Variational Inference for Bayesian Bridge Regression

Carlos Tadeu Pagani Zanini - UFRJ
Ronaldo Dias - Unicamp
Helio dos Santos Migon - UFRJ

Main Goal

- Semi-parametric Bayesian model with bridge penalty
- MCMC has high computational cost in bridge models for large dataset.
  Ex.: Mallick and Yi (2018), Polson et al (2014).
- Currently, Variational Inference for Bayesian Bridge approximates the posterior by an independent joint distribution (Mean Field) $\Rightarrow$ poor estimates!
- Automatic Differentiation Variational Inference (ADVI) by Kucukelbir et al. (2017) is more flexible!
- Proposal: Adapt/modify ADVI for Bayesian semi-parametric regression with bridge penalty.
- Non-parametric component modeled by B-splines.

Bridge Regularization

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \sum_{j=1}^{D} \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}$$

$$\underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\arg\min} \quad \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}) + \lambda \sum_{\ell=1}^{p_x} |\beta_\ell|^\alpha$$

- Response variable: $\mathbf{y} \in \mathbb{R}^n$
- Covariates with penalized effects: $\mathbf{X}$ (dimension $n \times p_x$)
- Penalized coefficients: $\boldsymbol{\beta} \in \mathbb{R}^{p_x}$
- Unpenalized coefficients: $\mathbf{Z}$ (dimension $n \times p_z$)
- Non-penalized coefficients: $\boldsymbol{\gamma} \in \mathbb{R}^{p_z}$
- Regularization parameter: $\lambda > 0$
- Type of Penalization: $\alpha > 0$.

Bridge penalization parameter: α

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

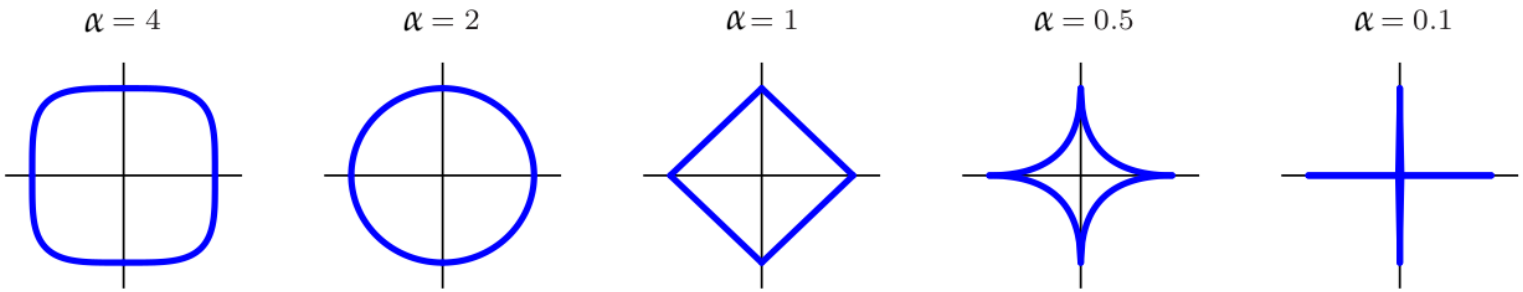


Figure: Ball $\sum_{j-1}^{p_x} |\beta_j|^\alpha \leq 1$ for different values of α ($p_x = 2$). For $\alpha < 1$, non-convex region. Figure from Hastie et. al (2015).

Bayesian Bridge Model

Bayeisan Model:

$$(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \phi) \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},\ \phi^{-1}\mathbf{I}_n),$$

$$(\beta_j \mid \lambda, \phi, \alpha) \overset{\text{iid}}{\sim} GG(0, \lambda^{-\frac{1}{\alpha}}\phi^{-\frac{1}{2}}, \alpha),$$

Generalized Gaussian prior for bridge penalty:

$$p(\boldsymbol{\beta} \mid \lambda, \phi, \alpha) = \prod_{j=1}^{p_x} \frac{\alpha \exp\left\{-\lambda\left(\phi^{\frac{1}{2}}|\beta_j|\right)^{\alpha}\right\}}{2\lambda^{-\frac{1}{\alpha}}\phi^{-\frac{1}{2}}\Gamma(\alpha^{-1})}$$

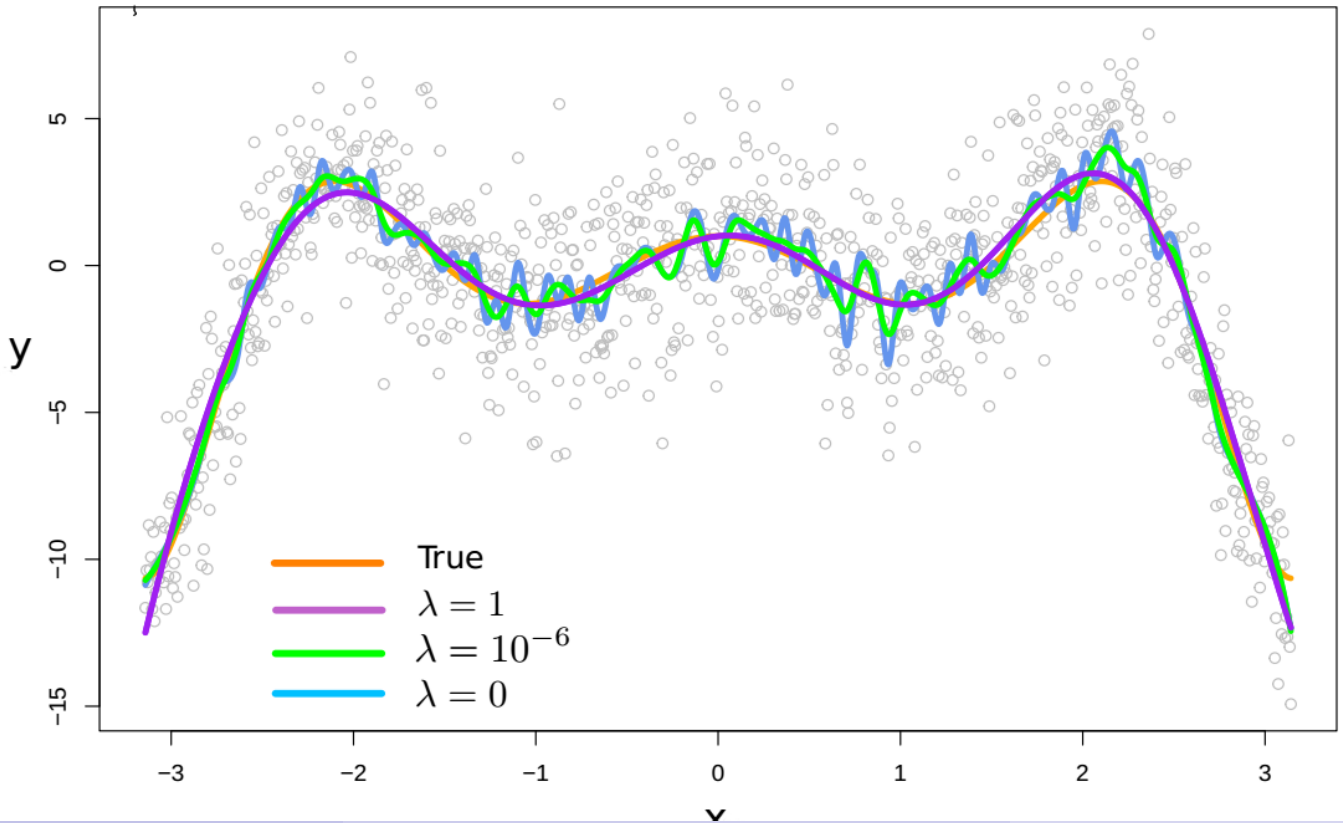
MAP is equivalent to the frequentist formulation:

$$\begin{aligned}
\text{MAP} &:= \arg\max_{\boldsymbol{\beta},\boldsymbol{\gamma}} \ p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y}, \phi, \alpha, \lambda) \iff \arg\max_{\boldsymbol{\beta},\boldsymbol{\gamma}} p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \phi) \times p(\boldsymbol{\beta} \mid \lambda, \phi, \alpha) \\
&= \arg\max_{\boldsymbol{\beta},\boldsymbol{\gamma}} \ -\frac{\phi}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}) - \lambda\phi^{\frac{\alpha}{2}}\sum_{j=1}^{p_x}|\boldsymbol{\beta}_j|^{\alpha} \\
&= \arg\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \ \frac{\phi}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}) + \lambda\phi^{\frac{\alpha}{2}}\sum_{j=1}^{p_x}|\boldsymbol{\beta}_j|^{\alpha} \\
&= \arg\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \ \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma})^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}) + \lambda^{*}\sum_{j=1}^{p_x}|\boldsymbol{\beta}_j|^{\alpha} \qquad (\lambda^{*} = \lambda\phi^{\frac{\alpha}{2}})
\end{aligned}$$

Regularization

- The penalty is important to avoid overfitting.
- The larger λ, the smoother the fitted curve is.

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



B-splines in a nutshell

- Curves form a basis for building functions via linear combinations.
- Each curve is formed by 4 cubic functions glued together smoothly.
- The curves are functions of a covariate $\Rightarrow$ Add columns in covariates matrix $\mathbf{X}$.

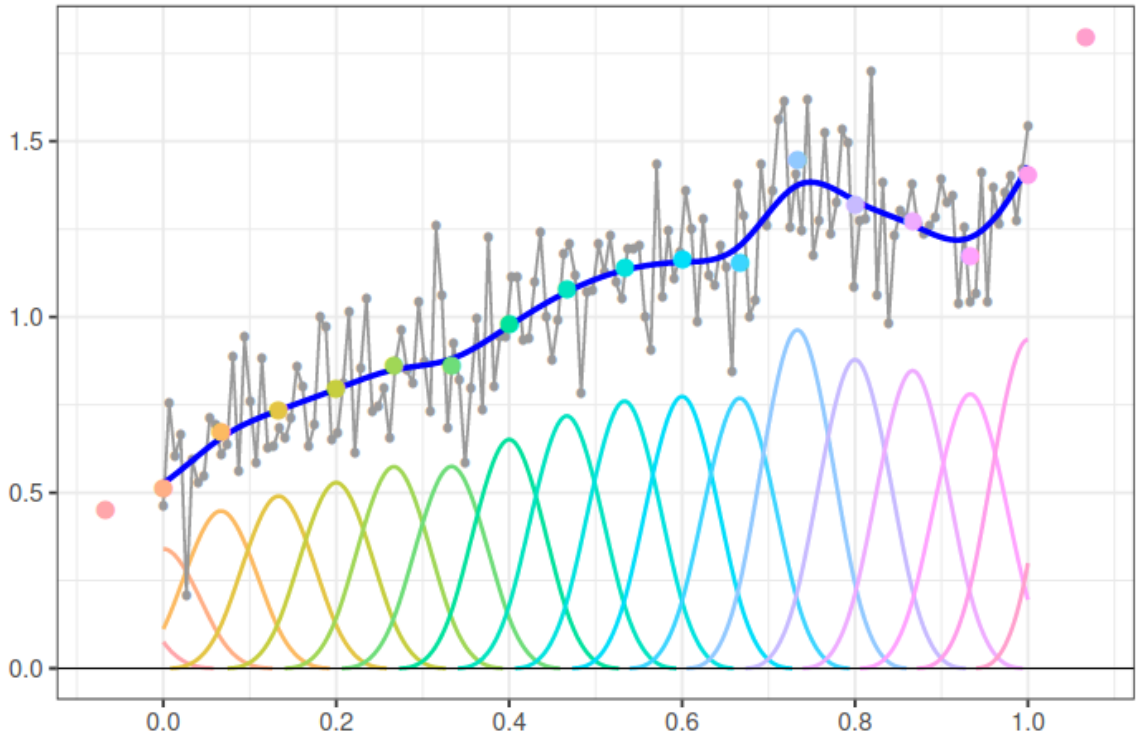


Figure: Illustration of a basis of B-splines. Figure extracted from Eilers and Marx (2021).

Variational Inference

- Target Distribution: $p(\boldsymbol{\theta} \mid \mathbf{y})$ (posterior)
- Variational Family : $Q = \{q_{\boldsymbol{\psi}}(\boldsymbol{\theta});\ \boldsymbol{\psi} \in \mathcal{V}\}$
- Objective: Find $q_{\boldsymbol{\psi}}(\boldsymbol{\theta})$ in Q that best approximates $p(\boldsymbol{\theta} \mid \mathbf{y})$:

$$\arg\min_{\boldsymbol{\psi}\in\mathcal{V}} KL\left(q_{\boldsymbol{\psi}}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta} \mid \mathbf{y})\right) \Leftrightarrow$$

$$\Leftrightarrow \arg\max_{\boldsymbol{\psi}\in\mathcal{V}} \underbrace{\mathbb{E}_{\boldsymbol{\theta}\sim q_{\boldsymbol{\psi}}(\boldsymbol{\theta})}\{[\log p(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}) - \log q_{\boldsymbol{\psi}}(\boldsymbol{\theta})]\}}_{ELBO(\mathbf{y},\boldsymbol{\psi})}.$$

Method limitations

- Optimization requires gradients
  $\nabla_{\boldsymbol{\psi}} ELBO(\mathbf{y}, \boldsymbol{\psi}) =$
  $\nabla_{\boldsymbol{\psi}} \mathbb{E}_{\boldsymbol{\theta} \sim q_{\boldsymbol{\psi}}(\boldsymbol{\theta})} \{[\log p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) - \log q_{\boldsymbol{\psi}}(\boldsymbol{\theta})]\},$
  which are often not available analytically.
- Mean-field hypothesis:
  $q_{\boldsymbol{\psi}}(\boldsymbol{\theta}) = \prod_j q_{\boldsymbol{\psi}_j}(\boldsymbol{\theta}_j)$ (independence).
- Flexible variational families do not admit analytic solution for $\nabla_{\boldsymbol{\psi}} ELBO(\mathbf{y}, \boldsymbol{\psi})$.
- Approximate ELBO via Monte Carlo!

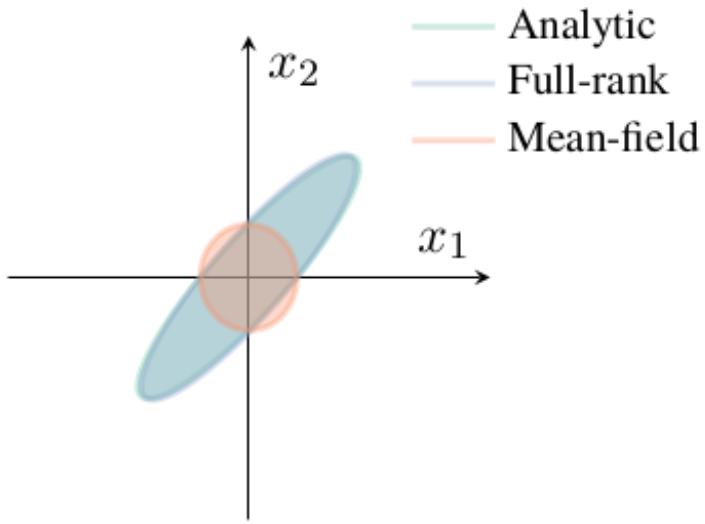


Figure: Example of Gaussian variational family with a mean-field hypothesis. Figure extracted from Kucukelbir et al (2017).

Exemplo: modelo de regressão logística

$$(y_i \mid \mathbf{x}_i, \beta_0, \ldots, \beta_d) \sim Bernoulli(p_i)$$

$$p_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_d x_{id})}$$

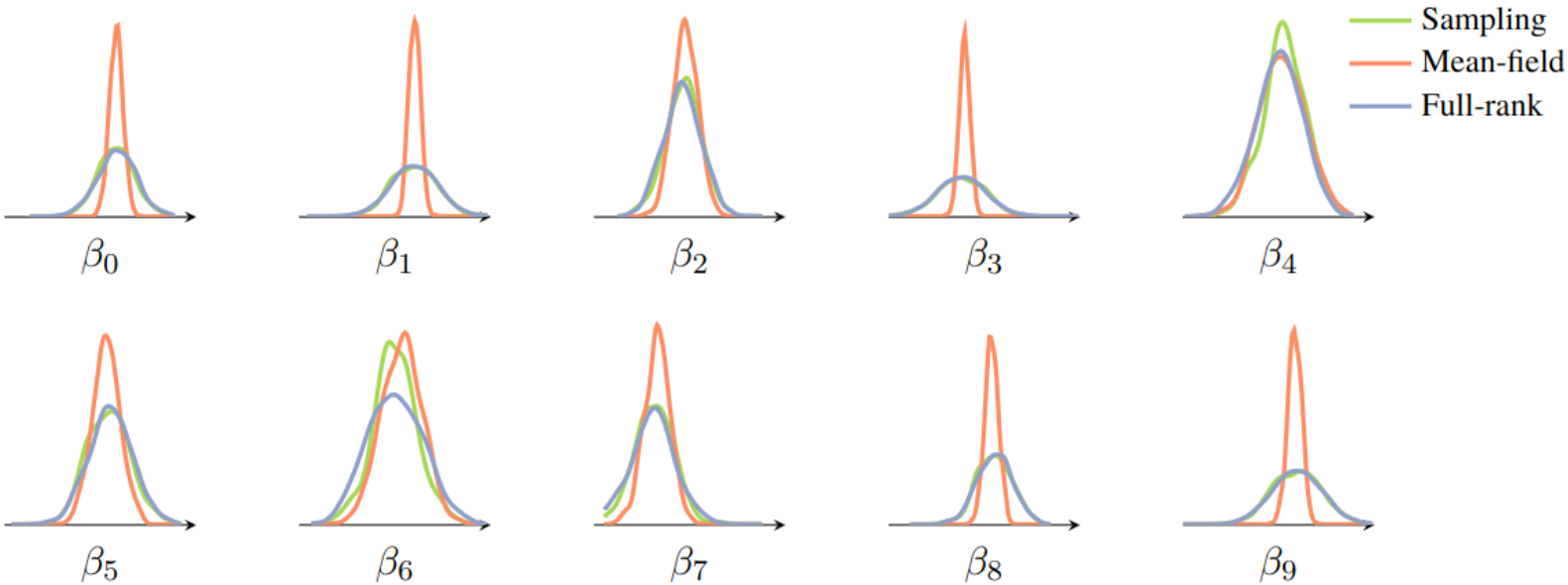


Figure 5: Comparison of marginal posterior densities for a logistic regression model. Each plot shows kernel density estimates for the posterior of each coefficient using 1000 samples. Mean-field ADVI underestimates variances for most of the coefficients.

Automatic Differentiation Variational Inference (ADVI)

- Parameter $\boldsymbol{\theta}$ typically has a support $\boldsymbol{\Theta} \neq \mathbb{R}^d$.
- Changing variables: $\boldsymbol{\xi} = T(\boldsymbol{\theta})$ such that $\boldsymbol{\xi}$ has support $\mathbb{R}^d$.
- The joint distribution is reparametrized in terms of $\boldsymbol{\xi}$:

$$\begin{aligned}\tilde{p}(\mathbf{y}, \boldsymbol{\xi}) &= p(\mathbf{y}, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=T^{-1}(\boldsymbol{\xi})} \times |J_{T^{-1}(\boldsymbol{\xi})}| \\ &= p(\mathbf{y}, T^{-1}(\boldsymbol{\xi})) \times |J_{T^{-1}(\boldsymbol{\xi})}|.\end{aligned}$$

- The variational distribution $q_{\boldsymbol{\psi}}(\boldsymbol{\xi}) = N(\boldsymbol{\xi}; \mathbf{m}, \mathbf{L}\mathbf{L}^\top)$ is specified in the transformed parameters.
- Variational parameters $\boldsymbol{\psi} = (\mathbf{m}, \mathbf{L})$.
- Reparameterization: $\boldsymbol{\xi} = \mathbf{m} + \mathbf{L}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})$.
- ADVI takes $\nabla_{\boldsymbol{\psi}} ELBO(\mathbf{y}, \boldsymbol{\psi})$ as an expected value to be approximated via Monte Carlo.

ADVI - ascending gradient method

- Reparametrization: $\boldsymbol{\xi} \sim N(\mathbf{m}, \mathbf{L}^\top\mathbf{L}) \Rightarrow \boldsymbol{\xi} = \mathbf{m} + \mathbf{L}\boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})$.
- Reparameterization allows "passing the derivative inside the expectation":

$$\nabla_{\boldsymbol{\psi}} ELBO(\mathbf{y}, \boldsymbol{\psi}) = \nabla_{\boldsymbol{\psi}} \mathbb{E}_{q_{\boldsymbol{\psi}}(\boldsymbol{\xi})} \left[\log \tilde{p}(\mathbf{y}, \boldsymbol{\xi}) - \log q_{\boldsymbol{\psi}}(\boldsymbol{\xi})\right]$$

$$= \nabla_{\boldsymbol{\psi}} \mathbb{E}_{\boldsymbol{\epsilon}\sim N(\mathbf{0},\mathbf{I})} \left[\log p(\mathbf{y}, T^{-1}(\boldsymbol{\xi})) + \log \mid J_{T^{-1}(\boldsymbol{\xi})} \mid - \log N(\boldsymbol{\xi};\ \mathbf{m}, \mathbf{L}\mathbf{L}^\top)\Big|_{\boldsymbol{\xi}=\mathbf{m}+\mathbf{L}\boldsymbol{\epsilon}}\right]$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}\sim N(\mathbf{0},\mathbf{I})} \left\{\nabla_{\boldsymbol{\psi}} \left[\log p(\mathbf{y}, T^{-1}(\boldsymbol{\xi})) + \log \mid J_{T^{-1}(\boldsymbol{\xi})} \mid - \log N(\boldsymbol{\xi};\ \mathbf{m}, \mathbf{L}\mathbf{L}^\top)\Big|_{\boldsymbol{\xi}=\mathbf{m}+\mathbf{L}\boldsymbol{\epsilon}}\right]\right.$$

$$\approx \frac{1}{M}\sum_{\ell=1}^{M} \nabla_{\boldsymbol{\psi}} \left[\log p(\mathbf{y}, T^{-1}(\boldsymbol{\xi})) + \log \mid J_{T^{-1}(\boldsymbol{\xi})} \mid - \log N(\boldsymbol{\xi};\ \mathbf{m}, \mathbf{L}\mathbf{L}^\top)\Big|_{\boldsymbol{\xi}=\mathbf{m}+\mathbf{L}\boldsymbol{\epsilon}^{(\ell)}}\right],$$

onde $\boldsymbol{\epsilon}^{(\ell)} \overset{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{I}),\ \ \ell = 1, \ldots, M$.

- Update via gradient method $(0 < \delta < 1)$:

$$\boldsymbol{\psi} \leftarrow \boldsymbol{\psi} + \delta \times \nabla_{\boldsymbol{\psi}} ELBO(\mathbf{y}, \boldsymbol{\psi}).$$

ADVI - Semi-parametric regression with splines

- Semi-parametric Bayesian model:

$$
\begin{aligned}
(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \phi) &\sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma},\ \phi^{-1}\mathbf{I}_n), \\
(\beta_j \mid \lambda, \phi, \alpha) &\overset{\text{iid}}{\sim} GG(0, \lambda^{-\frac{1}{\alpha}}\phi^{-\frac{1}{2}}, \alpha), \\
\lambda &\sim Gamma(a_\lambda, b_\lambda) \\
\phi &\sim Gamma(a_\phi, b_\phi) \\
\boldsymbol{\gamma} &\sim N(0, s_\gamma \mathbf{I}) \\
\alpha = 2.5 \times \eta, &\quad \eta \sim Beta(a_\alpha, b_\alpha)
\end{aligned}
$$

- original parameters:

$$\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \phi, \lambda, \alpha) \in \mathbb{R}^{p_x} \times \mathbb{R}^{p_z} \times \mathbb{R}^+ \times \mathbb{R}^+ \times (0;\ 2,5).$$

- Transformed parameters:

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

- Variational distribution $q_{\boldsymbol{\psi}}(\boldsymbol{\xi}) = N(\boldsymbol{\xi}; \mathbf{m}, \mathbf{L}\mathbf{L}^\top)$.
- Variational parameters: $\boldsymbol{\psi} = (\mathbf{m}, \mathbf{L})$.

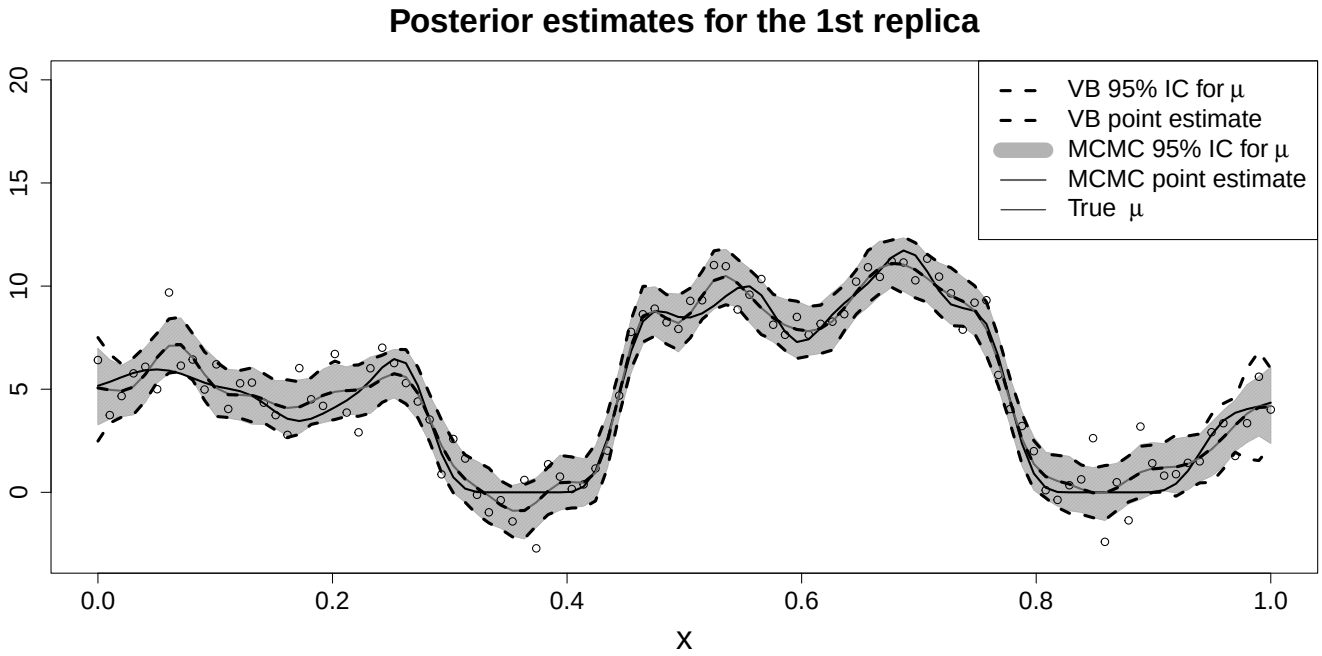
Simulation

Simulation Scheme

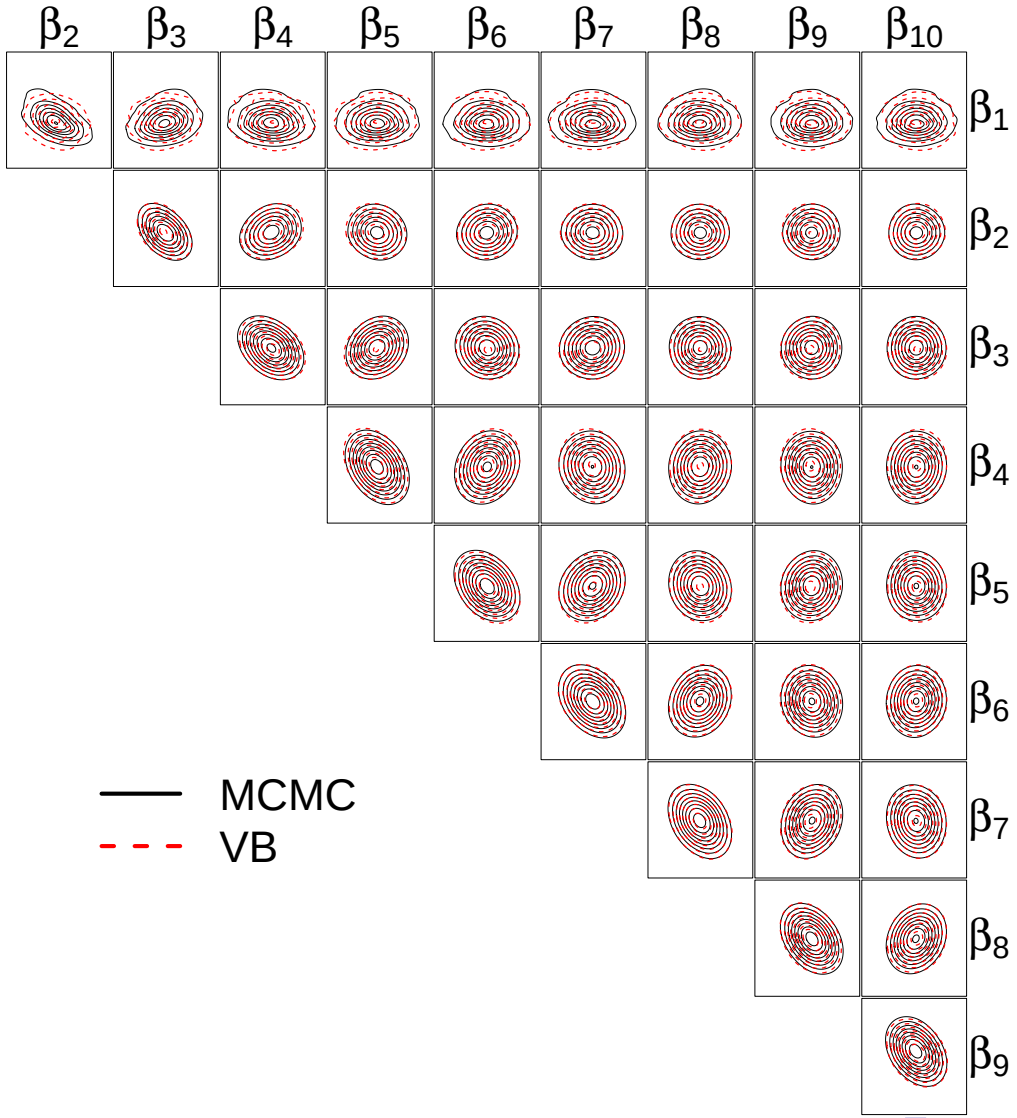
- 100 points.
- 100 simulated datasets.
- B-spline with 30 knots.

Prior Specification

- $p(\lambda) = Gamma(1, 1)$
- $p(\phi) = Gamma(1, 1)$
- $\alpha = 2.5\eta,\ \eta \sim Beta(1, 1)$
- $(\beta_j \mid \lambda, \phi, \alpha) \overset{\text{iid}}{\sim} GG(0, \lambda^{-\frac{1}{\alpha}} \phi^{-\frac{1}{2}}, \alpha)$



Simulation: bivariate posteriors for the B-splines coefficients



Computational times

VB Specifications

- ADAM Optimizer (Kingma and Ba, 2014)
- Learning rate: $\delta = 0.01$
- 100 Monte Carlo samples to estimate the gradients
- 1000 to 10000 iterations

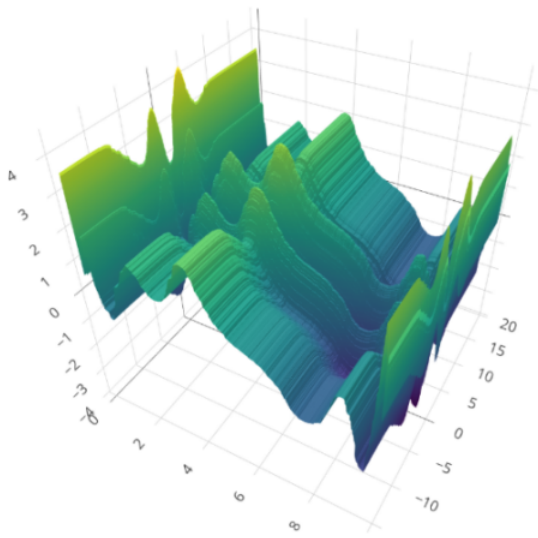
MCMC Specifications (Mallick and Yi, 2018)

- 5000 iterations

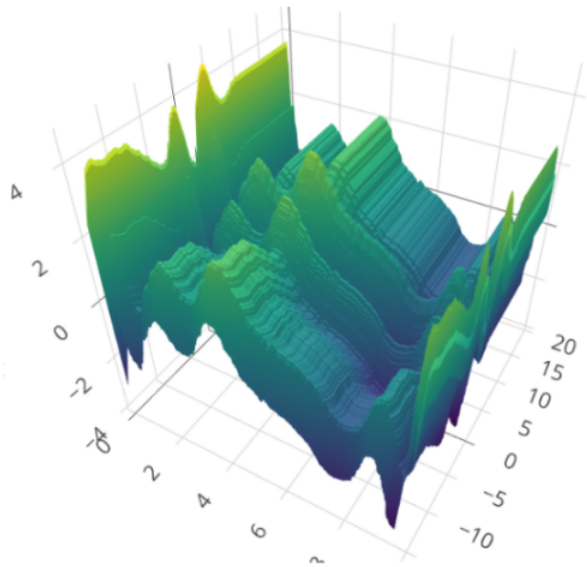
Table: Computational times (seconds) for MCMC and ADVI according to simulated data size (n). *: In this case, we ran 50000 iterations of MCMC instead of 5000 because of problems with the convergence of the MCMC chains.

n	MCMC	ADVI	ADVI (5000 iterations)
$1,000$	$219s^*$	$6s$	$15s$
$10,000$	$64s$	$6s$	$15s$
$50,000$	$294s$	$15s$	$15s$
$100,000$	$618s$	$13s$	$65s$
$500,000$	$2,744s$	$76s$	$76s$
$1,000,000$	$4,570s$	$143s$	$72s$

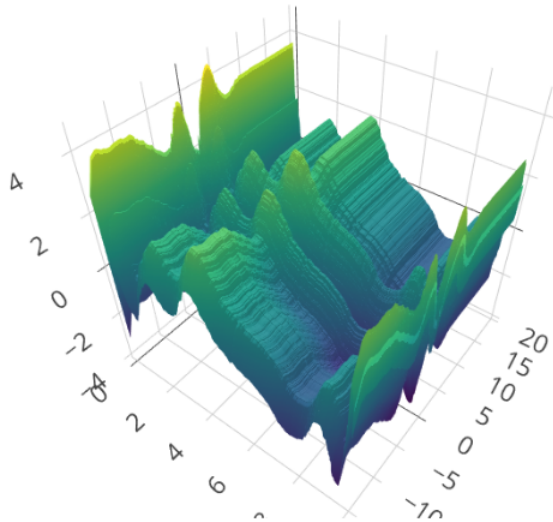
Multiple Covariates



Simulated curve



VB



MCMC

Figure: Simulated mean of y as a function of covariates x_1 and x_2. Posterior point estimate for the average of y under the proposed VI and MCMC approach.

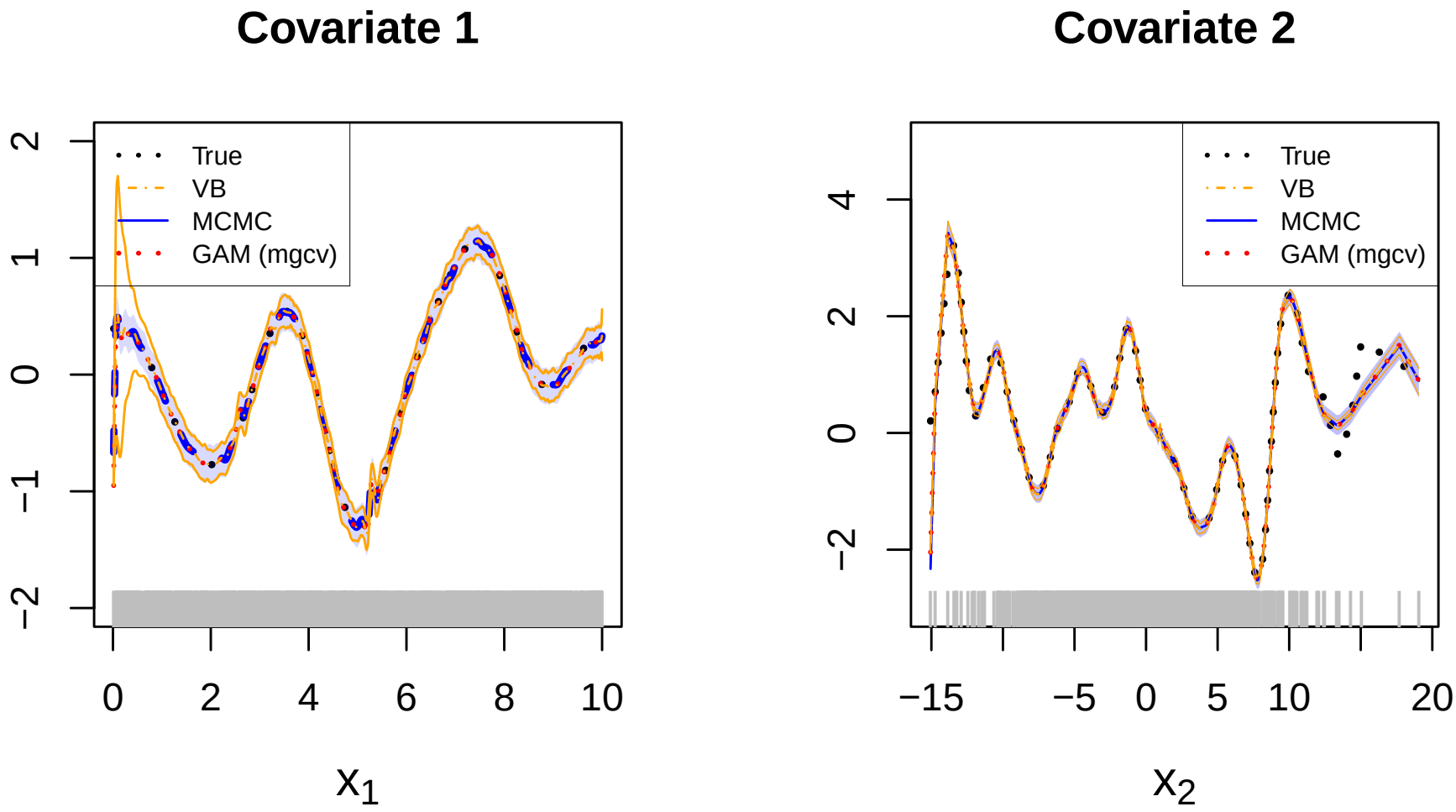


Figure: Comparison of MCMC and GAM when fitting the simulated data.

Application to real data

Dataset

- Energy load in Northeast Brazil
- 2014 to 2022
- Hourly scale measurements
- 69,717 observations
- Data available at https://dados.ons.org.br/dataset/carga-energia

Model

- Semi-parametric Bayesian regression with bridge penalty
- 1 knot every 100 hours (700 knots in total)
- Representation in Fourier bases (cosine and sine) for weekly periodicity
- 868 columns in total.

MCMC : 19hs 15min
VI: 14min

Full energy charge

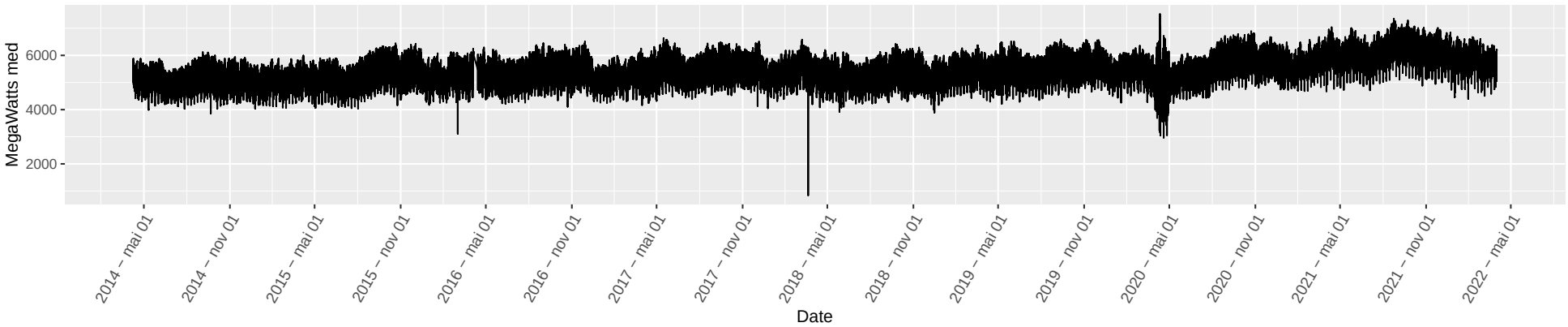


Figure: Full Energy Charge data (measured hourly).

Average response (first 100 times)

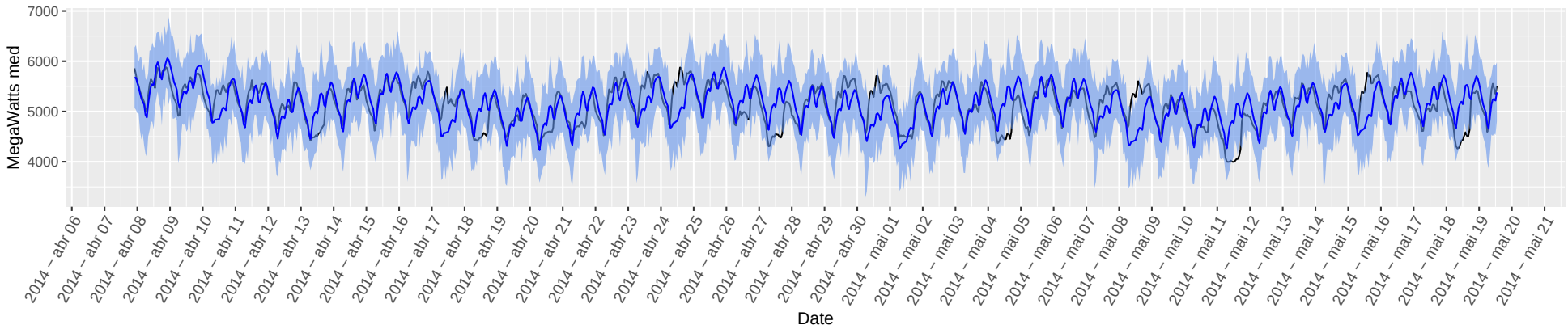


Figure: VB 95% credibility bands and posterior mean for the average response. Only the first 1000 observations of Energy Charge data are shown.

Average response (last 100 times)

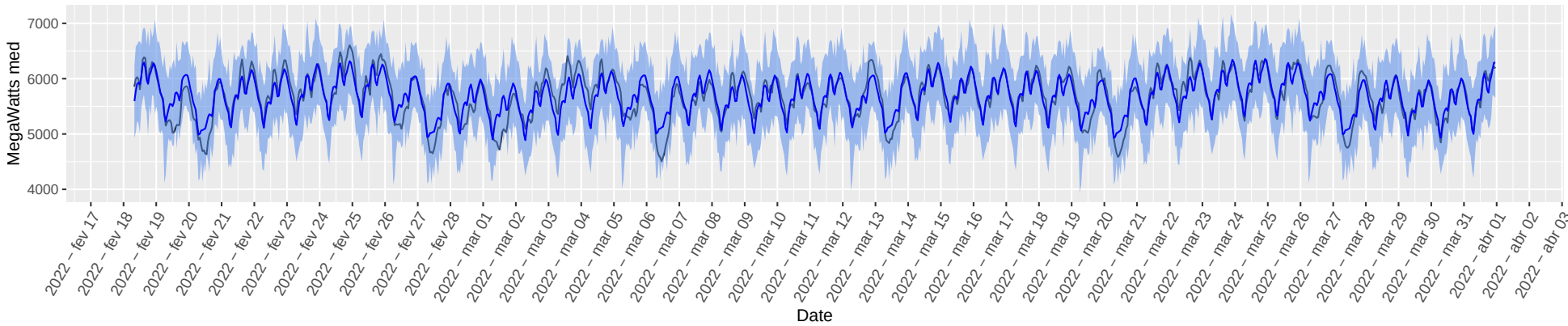


Figure: VB 95% credibility bands and posterior mean for the average response. Only the last 1000 observations of Energy Charge data are shown.

Conclusions and future work

- ADVI proposal for Bayesian inference in semi-parametric bridge regression models.
- Full Bayesian inference framed as an optimization problem.
- Stochastic gradients drastically reduce computational time compared to MCMC.
- Variational inference via ADVI estimates the posterior joint uncertainty well.
- It is not necessary to assume independence (mean-field!) in the variational family.
- Future work:
  1. Expand the distribution of the response variable beyond the Gaussian.
  2. Include other ways of penalizing base coefficients (e.g., p-splines).

Muito obrigado!
Thank you!

carloszanini@dme.ufrj.br

References

- Eilers, P., & Marx, B. (2021). Why P-splines?
- Hastie, T., Tibshirani, R., Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., & Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of machine learning research.*
- Mallick, H., & Yi, N. (2018). Bayesian bridge regression. *Journal of applied statistics*, 45(6), 988-1008.
- Polson, N. G., Scott, J. G., & Windle, J. (2014). The Bayesian bridge. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 713-733.
- Zanini, C. T. P., Migon, H. S., & Dias, R. (2024). Variational Inference for Bayesian Bridge Regression.