Synergizing Roughness Penalization and Basis Selection in Bayesian Spline Regression

Sunwoo Lim (with Seonghyun Jeong)

Oct 12, 2023

Outline

Bayesian Nonparametric regression by Splines

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

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

- B-Spline approximation: $f(\cdot) = \sum_{j=1}^{J} \theta_j B_j(\cdot)$, B_j : B-spline basis function.
- Matrix notation: $\boldsymbol{y} = \mathbf{B}_J \boldsymbol{\theta}_J + \boldsymbol{\epsilon}$, $\mathbf{B}_J \in \mathbb{R}^{n \times J}$ 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

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

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

$$\pi(\tilde{\boldsymbol{\theta}}_J \mid \lambda) \propto \frac{1}{\lambda^{\mathsf{rank}(\tilde{\mathbf{D}}_J^T \tilde{\mathbf{D}}_J)/2}} \exp\left(-\frac{1}{2\lambda} \tilde{\boldsymbol{\theta}}_J^T \tilde{\mathbf{D}}_J^T \tilde{\mathbf{D}}_J \tilde{\boldsymbol{\theta}}_J\right), \quad \tilde{\boldsymbol{\theta}}_J = (\theta_2, \ldots, \theta_J)^T \in \mathbb{R}^{J-1},$$

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

Penalizing $\tilde{\boldsymbol{\theta}}_J^T \tilde{\mathbf{P}}_J \tilde{\boldsymbol{\theta}}_J \approx$ penalizing $\int (f''(x))^2 dx$, for $\tilde{\mathbf{P}}_J = \tilde{\mathbf{D}}_J^T \tilde{\mathbf{D}}_J$.

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.

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

- $\tilde{\mathbf{B}}_J \in \mathbb{R}^{n\times(J-1)}$: basis matrix having $\tilde{B}_{j+1}(x_i)$ for the (i,j)th element.
- λ scales f around the mean.

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

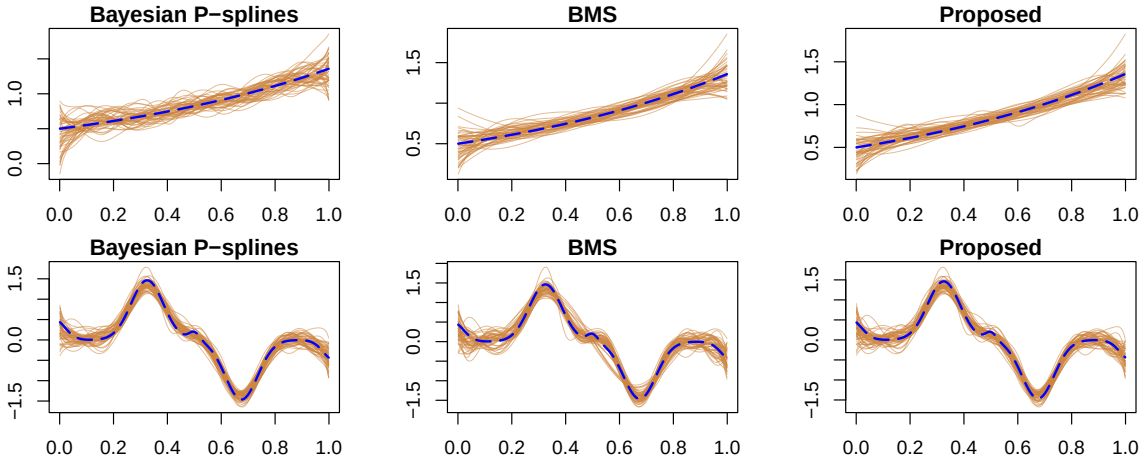

Figure: Pointwise means (orange), true f (blue), where $n = 200, \sigma = 0.5$.

Outline

Model specification

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

$$\begin{aligned} \tilde{\theta}_1 \mid \sigma^2 &\sim \mathsf{N}(0, \sigma^2\kappa^2), \\ \tilde{\boldsymbol{\theta}}_J \mid J, \sigma^2, \lambda, \tau &\sim \mathsf{N}_{J-1}\left(0, \lambda\sigma^2\left((1-\tau)\tilde{\mathbf{P}}_J + \tau n^{-1}\tilde{\mathbf{B}}_J^T\tilde{\mathbf{B}}_J\right)^{-1}\right). \end{aligned} \tag{3}$$

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

Prior for other parameters

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

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

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

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

Blocked Gibbs Sampling Algorithm

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

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

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

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

2. Draw σ^2 from $\sigma^2|J, \lambda, \tau, \mathbf{y} \sim \mathsf{IG}\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}\mathbf{y}^T\left[\mathbf{I}_n - \frac{1}{n+\kappa^2}\mathbf{1}_n\mathbf{1}_n^T - \tilde{\mathbf{B}}_J\boldsymbol{\Omega}_{J,\lambda,\tau}^{-1}\tilde{\mathbf{B}}_J^T\right]\mathbf{y}\right)$.
3. Draw $(\tilde{\theta}_1, \tilde{\boldsymbol{\theta}}_J)$ from $\pi(\tilde{\theta}_1, \tilde{\boldsymbol{\theta}}_J|\sigma^2, J, \lambda, \tau, \mathbf{y})$, where

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

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

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

Outline

Assumptions

1. The true function f_0 belongs to the α-Hölder space defined as

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

   with $\alpha \le q$ and $\lambda \lesssim \sqrt{\log n}$.

2. True variance parameter σ_0^2 satisfies $c^{-1} \le \sigma_0^2 \le c$ for some constant $c > 1$.

3. For fixed x_i, $i = 1, \ldots, n$, there exists a distribution function G such that

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

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

Theoretical Results

Def) Posterior contraction rate of f is the fastest rate ϵ_n of posterior contraction:

$$\Pi_n(d(f, f_0) > \epsilon_n \mid D_n) \xrightarrow{p} 0.$$

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

$$n^{-\frac{\alpha}{2\alpha+1}}(\log n)^c \quad \text{for some } c > 0$$

with respect to the empirical ℓ_2 norm.

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

Outline

Simulation setting

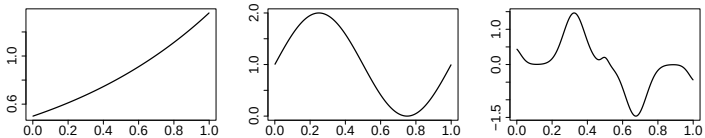

Figure: Test functions. f_1 (left), f_2 (center), f_3 (right)

Data Setting

- $x_i \overset{\text{iid}}{\sim} \text{U}(0,1), \quad n = 200, 500, 1000, 2000, \quad \sigma = 0.1, 0.5.$
- Compute MSE $= \frac{1}{n}\sum_{i=1}^{n}(f_i - f_{(\text{estimated})i})^2$ with 200 replications.

Nonparametric regression methods

- *BPS50* and *BPS30*: Bayesian P-splines with 50/30 interior knots.
- *PS50* and *PS30*: Frequentist P-splines, penalty parameter optimized via GCV.
- *BMS*: BMS via the mixtures of g-prior.
- *Proposed*: The proposed method.

Simulation results : f_1

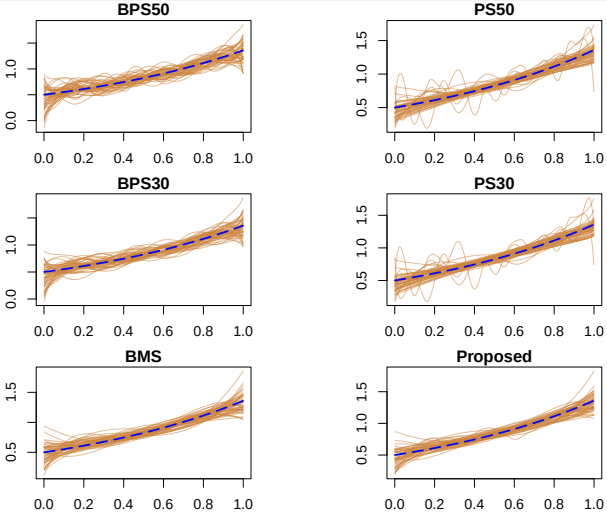

Figure: Estimates (orange), f (blue) , $n = 200, \sigma = 0.5$

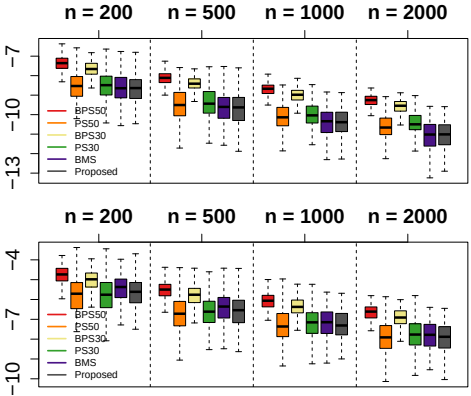

Figure: log MSEs, $\sigma = 0.1$(top), 0.5(bottom)

Simulation results : f_2

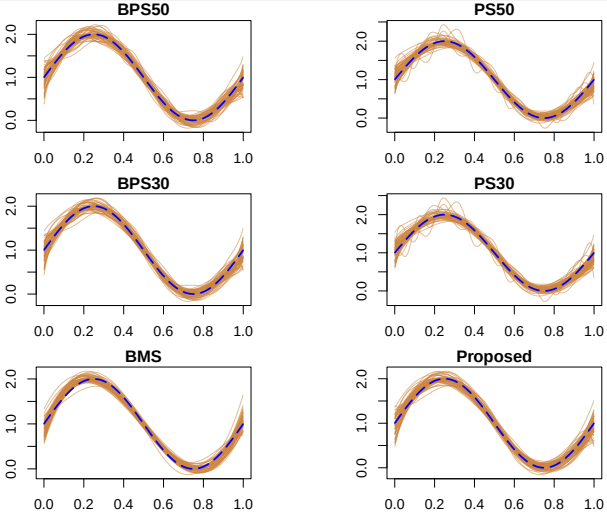

Figure: Estimates (orange), f (blue) , $n = 200, \sigma = 0.5$

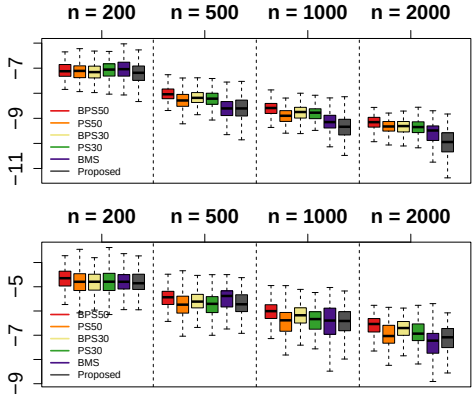

Figure: log MSEs, $\sigma = 0.1$(top), 0.5(bottom)

Simulation results : f_3

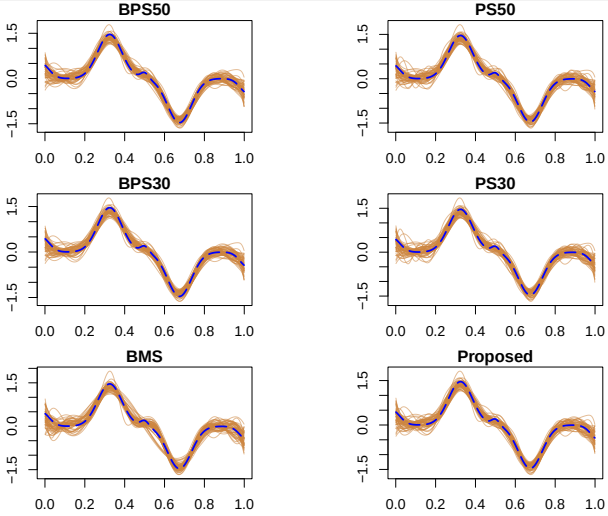

Figure: Estimates (orange), f (blue) , $n = 200, \sigma = 0.5$

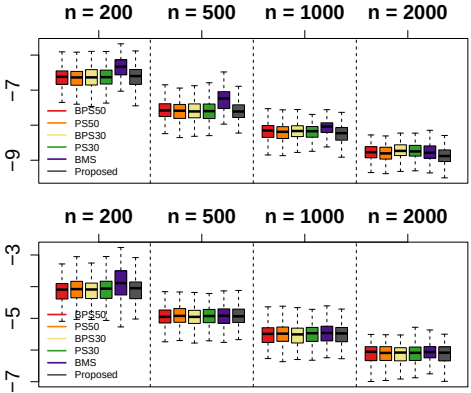

Figure: log MSEs, $\sigma = 0.1$(top), 0.5(bottom)

Discussion

- Bayesian and frequentist P-splines suffer from overfitting.
  Log MSE decreases slower than knot-selection based procedures as n grows.
- BMS suffers from **model misspecification**.
- Alleviates weaknesses of each method by convex combination of precision matrices.
- 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.