritmo MCMC não produzirá um modelo esparso
- Uma saída para tal é substituir a regularização L_1 por priori "*spike and slab*"
  - $\gamma_j = 1$ se $\beta_j \neq 0$ e $\gamma_j = 0$ caso contrário
  - Priori conveniente para γ: $p(\gamma) = \prod_{j=1}^{k} \pi_j^{\gamma_j}(1-\pi_j)^{(1-\gamma_j)}$
  - É comum escolher $\pi_1 = \cdots = \pi_k = \pi$
  - Uma priori mais informativa: $\pi = k_0/k$, k_0 é a crença a priori do número de coeficientes incluídos

- Prioris: $\gamma_j \sim Bernoulli(\pi_j)$ e $\beta_\gamma \sim \mathcal{N}\left(0, \nu^2 \left[\Sigma_\gamma^{-1}\right]^{-1}\right)$

Algoritmo MCMC-SVM (spike and slab)

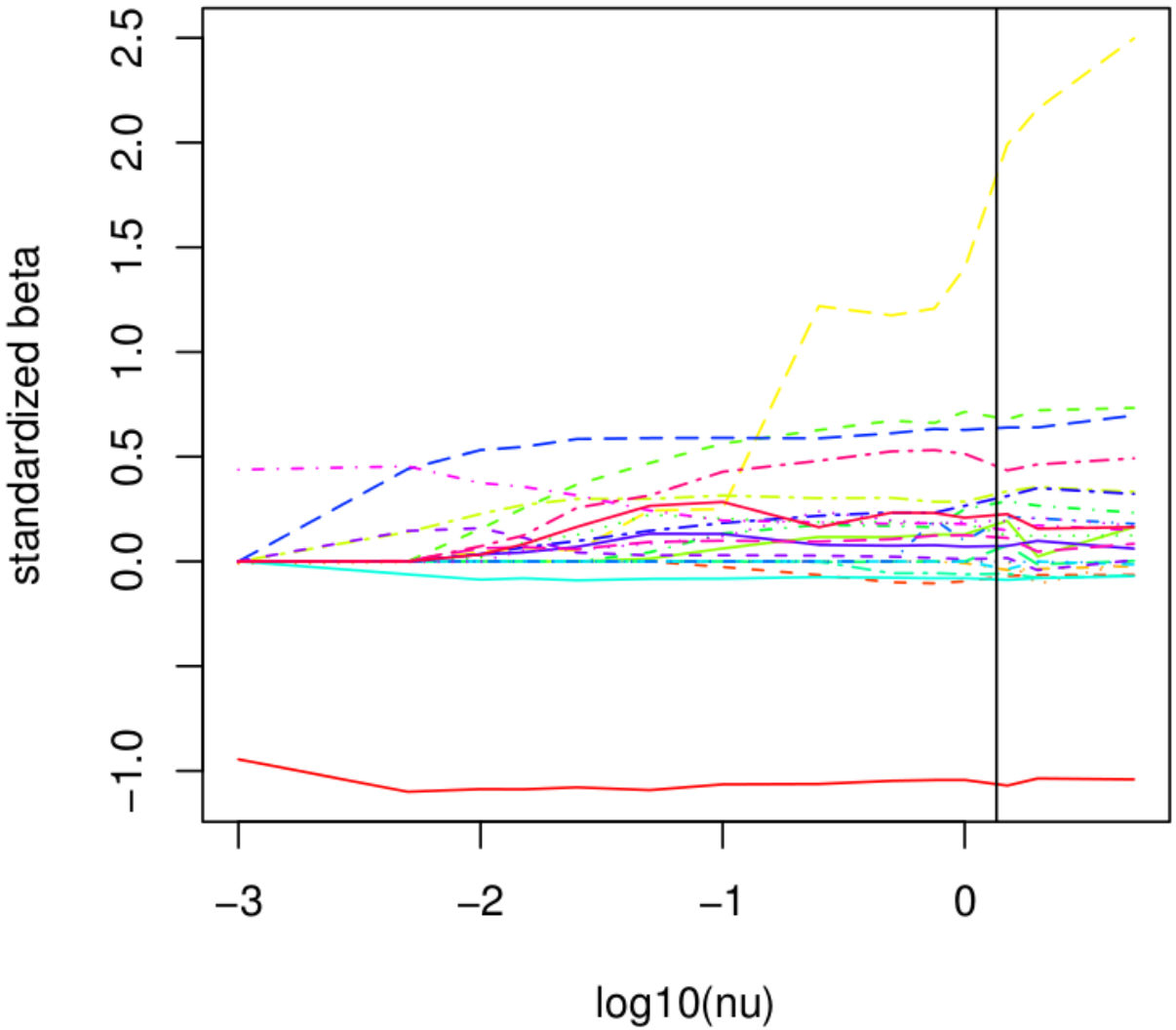
1. Amostrar $\lambda^{(g+1)}$ através de $\lambda_i^{-1}|\beta_\gamma^{(g)}, yi \sim \mathcal{IG}\left(|1 - y_i x_i^T \beta_\gamma^{(g)}|, 1\right)$
2. Amostrar γ através de $p(\gamma|y, \lambda^{(g+1)}, \nu) \propto p(\gamma)\frac{|\Sigma_\gamma^{-1}/\nu^2|^{1/2}}{|B_\gamma^{-1}|^{1/2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(1+\lambda_i - y_i x_{i,\gamma}^T b_\gamma)^2}{\lambda_i} - \frac{1}{2\nu^2} b_\gamma^T \Sigma_\gamma^{-1} b_\gamma\right)$
3. Amostrar $\beta_\gamma^{(g+1)}$ através de $\mathcal{N}\left(b_\gamma^{(g+1)}, B_\gamma^{(g+1)}\right)$

- Dados de spam (Hastie et al., 2009)

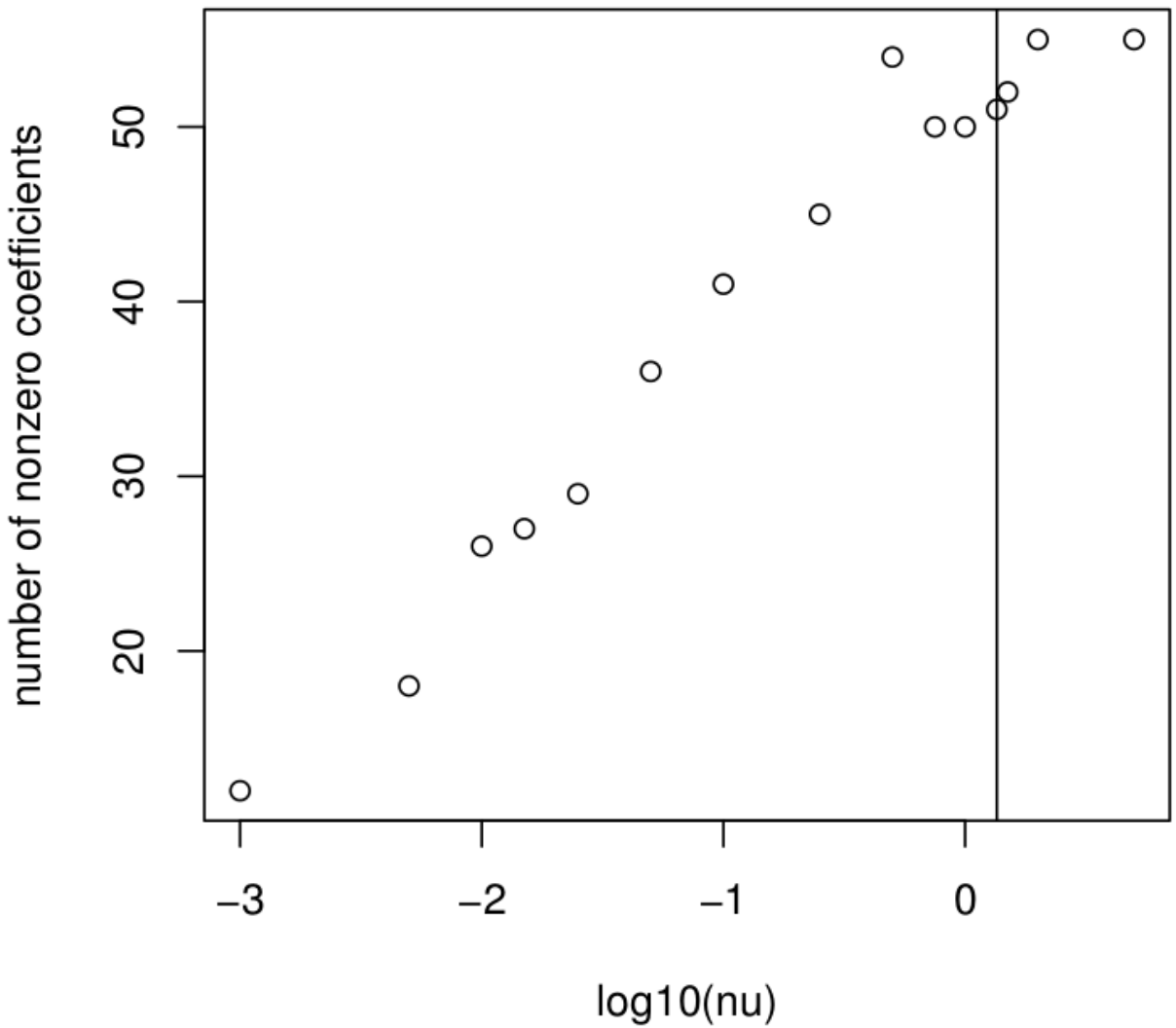
Exemplo: classificação de e-mails em *spam* e *ham* (1)

- Exemplo discutido na página 300 do ESL.
- Queremos classificar e-mails em duas categorias: *spam* (lixo eletrônico) e *ham* (legítimo).
- Temos 4601 e-mails, sendo que 1813 são marcados como *spam*; foram definidas 57 variáveis preditoras.
- Temos 48 preditoras quantitativas com as porcentagens das ocorrências de palavras específicas, tais como `business`, `address`, `internet`, `free` etc.
- Mais 6 preditoras quantitativas com as porcentagens das ocorrências de caracteres específicos, tais como `;`, `$`, `!` etc.
- O tamanho médio das sequências ininterruptas de letras maiúsculas (`CAPAVE`).
- O tamanho da maior sequência ininterrupta de letras maiúsculas (`CAPMAX`).
- A soma dos comprimentos das sequências ininterruptas de letras maiúsculas (`CAPTOT`).

- Coeficientes estimados (*lasso prior*)

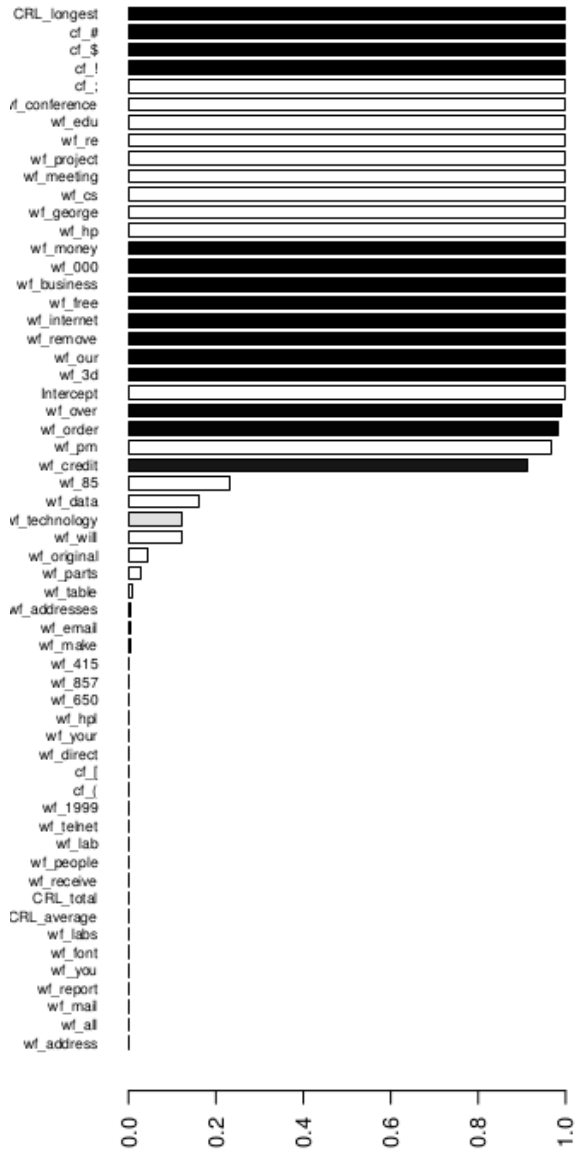


- $\nu \times$ "dimensão do modelo" (*lasso prior*)

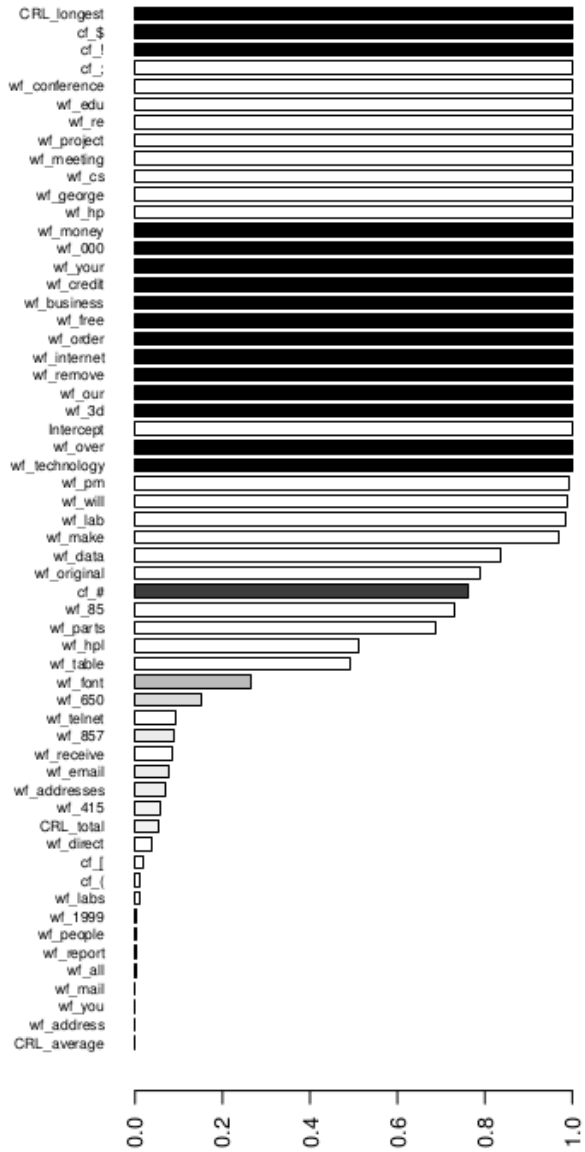


- Probabilidade de inclusão a posteriori (*spike and slab*)

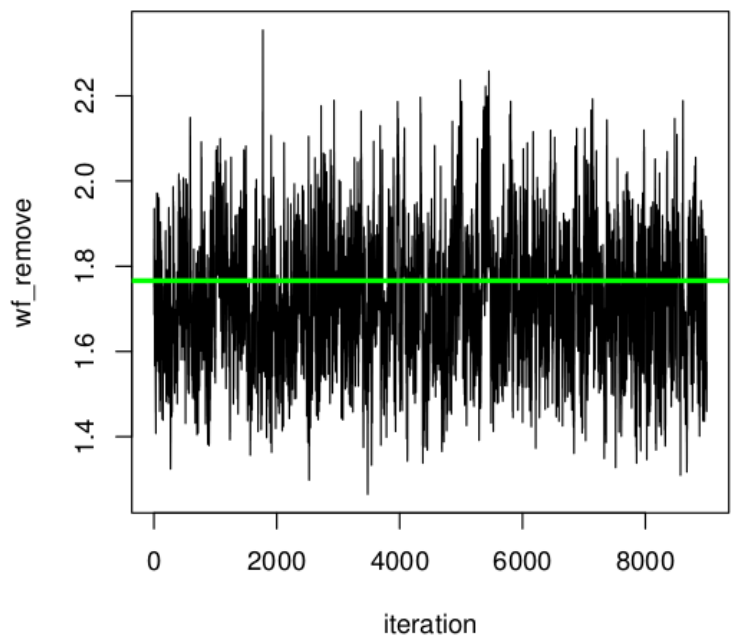
esparso ($\pi = 0.01$)

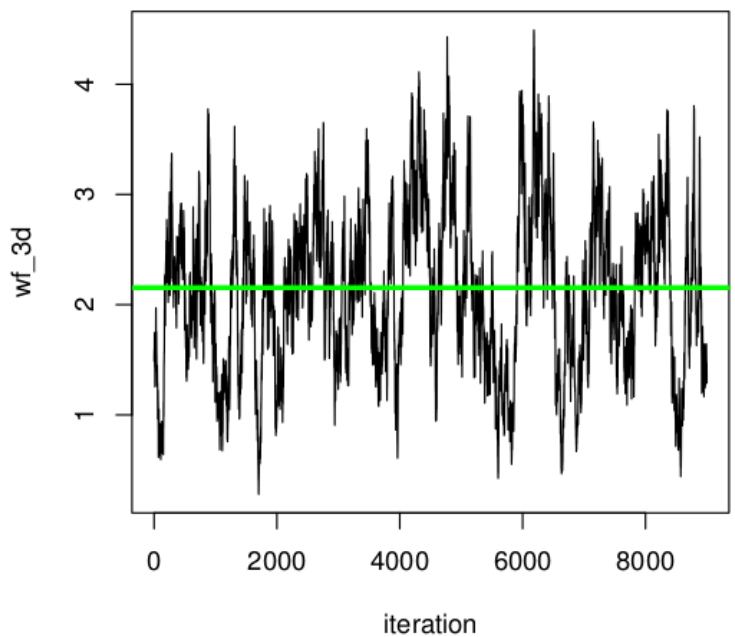


permissivo ($\pi = 0.5$)



- Trajetórias MCMC de 2 coeficientes que raramente "são" iguais a zero





- A função de perda hinge $\max(1 - y_i x_i^T \beta)$ para SVM parece tornar a análise Bayesiana complicada
- A representação de mistura de Polson & Scott para a pseudo-posteriori permite a análise Bayesiana
- Polson & Scott também desenvolvem no paper um algoritmo EM para estimar os coeficientes, sob a representação de mistura
- O uso de prioris *spike and slab* tem precedente de boa performance em problemas Bayesianos de seleção de variáveis. Performances similares são esperadas para SVM neste contexto
- A inclusão de incerteza nos coeficientes (no hiperplano)

- Andrews, D. F. and Mallows, C. L. (1974). "Scale Mixtures of Normal Distributions." Journal of the Royal Statistical Society, Series B: Methodological, 36: 99-102.
- Clyde, M. and George, E. I. (2004). "Model uncertainty." Statistical Science, 19: 81-94.
- Fan, J. and Li, R. (2001). "Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties." Journal of the American Statistical Association, 96(456): 1348?1360.
- George, E. I. and McCulloch, R. E. (1997). "Approaches for Bayesian Variable Selection." Statistica Sinica, 7: 339-374.
- Griffin, J. E. and Brown, P. J. (2005). "Alternative Prior Distributions for Variable Selection with very many more variables than observations."
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learn- ing. Springer, second edition.

- Holmes, C. C. and Held, L. (2006). "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression." Bayesian Analysis, 1(1): 145-168.
- Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian Variable Selection in Linear Regression (C/R: P1033-1036)." Journal of the American Statistical Association, 83: 1023-1032.
- Neal, R. M. (2003). "Slice Sampling." The Annals of Statistics, 31(3): 705-767.
- Pollard, H. (1946). "The representation of e?x as a Laplace integral." Bull. Amer. Math. Soc., 52(10): 908-910.
- West, M. (1987). "On Scale Mixtures of Normal Distributions." Biometrika, 74: 646-648.