The Bayesian Bridge

Lucas Tavares Short Cabral

19 de fevereiro de 2016

Seja $y = X\beta + \varepsilon$ para $\beta = (\beta_1, \ldots, \beta_p)$ desconhecidos. O estimador bridge $\hat{\beta}$ é o que minimiza a equação

$$Q_y(\beta) = \frac{1}{2}||y - X\beta||^2 + \nu \sum_{j=1}^{p} |\beta_j|^\alpha. \quad (1)$$

Com $\alpha = 1$ temos o LASSO, já para $\alpha = 2$ temos a regressão Ridge. A estimação é usualmente feita por meio do algoritmo EM.

Priori

A ideia do artigo é adotar uma perspectiva Bayesiana. Para isso, trata-se a moda global da $p(\beta|y) \propto exp\{-Q_y(\beta)\}$ como sendo a minimizadora da equação (1).
Assumindo uma verossimilhança Gausiana para y e uma priori β como um produto de exponênciais potência independentes:

$$p(\beta|\alpha,\nu) \propto \prod_{j=1}^{p} exp\left(-\left|\frac{\beta_j}{\tau}\right|^{\alpha}\right), \quad \tau = \nu^{-1/\alpha}. \tag{2}$$

Ao invés de minimiza a equação (1), amostra-se da distribuição posteriori conjunta de β e os hiperparâmetros do modelo.

1. Bridge Bayesiano vs Ridge Bayesiano e LASSO
2. Bridge Bayesiano vs Bridge Clássico
3. Bridge Bayesiano vs prioris que induzem esparsidade

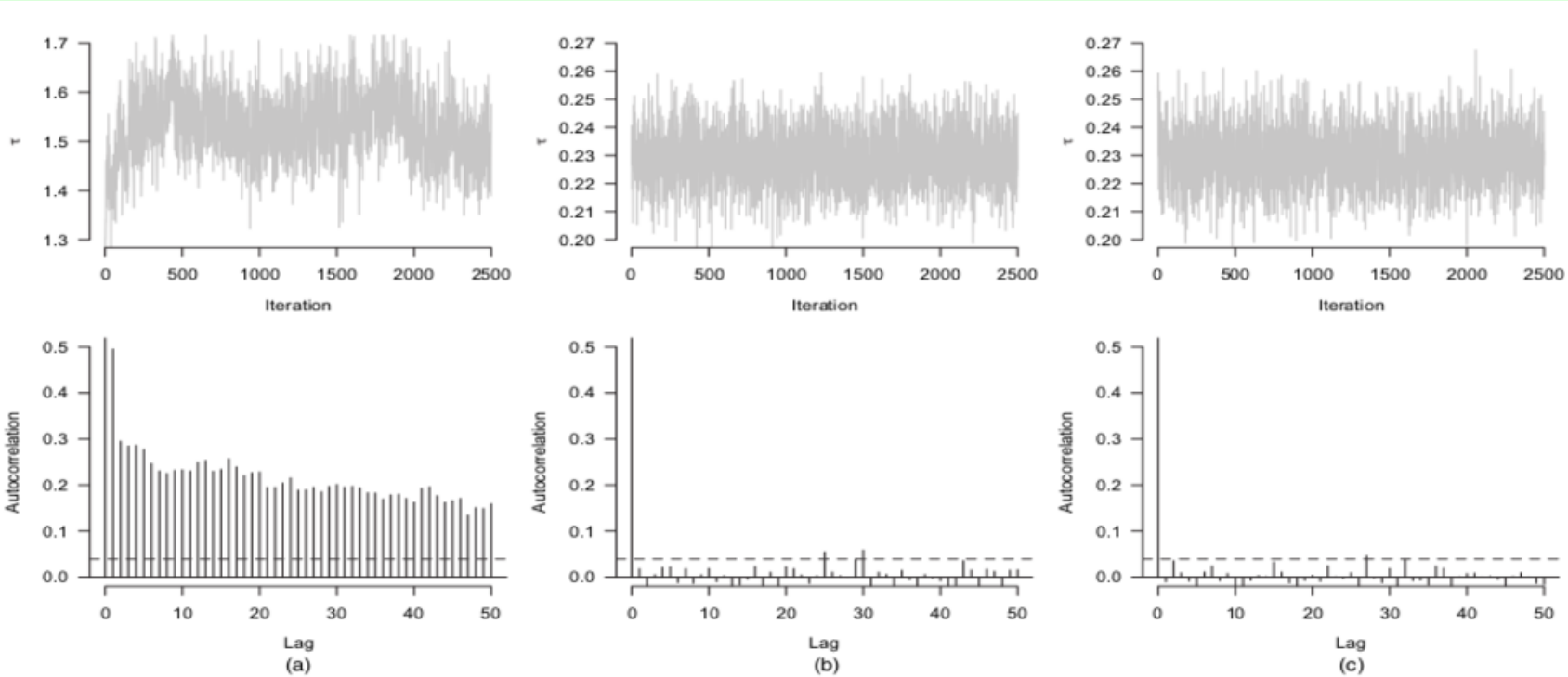


Fig. 1. Comparison of the simulation histories for τ, the global scale parameter, by using MCMC sampling for the bridge and the horseshoe on a 1000-dimensional orthogonal regression problem with $n = 1100$ observations (there were 100 non-zero entries in β simulated from a t_4-distribution, and 900 0s; because the priors have different functional forms, the τ-parameters in each model have a comparable role, but not a comparable scale, which accounts for the difference between the vertical axes): (a) horseshoe (with parameter expansion): (b) bridge (using Bartlett–Fejer kernels); (c) bridge (using normal mixtures)

Usando a representação de mistura de normais (West, 1987). Pode-se representar a exponencial potência como uma mistura de normais para um $\alpha \in (0, 2]$. Então,

$$exp(-|t|^{\alpha}) = \int_0^{\infty} exp(-st^2/2)g(s)ds. \quad (3)$$

A distribuição conjunta a partir de (1) e (3):

$$p(\beta, \Lambda|y) = C\, exp\left\{-\frac{1}{2}\beta'(\sigma^{-2}X'X + 2\nu^{2/\alpha}\Lambda)\beta + \beta'\sigma^{-2}X'y\right\} \prod_{j=1}^{p} p(\lambda_j), \quad (4)$$

onde $\Lambda = diag(\lambda_1, \ldots, \lambda_j)$ e $p(\lambda_j) = \lambda_j^{-1/2} g(\lambda_j)$, g denota a densidade estável do integrando da equação (3).

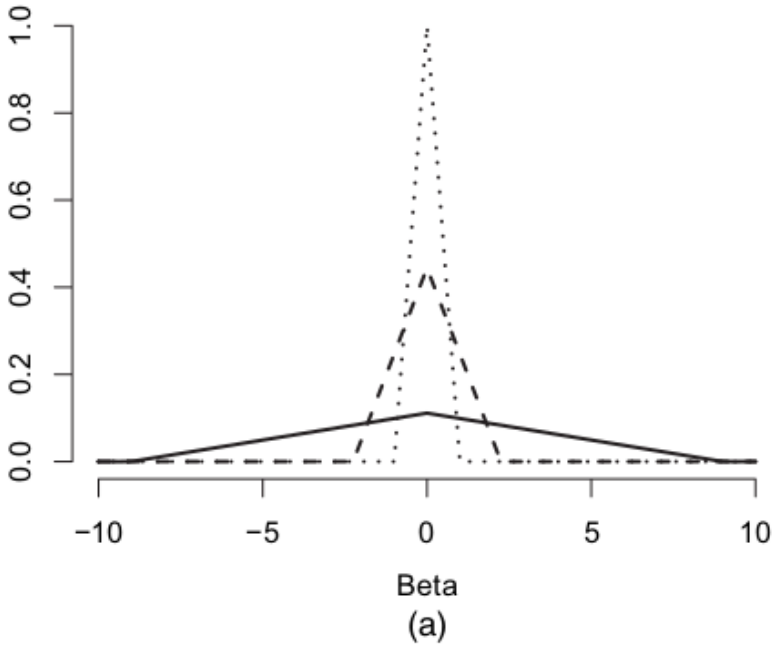
A posteriori condional λ_j dado β_j é um v.a. estável exponencialmente inclinada

$$p(\lambda_j|\beta_j) = \frac{exp(-\nu^{2/\alpha}\beta_j^2\lambda_j)p(\lambda_j)}{\mathbb{E}\{exp(-\nu^{2/\alpha}\beta_j^2\lambda_j)\}} \tag{5}$$

$$(y|\beta,\sigma^2) \sim N(X\beta,\sigma^2\mathbb{I}) \tag{6}$$

$$p(\beta_j|\tau,\omega_j,\alpha) = \frac{1}{\tau\omega_j^{1/\alpha}}\left\{1-\left|\frac{\beta_j}{\tau\omega_j^{1/\alpha}}\right|\right\}_+ \tag{7}$$

$$(\omega_j|\alpha) \sim \frac{1+\alpha}{2}\text{Gama}\left(2+\frac{1}{\alpha},1\right)+\frac{1-\alpha}{2}\text{Gama}\left(1+\frac{1}{\alpha},1\right) \tag{8}$$



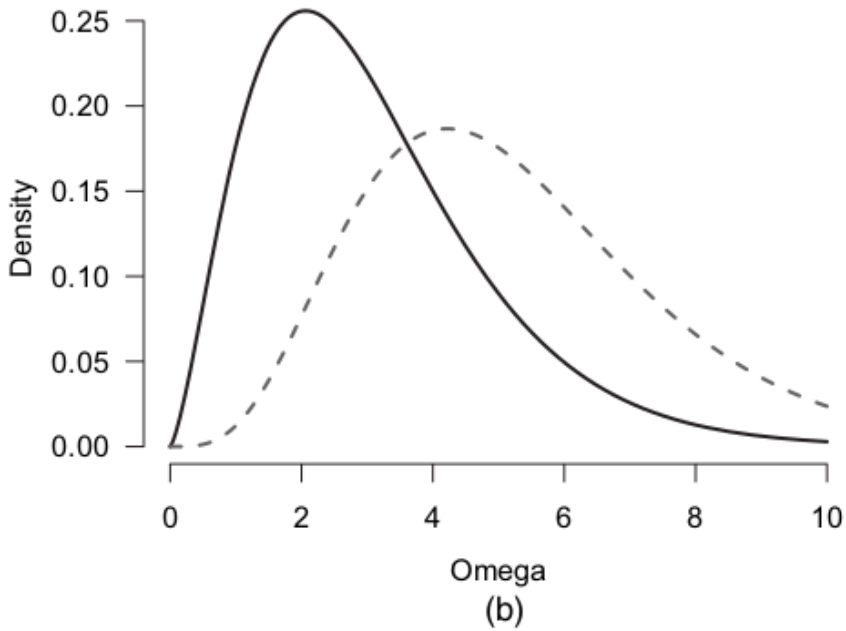


Fig. 2. (a) Triangular densities, or normalized Bartlett–Fejer kernels, of different widths ($\alpha = 0.5$) (· · · · · ·, $\omega = 1.0$; – – –, $\omega = 1.5$; ———, $\omega = 3.0$) and (b) two examples of mixing distributions for ω_j that give rise to exponential power marginals for β_j in conjunction with the Bartlett–Fejer kernel (———, $\alpha = 0.75$; – – –, $\alpha = 0.25$)

Table 1. Summary of performance for the two MCMC strategies in two design matrix scenarios†

Method	*Results for scenario 1 (strong collinearity)*			*Results for scenario 2 (orthogonal)*		
	Time (s)	*Median effective sampling rate*	*Minimum effective sampling rate*	*Time (s)*	*Median effective sampling rate*	*Minimum effective sampling rate*
Triangle	45.1	15	5	1.6	49559	36737
Normal	33.5	1816	536	5.5	17533	13048

†Scenario 1: diabetes data with all pairwise interactions and the design matrix on the original scale. Scenario 2: Boston housing data with an orthogonalized design matrix.

Amostrador de Gibbs:

Step 1: gerar $(u_j|\beta_j, \omega_j) \sim Unif(0, 1 - |\beta_j/\tau|\omega_j^{-1/\alpha})$

Step 2: gerar cada ω_j de uma mistura de gama truncada

Step 3: atualizar β de uma normal multivariada truncada proporcional a $N(\hat{\beta}, \sigma^2(X'X^{-1})\mathbb{I}(|\beta_j| \leq b_j), \quad \forall j.$

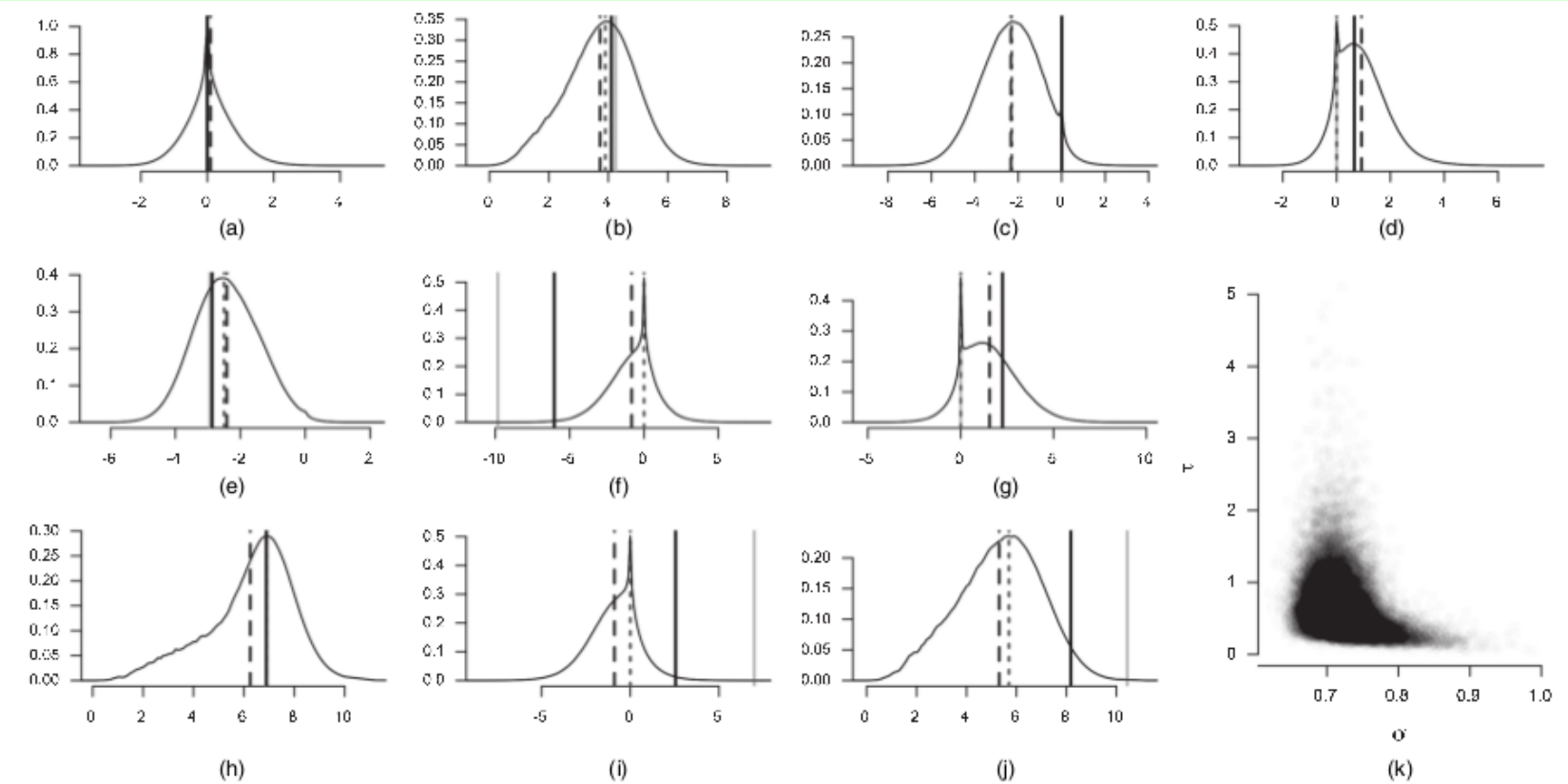


Fig. 5. Marginal posterior densities for the marginal effects of 10 predictors in the diabetes data (|, penalized likelihood solution with ν chosen by generalized cross-validation; ¦, result of stepwise Akaike information criterion selection starting from the full model; ¦, marginal posterior mean for β_j; ¦, mode of the marginal distribution for β_j under the fully Bayes posterior) (all predictors were standardized): (a) age; (b) blood pressure; (c) high density lipoprotein; (d) glucose; (e) female; (f) total cholesterol; (g) taurocholic acid; (h) body mass index; (i) low density lipoprotein; (j) triglyceride level; (k) joint posterior distribution of the scale components τ and σ

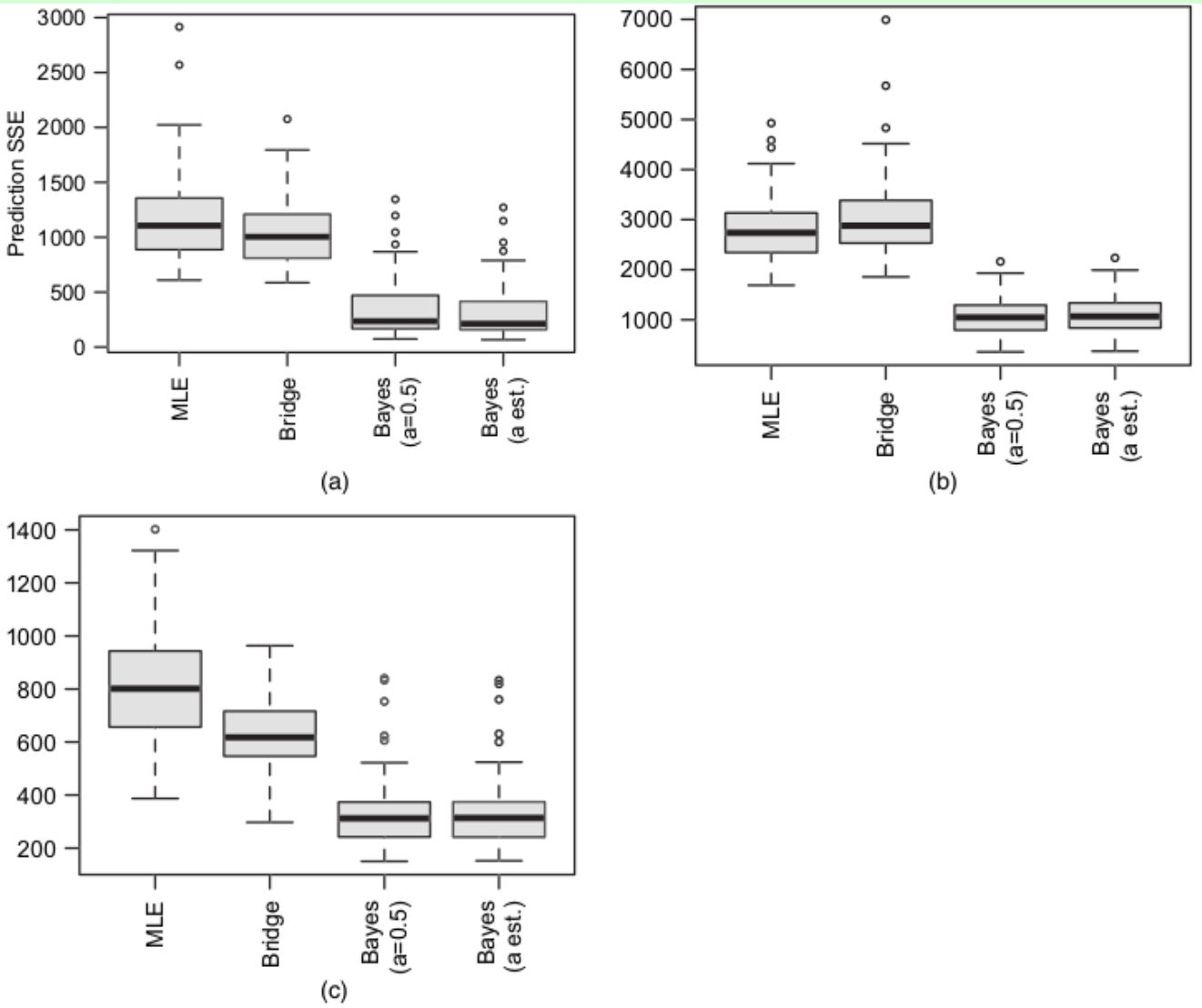


Fig. 6. Boxplots of the sum of squared errors in prediction hold-out data by using four methods for estimating β (maximum likelihood estimation, MLE, the classical bridge with $\alpha = 0.5$ and ν chosen by generalized cross-validation, the Bayesian bridge with $\alpha = 0.5$ and the Bayesian bridge with α estimated under a uniform prior): (a) Boston housing data; (b) near infrared glucose data; (c) ozone data