# Variational Inference for Bayesian Bridge Regression

*Zanini, C. T. P.*[1], *Migon, H. S.*[1], *Dias, R.*[2]

[1] UFRJ - Departamento de Métodos Estatísticos

[2] Unicamp - Instituto de Matemática Estatística e Ciência da Computação

**Abstract:** The bridge approach in regression models uses α-norm, with α>0, to define a penalization on large regression coefficients. Particular cases include lasso and ridge penalizations. This work implements Automatic Differentiation Variational Inference for Bayesian inference on semi-parametric regression models based on B-splines with bridge penalization. The inference procedure allows the use of small batches of data at each iteration, therefore drastically reducing computational time in comparison with MCMC. Full Bayesian inference is preserved so joint uncertainty estimates for all model parameters are available.

**Related publications:**

– Zanini C. T. P., Migon, H. S., Dias, R. *Variational Inference for Bayesian Bridge Regression.* **(under review)**

UFRJ UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

UNICAMP

4th Conference on Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

# Bayesian bridge semi-parametric regression

- Bridge penalization uses $\alpha$-norm:

$$\underset{\beta,\gamma}{\arg\min}\left\{\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\gamma)^{\top}(\mathbf{y}-\mathbf{X}\beta-\mathbf{Z}\gamma)+\lambda\sum_{j=1}^{p_x}|\beta_j|^{\alpha}\right\}$$

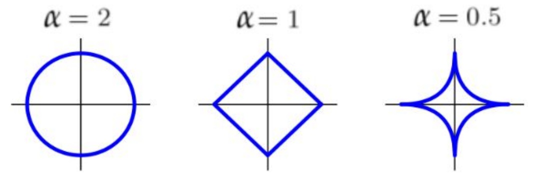
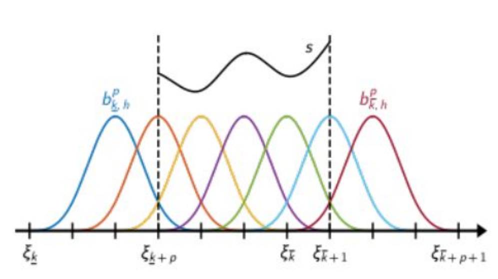



**Fig. 1**: Constraint regions $\sum_{i=1}^{p_x}|\beta_j|^{\alpha} \leq 1$ for different values of $\alpha$. Figure extracted from Hastie et al. (2015) (adapted).

- Penalized **spline** functions of covariates: $\mathbf{X}\beta$
- Unpenalized effects of covariates: $\mathbf{Z}\gamma$
- Higher $\lambda \Rightarrow$ smoother effects

Bayesian formulation:

$$(\mathbf{y} \mid \beta,\gamma,\phi) \sim N(\mathbf{X}\beta+\mathbf{Z}\gamma,\ \phi^{-1}\mathbf{I}_n),$$
$$(\beta_j \mid \lambda,\phi,\alpha) \overset{\text{iid}}{\sim} GG(0,\lambda^{-\frac{1}{\alpha}}\phi^{-\frac{1}{2}},\alpha),\ j=1,\ldots,p_x,$$
$$\phi \sim Ga(a_\phi,b_\phi),\gamma \sim N(\mu_\gamma,\Sigma_\gamma),\lambda \sim Ga(a_\lambda,b_\lambda),$$




UFRJ UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
UNICAMP



4th Conference on Statistics and Data Science
Salvador, Brazil (online)
December 1-3, 2022

# ADVI for Bayesian Bridge

- Original parameters:

$$\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi, \lambda, \alpha) \in \mathbb{R} \times \mathbb{R}^{+}\mathbb{R}^{+} \times (0, 2.5).$$

- Target: $p(\boldsymbol{\theta} \mid \mathbf{y})$
- Variational family: $Q = \{q_{\psi}(\boldsymbol{\theta});\ \ \psi \in \mathcal{V}\}$
- Objective: find $q_{\psi}(\boldsymbol{\theta})$ in $Q$ that is closest to $p(\boldsymbol{\theta} \mid \mathbf{y})$:

$$\underset{\psi \in \mathcal{V}}{\arg\min}\ KL\left(q_{\psi}(\boldsymbol{\theta}) \mid\mid p(\boldsymbol{\theta} \mid \mathbf{y})\right)$$

- Transformed parameters:

$$\boldsymbol{\xi} = T(\boldsymbol{\theta}) = (\boldsymbol{\beta}, \log \phi, \log \lambda, \text{logit}^{-1}(\alpha/2.5)) \in \mathbb{R}^{p+3}.$$

- Target: $\widetilde{p}(\boldsymbol{\xi} \mid \mathbf{y})$
- Variational distribution: $\widetilde{q}_{\psi}(\boldsymbol{\xi}) = N(\boldsymbol{\xi};\ \mathbf{m}, \mathbf{L}\mathbf{L}^{\top})$.
- Variational parameters: $\psi = (\mathbf{m}, \mathbf{L})$.
- Objective: find $q_{\psi}(\boldsymbol{\xi})$ in $Q$ that is closest to $p(\boldsymbol{\xi} \mid \mathbf{y})$:

$$\underset{\psi \in \mathcal{V}}{\arg\min}\ KL\left(\widetilde{q}_{\psi}(\boldsymbol{\xi}) \mid\mid \widetilde{p}(\boldsymbol{\xi} \mid \mathbf{y})\right)$$

- Optimization requires gradient of K.L.
- Stochastic gradients do not require reading full dataset at every iteration.
- Automatic Differentiation Variational Inference (ADVI) (Kucukelbir, 2017) allows full dependence structure in the variational family.

# Results and Conclusions

## Simulation scenarios

- 1,000 to 1,000,000 data points
- 30 bspline knots
- Very similar to MCMC but **much faster**!

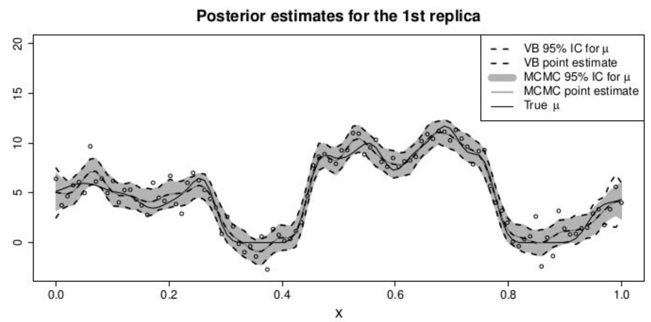
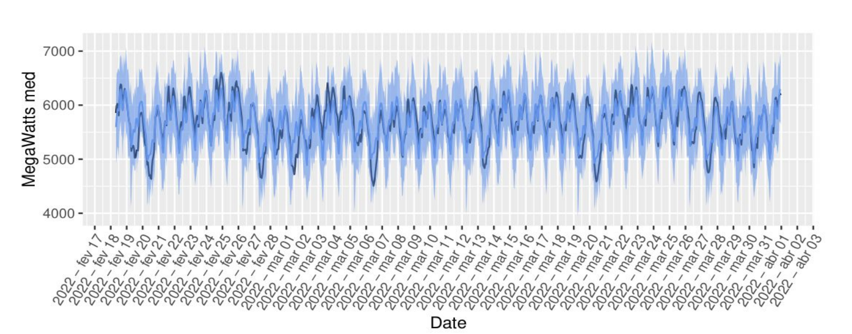



**Fig. 2:** ADVI and MCMC posterior confidence bands for the true B-spline curve $\mu = f_\beta(x)$ simulated in the first replica.

## Real data:

- Energy charges in Northern Brazil
- 1 bspline knot at every 100 hours (700 in total)
- 9717 hourly observations and 806 columns




- Proposed ADVI: less than **14min** to complete 2000 iterations with 100 M.C. samples to estimate gradients.
- MCMC: **19.23** hours to run the same 2000 iterations.

UFRJ UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

UNICAMP

4th Conference on **Statistics and Data Science**
Salvador, Brazil (online)
December 1-3, 2022