

# Synergizing Roughness Penalization and Basis Selection in Bayesian Spline Regression

Sunwoo Lim (joint work with Seonghyun Jeong)

Nov 1, 2023

# Outline

1 Problem Description

2 Proposed method

3 Theoretical Analysis

4 Simulation study

## Bayesian Nonparametric regression by Splines

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n, \quad \sigma^2 > 0.$$

$f : [0, 1] \rightarrow \mathbb{R}$ ,  $x_i \in [0, 1]$  are fixed.

- B-Spline approximation:  $f(\cdot) = \sum_{j=1}^J \theta_j B_j(\cdot)$ ,  $B_j$ : B-spline basis function.
- Matrix notation:  $\mathbf{y} = \mathbf{B}_J \boldsymbol{\theta}_J + \boldsymbol{\epsilon}$ ,  $\mathbf{B}_J \in \mathbb{R}^{n \times J}$  has  $B_j(x_i)$  as  $(i, j)$ th element.
- Careful design of the complexity is crucial.
- Two popular smoothing ideas: **Bayesian P-splines & Basis selection.**

## Bayesian P-splines (Lang & Brezger, 2004)

Bayesian P-splines place plenty of equidistant knots and add a roughness penalty.

$$\pi(\boldsymbol{\theta}_J \mid \lambda) \propto \frac{1}{\lambda^{\text{rank}(\mathbf{D}_J^T \mathbf{D}_J)/2}} \exp\left(-\frac{1}{2\lambda} \boldsymbol{\theta}_J^T \mathbf{D}_J^T \mathbf{D}_J \boldsymbol{\theta}_J\right),$$

$$\mathbf{D}_J = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix} \in \mathbb{R}^{(J-2) \times J}. \quad (1)$$

$\mathbf{D}_J$ : matrix representation of second order finite difference  $\Delta^2$ :

$$\Delta^2(\theta_j) = \Delta(\Delta(\theta_j)) = (\theta_j - \theta_{j-1}) - (\theta_{j-1} - \theta_{j-2}) = \theta_j - 2\theta_{j-1} + \theta_{j-2}.$$

$\boldsymbol{\theta}_J^T \mathbf{P}_J \boldsymbol{\theta}_J \approx \|f''\|_2^2$  : finite difference approximation of the roughness penalty.

## Bayesian P-splines

Often represent  $f(\cdot) = \tilde{\theta}_1 + \sum_{j=2}^J \tilde{\theta}_j \tilde{B}_j$ ,  $\tilde{B}_j$ : mean-centered

$$\pi(\tilde{\theta}_1) \propto 1,$$

$$\begin{aligned} \pi(\tilde{\boldsymbol{\theta}}_J | \lambda) &\propto \frac{1}{\lambda^{\text{rank}(\tilde{\mathbf{D}}_J^T \tilde{\mathbf{D}}_J)/2}} \exp\left(-\frac{1}{2\lambda} \tilde{\boldsymbol{\theta}}_J^T \tilde{\mathbf{D}}_J^T \tilde{\mathbf{D}}_J \tilde{\boldsymbol{\theta}}_J\right), \quad \tilde{\boldsymbol{\theta}}_J = (\theta_2, \dots, \theta_J)^T \in \mathbb{R}^{J-1}, \\ \tilde{\mathbf{D}}_J &= \begin{pmatrix} -2 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix} \in \mathbb{R}^{(J-2) \times (J-1)}. \end{aligned} \tag{2}$$

Note)  $\tilde{\mathbf{D}}_J$  is constructed by removing the first column of  $\mathbf{D}_J$ .

**Pros:** Efficient computation, scalable to additive models.

**Cons:** Prior variance of  $f$  not invariant to  $J$ , overfitting, not rate-adaptive.

## Bayesian Basis Selection

Find optimal  $J$  by Bayesian Model Selection (BMS), without roughness penalty.

$$\begin{aligned} \pi(\tilde{\theta}_1) &\propto 1, \\ \tilde{\theta}_J \mid J, \lambda, \sigma^2 &\sim N_{J-1}(0, \lambda\sigma^2 n(\tilde{\mathbf{B}}_J^T \tilde{\mathbf{B}}_J)^{-1}). \end{aligned} \tag{3}$$

- $\tilde{\mathbf{B}}_J \in \mathbb{R}^{n \times (J-1)}$ : basis matrix having  $\tilde{B}_{j+1}(x_i)$  for the  $(i, j)$ th element.
- $\lambda$  scales  $f$  around the mean. The posterior mean of  $\theta_J$  is a scalar multiple of  $\hat{\theta}_J$ .

### Pros

- ① Close relationship with familiar information criteria (e.g, AIC, BIC).
- ② Rate-adaptive estimation.

### Cons

- ① Bias in approximating  $f$  solely by the knot specification.
- ② Free-knot splines not advantageous unless  $f$  has varying smoothness.

# Motivating Example

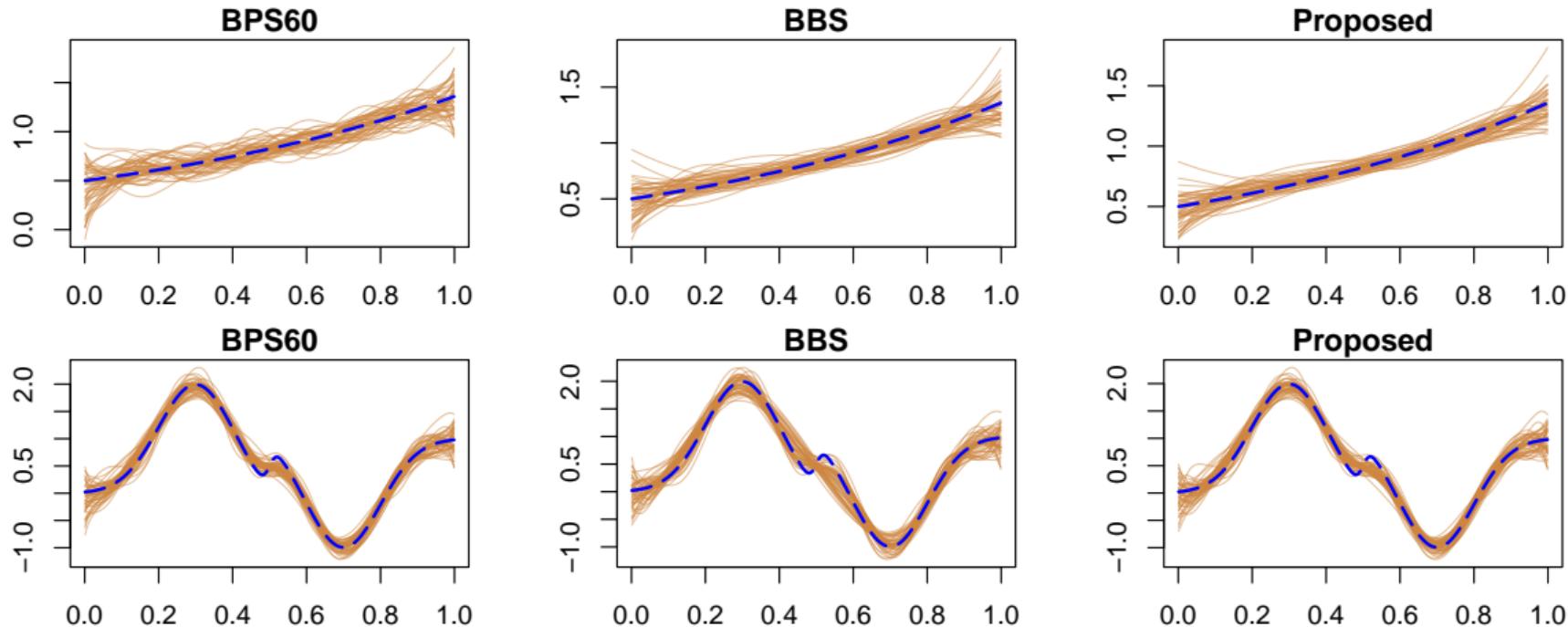


Figure: Pointwise means (orange), true  $f$  (blue), where  $n = 200, \sigma = 0.5$ .

# Outline

1 Problem Description

2 Proposed method

3 Theoretical Analysis

4 Simulation study

## Model specification

We propose the following fused prior distribution on  $(\tilde{\theta}_1, \tilde{\boldsymbol{\theta}}_J)$ ,

$$\begin{aligned} \tilde{\theta}_1 | \sigma^2 &\sim N(0, \sigma^2 \kappa^2), \\ \tilde{\boldsymbol{\theta}}_J | J, \sigma^2, \lambda, \tau &\sim N_{J-1} \left( 0, \lambda \sigma^2 \left( (1 - \tau) \tilde{\mathbf{P}}_J + \tau n^{-1} \tilde{\mathbf{B}}_J^T \tilde{\mathbf{B}}_J \right)^{-1} \right). \end{aligned} \quad (4)$$

- $\kappa > 0$ : large fixed constant, mimicking  $\tilde{\theta}_1 \propto 1$ .
- Weight parameter  $\tau \in (0, 1)$ : balance between **penalization** and **BBS**.

## Prior for other parameters

$\pi(J) \propto \nu^J, \quad J = 4, \dots, n.$  Truncated geometric prior is used.

$\pi(\sigma^2) \sim IG(a_\sigma, b_\sigma), \quad a_\sigma, b_\sigma > 0.$  Prior used for Bayesian P-splines and BBS.

$\pi(\lambda) \sim Exp(c_\lambda), \quad c_\lambda > 0.$

$\pi(\tau) \sim U(\delta, 1 - \delta), \quad \delta > 0, \quad \delta \approx 0.$

## Blocked Gibbs Sampling Algorithm

Sample from  $\pi(J, \sigma^2, \tilde{\theta}_1, \tilde{\theta}_J, \lambda, \tau | \mathbf{y})$ . Alternate between full conditionals of  $(J, \sigma^2, \tilde{\theta}_1, \tilde{\theta}_J)$ ,  $\lambda$ , and  $\tau$ .

- ➊ Draw  $J$  from  $\pi(J|\lambda, \tau, \mathbf{y}) \propto \pi(J)p(\mathbf{y}|J, \lambda, \tau)$ . The marginal likelihood is

$$p(\mathbf{y}|J, \lambda, \tau) \propto \left| \mathbf{I}_{J-1} - \boldsymbol{\Omega}_{J,\lambda,\tau}^{-1} \tilde{\mathbf{B}}_J^T \tilde{\mathbf{B}}_J \right|^{1/2} \times \left( b_\sigma + \frac{1}{2} \mathbf{y}^T \left[ \mathbf{I}_n - \frac{1}{n+\kappa^2} \mathbf{1}_n \mathbf{1}_n^T - \tilde{\mathbf{B}}_J \boldsymbol{\Omega}_{J,\lambda,\tau}^{-1} \tilde{\mathbf{B}}_J^T \right] \mathbf{y} \right)^{-(a_\sigma + n/2)},$$

$\boldsymbol{\Omega}_{J,\lambda,\tau} = \frac{1-\tau}{\lambda} \tilde{\mathbf{P}}_J + (n + \frac{\tau}{\lambda}) n^{-1} \tilde{\mathbf{B}}_J^T \tilde{\mathbf{B}}_J$ . Use Metropolis-Hastings, proposing  $J+1$  or  $J-1$  w.p  $\frac{1}{2}$ .

- ➋ Draw  $\sigma^2$  from  $\sigma^2 | J, \lambda, \tau, \mathbf{y} \sim \text{IG}\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2} \mathbf{y}^T \left[ \mathbf{I}_n - \frac{1}{n+\kappa^2} \mathbf{1}_n \mathbf{1}_n^T - \tilde{\mathbf{B}}_J \boldsymbol{\Omega}_{J,\lambda,\tau}^{-1} \tilde{\mathbf{B}}_J^T \right] \mathbf{y}\right)$ .

- ➌ Draw  $(\tilde{\theta}_1, \tilde{\theta}_J)$  from  $\pi(\tilde{\theta}_1, \tilde{\theta}_J | \sigma^2, J, \lambda, \tau, \mathbf{y})$ , where

$$\tilde{\theta}_1 | \sigma^2, \mathbf{y} \sim \mathcal{N}\left(\frac{\sum_{i=1}^n y_i}{n + \kappa^{-2}}, \frac{\sigma^2}{n + \kappa^{-2}}\right),$$

$$\tilde{\theta}_J | \sigma^2, J, \lambda, \tau, \mathbf{y} \sim \mathcal{N}_{J-1}\left(\boldsymbol{\Omega}_{J,\lambda,\tau}^{-1} \tilde{\mathbf{B}}_J^T \mathbf{y}, \sigma^2 \boldsymbol{\Omega}_{J,\lambda,\tau}^{-1}\right).$$

- ➍ Draw  $\lambda$  from  $\pi(\lambda | J, \tau, \sigma^2, \tilde{\theta}_1, \tilde{\theta}_J, \mathbf{y}) \propto \pi(\lambda) \pi(\tilde{\theta}_J | J, \sigma^2, \lambda, \tau)$  by slice sampling.

- ➎ Draw  $\tau$  from  $\pi(\tau | J, \lambda, \sigma^2, \tilde{\theta}_1, \tilde{\theta}_J, \mathbf{y}) \propto \pi(\tau) \pi(\tilde{\theta}_J | J, \sigma^2, \lambda, \tau)$  by time-efficient grid sampling.

# Outline

1 Problem Description

2 Proposed method

3 Theoretical Analysis

4 Simulation study

# Assumptions

- ① The true function  $f_0$  belongs to the  $\alpha$ -Hölder space defined as

$$\mathcal{H}_\lambda^\alpha([0, 1]) = \left\{ f : \mathcal{X} \rightarrow \mathbb{R}; \max_{0 \leq k \leq \lfloor \alpha \rfloor} \sup_{x \in [0, 1]} |f^{(k)}(x)| + \sup_{x, y \in [0, 1]: x \neq y} \frac{|f^{(\lfloor \alpha \rfloor)}(x) - f^{(\lfloor \alpha \rfloor)}(y)|}{|x - y|^{\alpha - \lfloor \alpha \rfloor}} \leq \infty \right\}$$

with  $\alpha \in (0, l + 1]$ .  $l$  denotes the spline polynomial degree.

- ② True variance parameter  $\sigma_0^2$  satisfies  $c^{-1} \leq \sigma_0^2 \leq c$  for some constant  $c > 1$ .
- ③ For fixed  $x_i$ ,  $i = 1, \dots, n$ , there exists a distribution function  $G$  such that

$$\|G_n - G\|_\infty = o(n^{-1/(2\alpha+1)} (\log n)^{-2\alpha/(2\alpha+1)}),$$

where  $G_n(\cdot) = n^{-1} \sum_{i=1}^n 1\{\cdot \leq x_i\}$  is empirical distribution function of  $x_i$ ,  $i = 1, \dots, n$ .

## Theoretical Results

**Def) Posterior contraction rate** of  $f$  is the fastest rate  $\epsilon_n$  of posterior contraction:

$$\Pi_n(d(f, f_0) > M_n \epsilon_n \mid D_n) \xrightarrow{P} 0, \quad \forall M_n \rightarrow \infty.$$

**Theorem)** Under the assumptions, our method achieves *adaptive* posterior contraction rate

$$n^{-\alpha/(2\alpha+1)} (\log n)^{1/(2\alpha+1)+1}$$

with respect to the empirical  $\ell_2$  norm.

**Note)** minimax rate of convergence in nonparametric regression is  $n^{-\frac{\alpha}{2\alpha+1}}$ .

# Outline

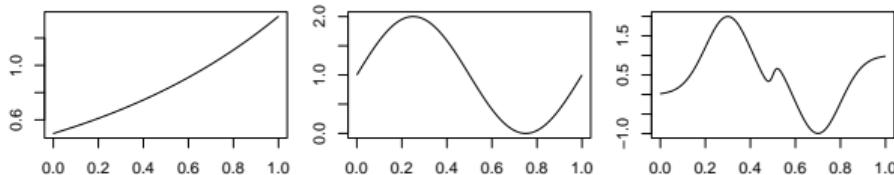
1 Problem Description

2 Proposed method

3 Theoretical Analysis

4 Simulation study

## Simulation setting



**Figure:** The test functions used.  $f_1$  (left),  $f_2$  (center),  $f_3$  (right).

## Data Setting

- $x_i \stackrel{\text{iid}}{\sim} U(0, 1)$ ,  $n \in \{200, 500, 1000\}$ ,  $\sigma^2 \in \{0.1^2, 0.5^2\}$ .
- Compute  $\text{MSE} = \frac{1}{n} \sum_{i=1}^n (f_i - f_{(\text{estimated})i})^2$  with 200 replications.

## Nonparametric regression methods

- *PS30* and *PS60*: Frequentist P-splines with 30 / 60 interior knots.
- *BPS30* and *BPS60*: Bayesian P-splines with 30 / 60 interior knots.
- *BBS*: Bayesian basis selection via the mixtures of  $g$ -prior.
- *Proposed*: The proposed method.

# Simulation results : $f_1$

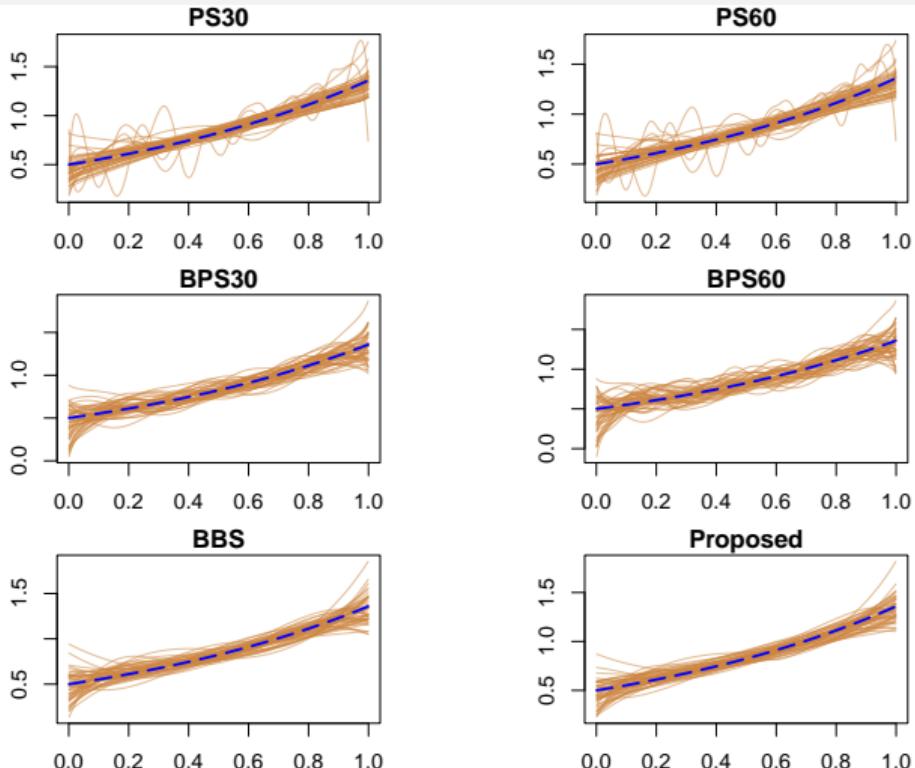


Figure: Estimates (orange),  $f$  (blue) ,  $n = 200, \sigma = 0.5$

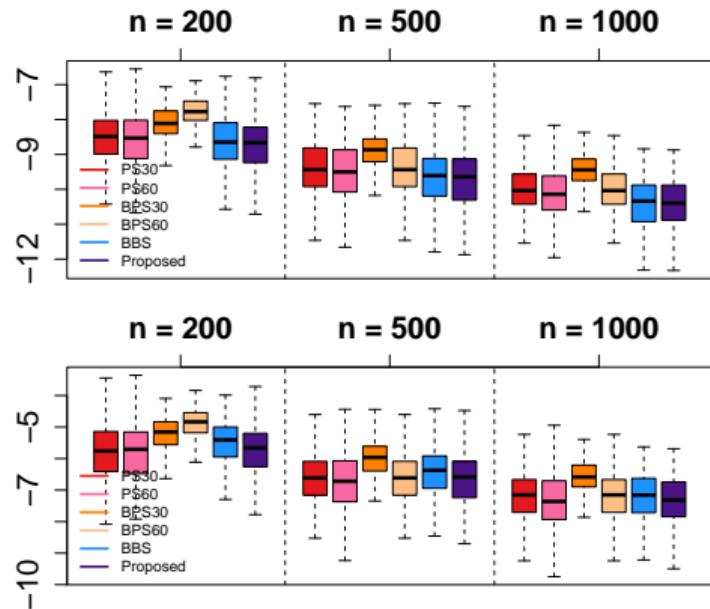
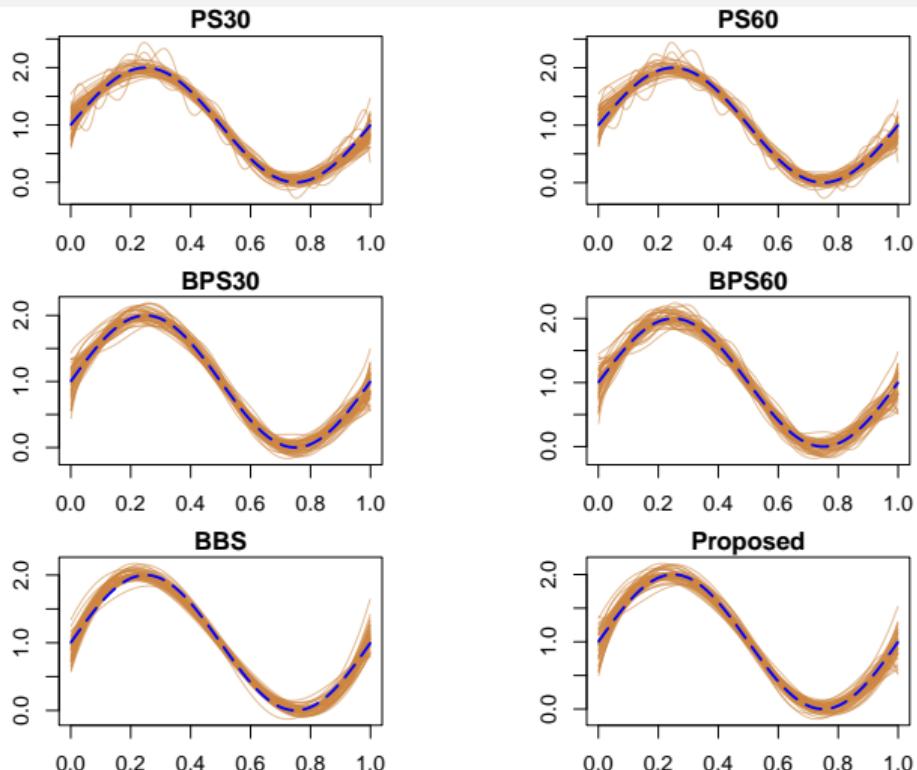
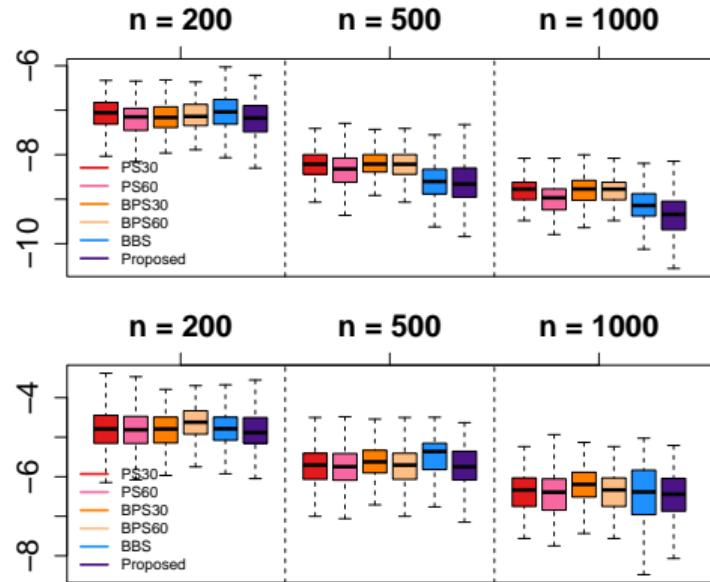
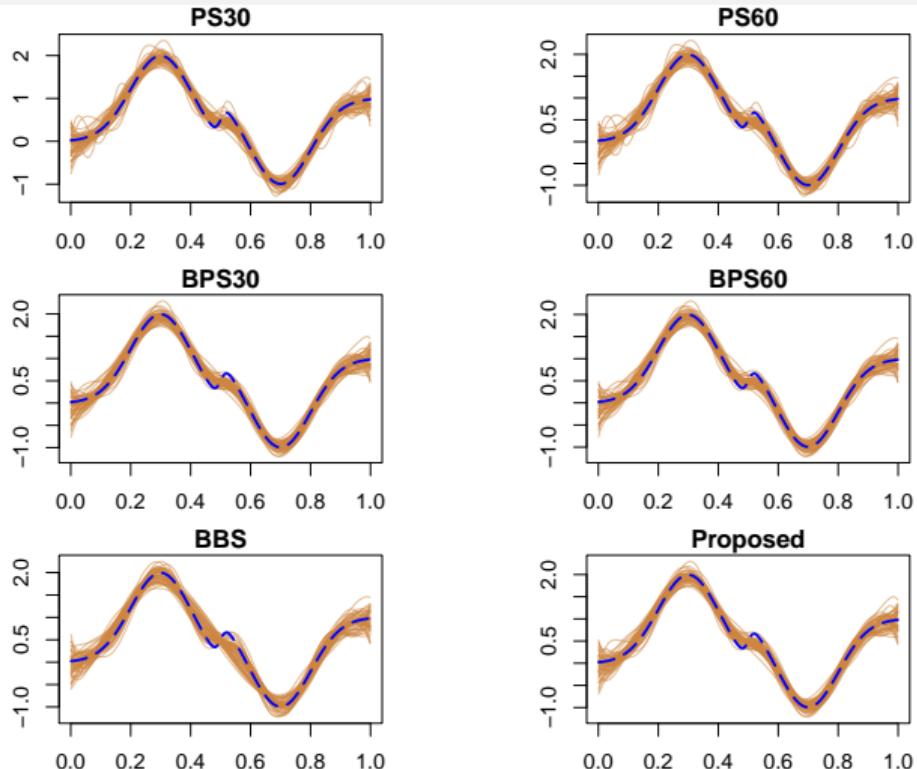


Figure: log MSEs,  $\sigma = 0.1$ (top),  $0.5$ (bottom)

Simulation results :  $f_2$ Figure: Estimates (orange),  $f$  (blue) ,  $n = 200, \sigma = 0.5$ Figure: log MSEs,  $\sigma = 0.1$ (top),  $0.5$ (bottom)

Simulation results :  $f_3$ Figure: Estimates (orange),  $f$  (blue) ,  $n = 200, \sigma = 0.5$ Figure: log MSEs,  $\sigma = 0.1$ (top),  $0.5$ (bottom)

# Discussion

- Bayesian and frequentist P-splines suffer from overfitting.  
Log MSE decreases slower than knot-selection based procedures as  $n$  grows.
- BBS suffers from **model misspecification**.  
Inaccurate coverage probabilities is a piece of evidence for this fact.
- Alleviates weaknesses of each method by convex combination of precision matrices.  
Key idea) Prior precision : amount of penalty.
- Desirable empirical and theoretical properties.
- Drawback: erratic behavior around boundaries by using B-splines.

## References

- Bach, P., & Klein, N. (2021). Posterior Concentration Rates for Bayesian Penalized Splines.
- Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties.
- George, E., & Foster, D. P. (2000). Calibration and empirical Bayes variable selection.
- Kang, G., & Jeong, S. (2023). Model selection-based estimation for generalized additive models using mixtures of g-priors: Towards systematization.
- Lang, S., & Brezger, A. (2004). Bayesian P-splines.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection.
- Shen, W., Ghosal, S. (2015). Adaptive Bayesian procedures using random series priors.
- Ventrucci, M., Rue, H. (2016). Penalized complexity priors for degrees of freedom in Bayesian P-splines.
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions.