Estimation for generalized additive models using Bayesian model selection with mixtures of g-priors

Gyeonghun Kang [1] Seonghyun Jeong [1,2]

[1] Department of Statistics and Data Science, Yonsei University

[2] Department of Applied Statistics, Yonsei University

Nov 30, 2022

Table of Contents

Main contributions

- Reviews and extends estimation methods for generalized additive models via Bayesian model selection using g-priors.
- Proposes a simple slice sampler algorithm to draw from a class of generalized beta distributions.
- Discusses a new perspective on the behavior of mixtures of g-priors, focusing on the Bayes factors as a penalty function with respect to model complexity and goodness-of-fit.

I. Basic concepts

Generalized Additive Models (GAM)

- Given $x_i \in (x_{i1}, \dots, x_{ip})^T \in \mathbb{R}^p$, we assume $y_i \in \mathbb{R}$ has the density

$$p(y_i; \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right), \quad i = 1, \dots, n, \tag{1}$$

  where $E(y_i) = b'(\theta_i)$, $V(y_i) = \phi b''(\theta_i)$ for twice differentiable $b(\cdot)$.

- GAM [Hastie and Tibshirani, 1986] is an extension of a linear model where,

$$h\big(E(y_i \mid X_i)\big) = \eta_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}), \quad \left(\sum_{i=1}^{n} f_j(x_{ij}) = 0\right) \tag{2}$$

  $h(\cdot)$ is a smooth, monotonic **link** function with domain $(-\infty, \infty)$ (e.g. logit link $\log \frac{p}{1-p} \in \mathbb{R}$).

- (1) can be re-expressed in terms of the additive predictors $\theta_i = (h \circ b')^{-1}(\eta_i)$.

Generalized Additive Models (GAM)

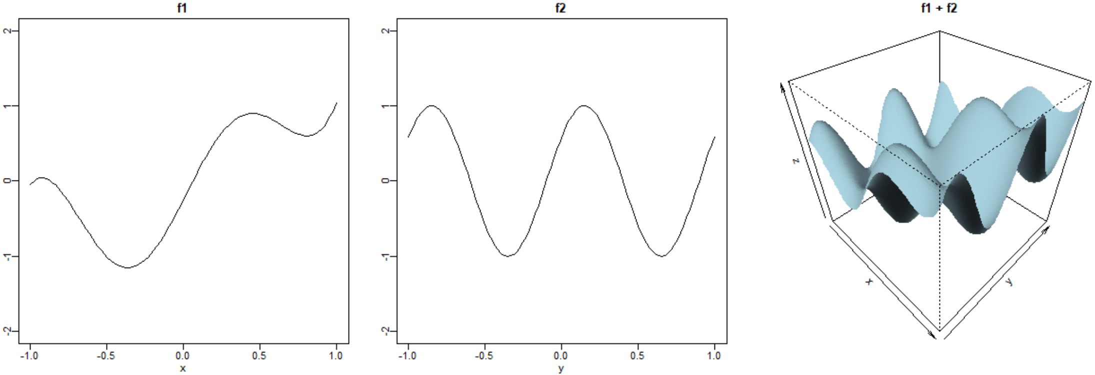

Figure: Nonlinear, but no interactions.

Basis Expansion and Splines

- We parameterize f_j by spline basis representation; for basis functions $b_{j1}(\cdot), \ldots, b_{jK_j}(\cdot)$,

$$f_j(\cdot) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(\cdot)$$

  where $b_{jk}(\cdot)$ are centered to meet the identifiability condition.

- Many choices of basis function exist, including piecewise polynomials on the interval spanned by knots $x_{(1)} = \xi^L < \xi_1, \ldots, \xi_{L_j} < x_{(n)} = \xi^R$, dth continuously differentiable at each knot ξ_k constitute dth order spline.
- Basis expansion renders design matrix $B_\xi \in \mathbb{R}^{n \times J}$ and spline coefficients $\beta_\xi \in \mathbb{R}^J$, with which (2) is represented as $\eta_\xi = \alpha 1_n + B_\xi \beta_\xi$. ($J = \sum_{j=1}^p K_j$)

Basis Expansion and Splines

- We deploy a modification of **natural cubic spline basis functions** in Hastie et al. [2009] for straightforward application of Bayesian variable selection methods.
- Specifically, for boundary knots $\{t^L, t^U\}$ and a set of M interior knots $t = \{t_1, \ldots, t_M\}$ satisfying $-\infty < t^L < t_1 < \cdots < t_M < t^U < \infty$, we define $N_k : \mathbb{R} \to \mathbb{R}$, $k = 1, \ldots, M+1$, as

$$
\begin{aligned}
N_1(u) &= u, \\
N_k(u) &= N(u; t^L, t^U, t_k) \\
&:= \frac{(u - t_k)_+^3 - (u - t^U)_+^3}{t^U - t_k} - \frac{(u - t^L)_+^3 - (u - t^U)_+^3}{t^U - t^L}, \quad k = 1, \ldots, M.
\end{aligned}
\tag{3}
$$

  Together with the constant term $N_0(u) = 1$, the basis functions in (3) generate piecewise cubic functions with the restriction that the spline function is linear beyond the boundary knots $\{t^L, t^U\}$.
- **With (3), adding a new interior knot-point $t_* \in (t^L, t^U)$ is equivalent to adding the corresponding basis term $N(u; t^L, t^U, t_*)$, likewise for elimination.**

Basis Expansion and Splines

- How to choose the number and location of knots $\xi = \{\xi_{jk} \mid j = 1, \ldots, p, k = 1, \ldots, L_j\}$?

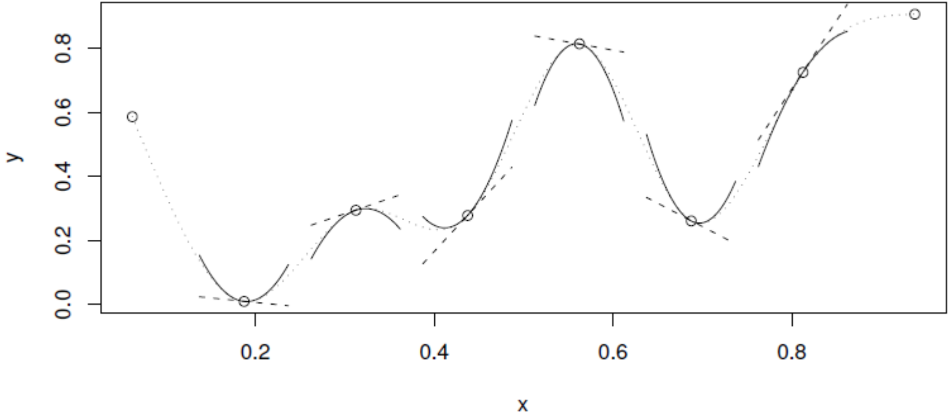

Figure: Wood [2017], p196

Knot placement through Bayesian Model Selection

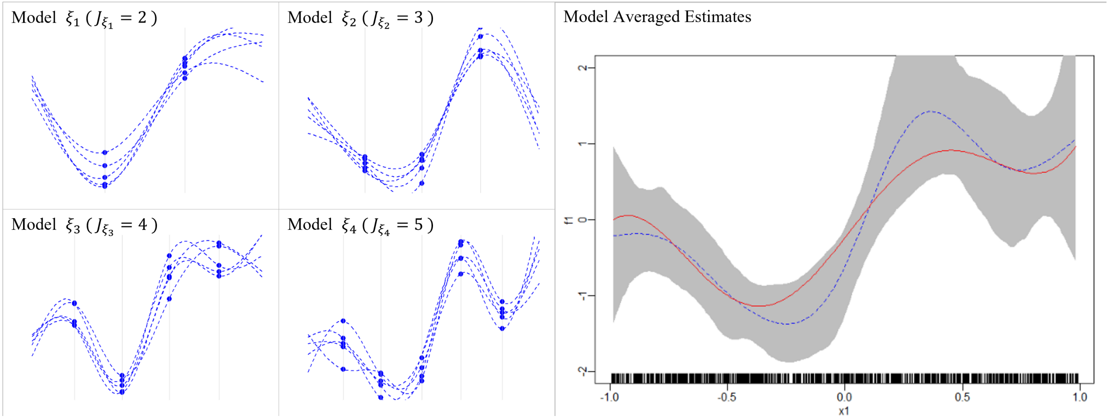

Figure: Bernoulli, $n = 1000$, *(Left)* Different knot configurations ξ, *(Right)* Model-avreaged posterior pointwise mean (blue dashed) versus the true function (red). 95% posterior pointwise credible interval shaded in gray

Knot placement through Bayesian Model Selection

- We let data choose $\xi \in \Xi$ via **Bayesian model selection** [Kass and Raftery, 1995].

$$\pi(\xi, \beta_\xi, y) = \underbrace{f(y \mid \beta_\xi, \xi)}_{\text{data distribution (1)}} \times \underbrace{\pi(\beta_\xi \mid \xi)}_{\text{splines under } \xi} \times \underbrace{\pi(\xi)}_{\text{number and locations}}$$

- The marginal posterior of ξ is

$$\pi(\xi \mid y) = \frac{\overbrace{\pi(\xi)}^{\text{prior}} \quad \overbrace{f(y \mid \xi)}^{\text{model evidence}}}{\int f(y \mid \xi)\, d\Pi(\xi)}$$

1. $\pi(\xi)$ reflects the assumptions on Ξ and the prior therewithin.
2. $f(y \mid \xi)$ is the **Model Evidence**, the marginal likelihood of ξ

$$f(y \mid \xi) = \int f(y \mid \beta_\xi, \xi)\pi(\beta_\xi \mid \xi)\, d\beta_\xi$$

Knot placement through Bayesian Model Selection

- $\pi(\xi \mid y)$ is the weight of ξ in obtaining **model-averaged estimate** of some functional of interest $\mathcal{L} : (f_1, \ldots, f_p) \mapsto \mathcal{L}(f_1, \ldots, f_p)$ (e.g., pointwise evaluation, credible interval, etc.)

$$\pi\big(\mathcal{L}(f_1, \ldots, f_p) \mid Y\big) = \int_{\Xi} \pi\big(\mathcal{L}(f_1, \ldots, f_p) \mid \xi, Y\big)\, d\Pi(\xi \mid Y).$$

- In the rest of the slides we discuss:

$$\pi(\xi \mid y) \propto \pi(\xi) f(y \mid \xi)$$

1. **Priors for knots** 1) even, 2) variable-selection, 3) free knot splines for $\pi(\xi)$
2. **Mixtures of g-priors** for $\pi(\beta_\xi \mid \xi)$
   - Our choice of $\pi(\beta_\xi \mid \xi)$ for tractable approximation to $f(y \mid \xi)$
   - The effect of different priors on g on knot selection behavior

II. Prior specification

Locally Orthogonal g-prior [Li and Clyde, 2018]

We consider locally orthogonal g-prior introduced in Li and Clyde [2018],

$$\begin{aligned} \beta_\xi \mid g, \xi &\sim N\big(0, g(\tilde{B}_\xi^T J_n(\hat{\eta}_\xi)\tilde{B}_\xi)^{-1}\big), \\ J_n(\hat{\eta}_\xi) &= \mathrm{diag}(-Y_i\theta''(\hat{\eta}_{\xi,i}) + (b\circ\theta)''(\hat{\eta}_{\xi,i}), i = 1, \dots, n) \\ \tilde{B}_\xi &= \big[I_n - \mathrm{tr}(J_n(\hat{\eta}_\xi))^{-1} 1_n 1_n^T J_n(\hat{\eta}_\xi)\big] B_\xi \end{aligned} \tag{4}$$

- $\hat{\eta}_\xi = (\hat{\eta}_{\xi,1}, \cdots, \hat{\eta}_{\xi,n})^T = \hat{\alpha}_\xi 1_n + B_\xi \hat{\beta}_\xi$ where $\hat{\alpha}_\xi$, $\hat{\beta}_\xi$ are the MLE under ξ
- $J_n(\hat{\eta}_\xi)$ is the observed information matrix of η_ξ evaluated at $\hat{\eta}_\xi$
- $\tilde{B}_\xi$ consists of columns of B_ξ centered by the weighted average per the diagonal element of $J_n(\hat{\eta}_\xi)$
- With $\pi(\alpha) \propto 1$, (4) leads to tractable model evidence through integrated Laplace approximation [Wang and George, 2007]

$$p(Y \mid g, \xi) = p(Y \mid \hat{\eta}_\xi)\mathrm{tr}(J_n(\hat{\eta}_\xi))^{-1/2}(g+1)^{-J_\xi/2} \exp\left(-\frac{Q_\xi}{2(g+1)}\right), \tag{5}$$

How to choose g?

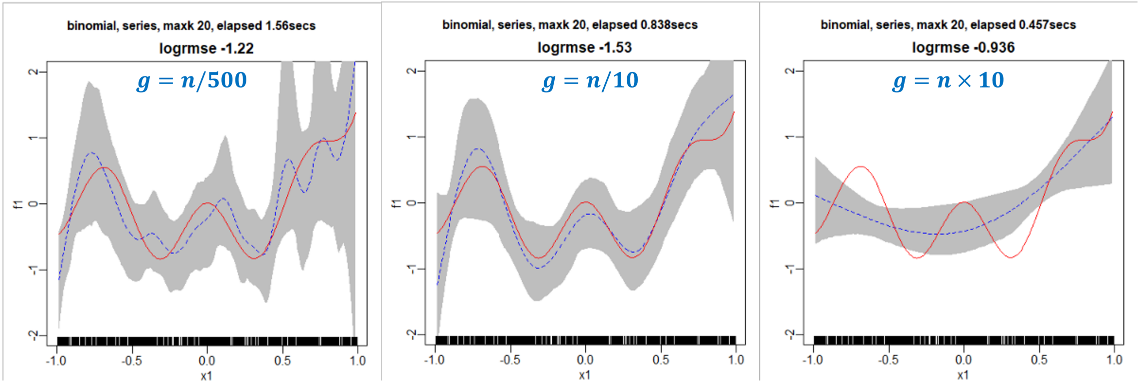

Figure: Bernoulli, $n = 1000$

Mixtures of g priors

- **Ideally we want g to be data-dependent**
  1. Constant: $g = n$ [Kass and Wasserman, 1995] or use optimal g (Empirical Bayes)
  2. Hyperprior: $g \sim \pi(g) \rightarrow$ **Mixtures of g-priors**
- We consider $(g+1)^{-1} \sim tCCH(a/2, b/2, r, s/2, \nu, \kappa)$, $a, b, \kappa > 0$ for $r, s \in \mathbb{R}$, $\nu \geq 1$, the truncated Compound Hypergeometric distribution, [Gordy, 1998]

$$
\begin{aligned}
u &\sim tCCH(a, b, z, s, \nu, \kappa) \\
f(u) &= \frac{\nu(\nu u)^{a-1}(1-\nu u)^{b-1}[\kappa + (1-\kappa)\nu u]^{-r} e^{-su}}{e^{-s/\nu}\Phi_1(b, r, a+b, s/\nu, 1-\kappa)B(a,b)} 1_{\{0<u<1/\nu\}}
\end{aligned}
\tag{6}
$$

where $B(\cdot,\cdot)$ is the beta function and $\Phi_1(\alpha, \beta, \gamma, x, y) = B(\alpha, \gamma-\alpha)^{-1} \int_0^1 u^{a-1}(1-u)^{\gamma-\alpha-1}(1-yu)^{-\beta} \exp(xu)\, du$ is the confluent hypergeometric function of two variables [Humbert, 1922].

Mixtures of g priors

- Mixtures of g-priors proposed in various literatures belong to tCCH distribution [Li and Clyde, 2018];

	a	b	r	s	ν	κ	Concentration
Uniform	2	2	0	0	1	1	$g = O(1)$
Hyper-$g^{(1)}$	1	2	0	0	1	1	$g = O(1)$
Hyper-$g/n^{(1)}$	1	2	1.5	0	1	n^{-1}	$g = O(n)$
Beta-prime$^{(2)}$	0.5	$n - J_\xi - 1.5$	0	0	1	1	$g = O(n)$
ZS-adapted$^{(3)}$	1	2	0	$n+3$	1	1	$g = O(n)$
Robust$^{(4)}$	1	2	1.5	0	$\frac{n+1}{J_\xi+1}$	1	$g = O(n)$
Intrinsic$^{(5)}$	1	1	1	0	$\frac{n+J_\xi+1}{J_\xi+1}$	$\frac{n+J_\xi+1}{n}$	$g = O(n)$

Table: Distributions belonging to the tCCH family. (1) Liang et al. [2008], (2) Maruyama and George [2011], (3) Held et al. [2015], (4) Bayarri et al. [2012], (5) Womack et al. [2014]

Mixtures of g priors

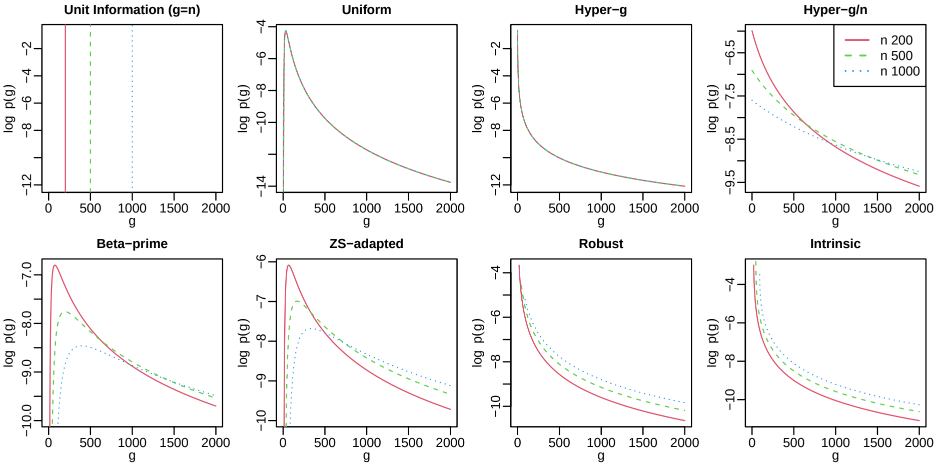

Figure: Distributions belonging to the tCCH family for $n = 200, 500, 1000$, with $J_\xi = 10$ if required.

Mixtures of g priors

The resulting model evidence $p(Y \mid \xi) = \int p(Y \mid \xi, g) d\Pi(g)$ is then expressed as

$$
\begin{aligned}
p(Y \mid \xi) = p(Y \mid \hat{\eta}_\xi) \mathrm{tr}(J_n(\hat{\eta}_\xi))^{-1/2} \nu^{-J_\xi/2} \exp\left(-\frac{Q_\xi}{2\nu}\right) \frac{B((a+J_\xi)/2, b/2)}{B(a/2, b/2)} \\
\times \Phi_1\left(\frac{b}{2}, r, \frac{a+b+J_\xi}{2}, \frac{s+Q_\xi}{2\nu}, 1-\kappa\right) \Big/ \Phi_1\left(\frac{b}{2}, r, \frac{a+b}{2}, \frac{s}{2\nu}, 1-\kappa\right),
\end{aligned}
\tag{7}
$$

We use the Gaussian-Kronrod quadrature routine available in the Boost `C++` library for Φ_1. The approximate posteriors conditional on ξ are given by[1]

$$
\begin{aligned}
\frac{1}{g+1} \mid Y, \xi &\sim \mathrm{tCCH}\left(\frac{a+J_\xi}{2}, \frac{b}{2}, r, \frac{s+Q_\xi}{2}, \nu, \kappa\right), \\
\alpha \mid Y, g, \xi &\sim \mathrm{N}\left(\hat{\alpha}_\xi, \mathrm{tr}(J_n(\hat{\eta}_\xi))^{-1}\right), \\
\beta_\xi \mid Y, g, \xi &\sim \mathrm{N}\left(\frac{g}{g+1}\hat{\beta}_\xi, \frac{g}{g+1}(\tilde{B}_\xi^T J(\hat{\eta}_\xi)\tilde{B}_\xi)^{-1}\right).
\end{aligned}
$$

[1] tCCH is sampled using the general slice sampler as in Edwards and Sokal [1988], Damlen et al. [1999]

Priors for knots

We specify priors on the space of knots $\xi = \{\xi_{jk} \mid j = 1, \ldots, p, k = 1, \ldots, L_j\}$ on the knot space Ξ

- We only consider knots satisfying $\mathrm{rank}(B_\xi) = J_\xi$ and $J_\xi < n$.
- Even so, the model space Ξ is intrinsically infinite dimensional as ξ_j can be any set of singletons on the interval spanned by x_j.
- Further restrictions on Ξ make for faster computation at a cost of reduced estimation quality.
  1. Ξ_{EK}: knots are predetermined to be evenly spaced. **(Even-knot)**
  2. Ξ_{VS}: knots are selected from a grid of equidistant points. **(VS-knot)** [Denison et al., 1998]
  3. Ξ_{FK}: knots are placed freely **(Free-knot)** [DiMatteo et al., 2001]

  The restrictions differ in rules for the mapping $L_j \mapsto \xi_j$ and the following inclusion relation holds[2];

$$\Xi_{EK} \subset \Xi_{VS} \subset \Xi_{FK}$$

[2] if we choose ξ_{jk} from the unique values of x_j for Ξ_{EK}, Ξ_{VS}

Priors for knots

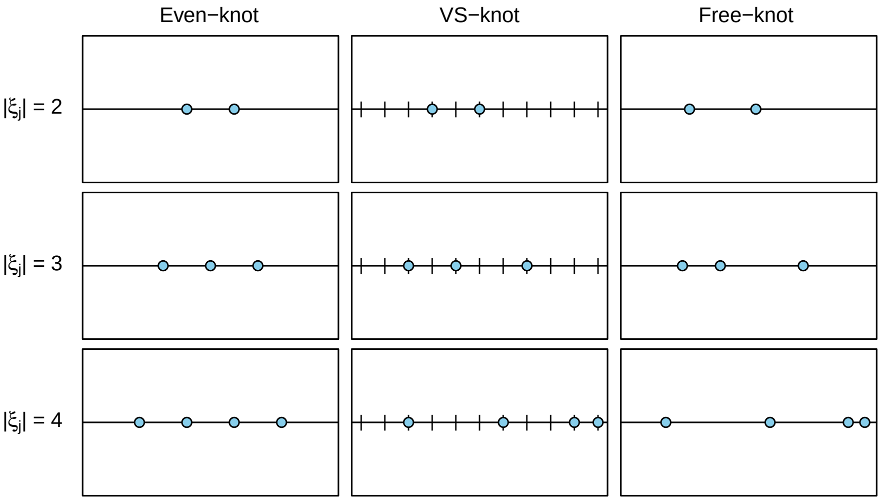

Figure: A graphical illustration of three knot placement strategies.

III. Mixtures of g-priors and penalty functions

Bayes factor as penalty functions

- The Bayes factor of two knots ξ_1, ξ_2 is defined as $BF[\xi_1;\xi_2] = p(Y \mid \xi)/p(Y \mid \xi_2)$.

Proposition

For the model (1) *and* (2) *with the prior in* (16), *consider two knots* ξ_i *and* ξ_j *where* $\hat{\eta}_{\xi_i} = \hat{\eta}_{\xi_j}$ *and* $J_{\xi_i} = J_{\xi_j} + k$ $(k \in \mathbb{N})$. *The Bayes factor of* ξ_i *to* ξ_j *is*

$$BF[\xi_i;\xi_j] = \begin{cases} (1+h)^{-k/2} & \text{if } g \sim \delta_h(g) \\ \mathrm{E}[(1+g)^{-k/2} \mid \xi_j, Y] & \text{if } g \sim tCCH(a/2, b/2, r, s/2, \nu, \kappa) \end{cases}$$

where $p(u \mid \xi_j, Y)$ *is the posterior distribution of* u *given* Y *under the model* ξ_j. *The result also holds for Gaussian additive model if either* $\kappa = 1$ *or* $s = 0$.

- $\log BF[\xi_i;\xi_j]$ indicates the amount of penalty imposed on the model ξ_i that uses additional k number of knots to no avail.

Bayes factor as penalty functions

Our main observations are that for both Gaussian and exponential family distribution,

- The penalty function $\log BF[\xi_i;\xi_j]$ of mixtures of g-priors **depends both on the model size** J_ξ **and the goodness-of-fit** R^2_ξ; the penalty gets weaker as the model size increases or the fit deteriorates. [3]
- Compared to $g = n$, mixtures of g-priors in Table 1 tend to **add more knots, especially if the current model has poor fit with many knots.**
- Within the mixtures of g-priors, $g = \mathcal{O}(1)$ priors allow comparatively more knots than $g = \mathcal{O}(n)$, markedly so in the region of poor fit.

To demonstrate these, we let $\xi_2 = \xi_1 + 1$ and plot $\log BF[\xi_1;\xi_2] > 0$, in which case $\log BF[\xi_1;\xi_2]$ denotes **the penalty on the model complexity** (greater penalty with higher $\log BF[\xi_1;\xi_2]$).

[3] We define $R^2_{\xi,pseudo} = 1 - \exp(-Q_\xi/n)$ to measure the goodness-of-fit for the exponential family models. Under some mild assumptions Q_ξ is asymptotically equivalent to the usual deviance statistics.

Bayes factor as penalty functions

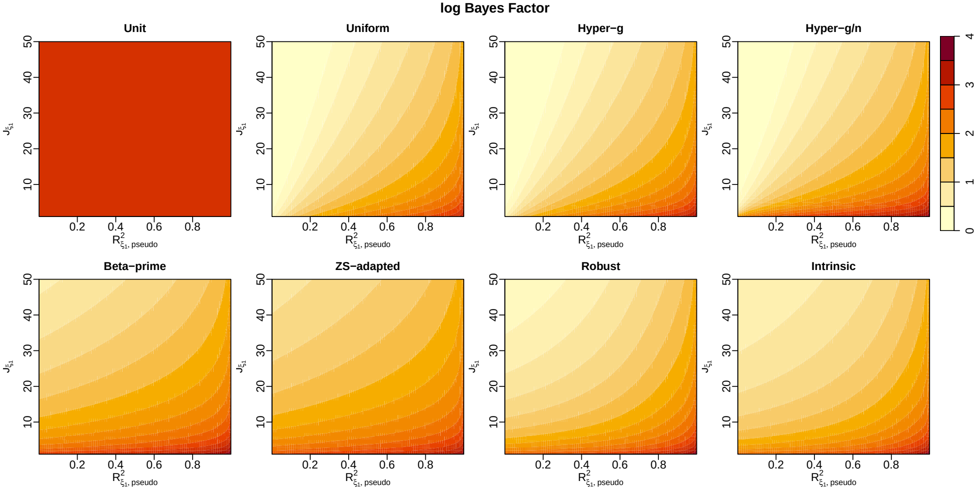

Figure: The log Bayes factor, $\log BF[\xi_1;\xi_2]$, of the exponential family model as a function of J_{ξ_1} and $R^2_{\xi_1,\text{pseudo}} (= R^2_{\xi_2,\text{pseudo}})$ for $n = 200$.

Bayes factor as penalty functions

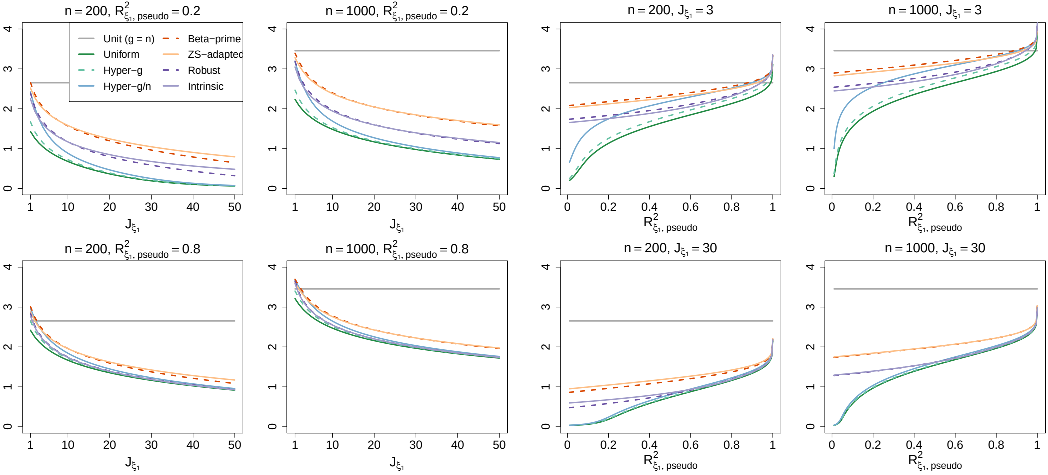

Figure: The log Bayes factor, $\log BF[\xi_1;\xi_2]$, of the exponential family model as a function of J_{ξ_1} and $R^2_{\xi_1,\text{pseudo}} (= R^2_{\xi_2,\text{pseudo}})$ for $n = 200, 1000$.

IV. Numerical Study

Comparison Among The Mixtures Of g-priors

- Throughout the simulations, we use the following uncentered functions $f_j^* : [-1, 1] \mapsto \mathbb{R}$, $j = 1, 2, 3$ as test functions:

$$\begin{aligned} f_1^*(x) &= 0.5(2x^5 + 3x^2 + \cos(3\pi x) - 1), \\ f_2^*(x) &= \frac{21(3x+1.5)^3}{8000} + \frac{21(3x-2.5)^2 e^{3x+1.5}}{400} \sin\left(\frac{(3x+1.5)^2\pi}{3.2}\right) \mathbf{1}_{(-0.5<x<0.85)}, \\ f_3^*(x) &= x. \end{aligned} \tag{8}$$

- For each j, we sample $\eta_i = f_j^*(x_i) = \alpha + f_j(x_i)$ **(univariate)** where $x_i \sim$ Unif$(-1, 1)$ so that f_j is the centered version of f_j^* and α the induced intercept. The test dataset is generated from a nonlinear logistic regression model: $Y_i \sim$ Bernoulli$\left(e^{\eta_i}/(1+e^{\eta_i})\right)$, $i = 1, \cdots, n$.
- We estimate f_j using the VS-knot spline of 30 knot candidates using unit information and mixtures of g-priors in Table 1, compare RMSE and coverage probabilities of 95% pointwise credible interval.

Comparison Among The Mixtures Of g-priors

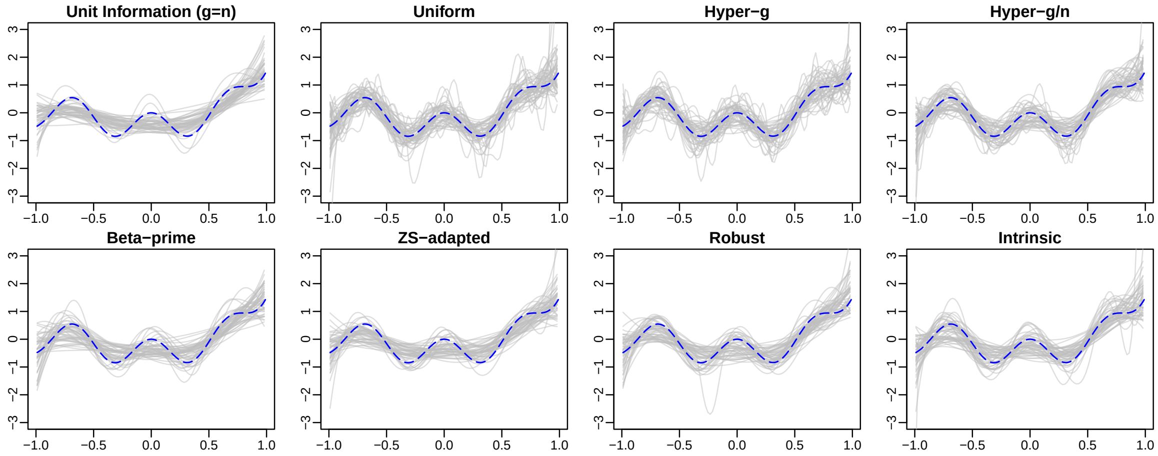

Figure: Pointwise posterior means of randomly chosen 50 replications (gray solid) and the true functions (blue dashed), $n = 500$

Comparison Among The Mixtures Of g-priors

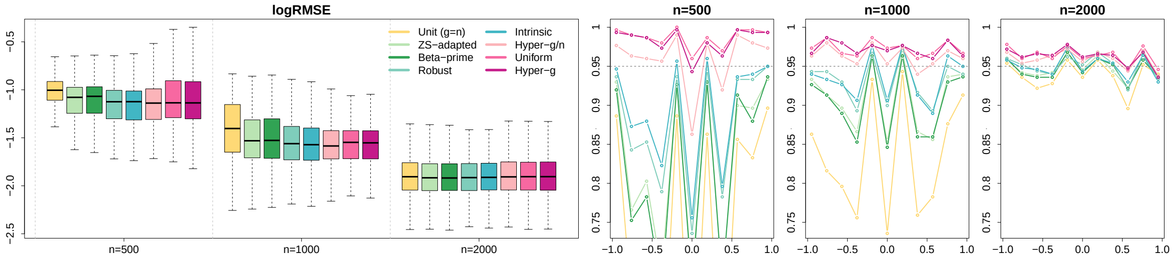

Figure: The log RMSE and the coverage probabilities for f_1 in the nonparametric logistic regression models with $n = 500, 1000, 2000$.

Comparison Among The Mixtures Of g-priors

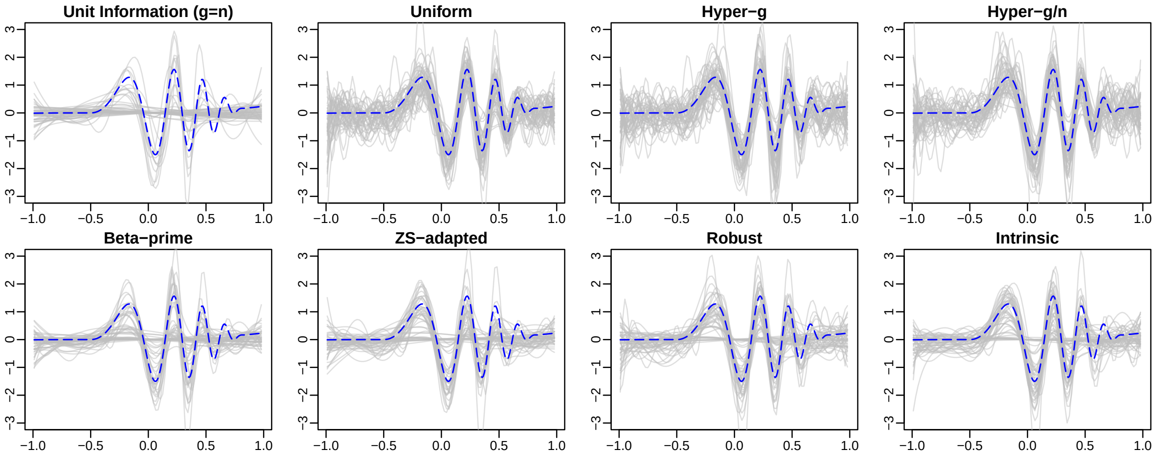

Figure: Pointwise posterior means of randomly chosen 50 replications (gray solid) and the true functions (blue dashed), $n = 500$

Comparison Among The Mixtures Of g-priors

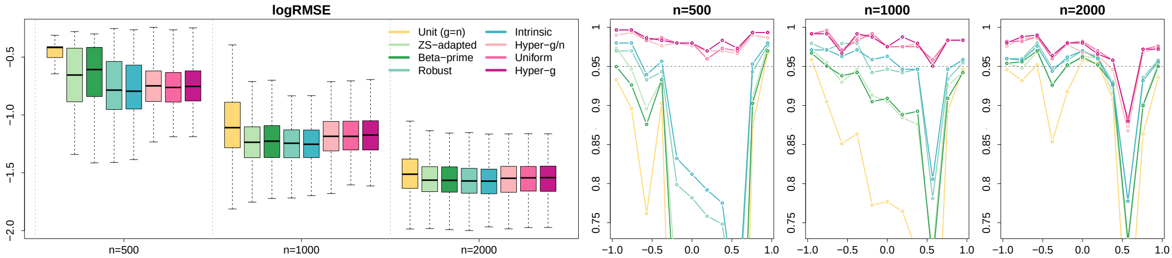

Figure: The log RMSE and the coverage probabilities for f_2 in the nonparametric logistic regression models with $n = 500, 1000, 2000$.

Comparison Among The Mixtures Of g-priors

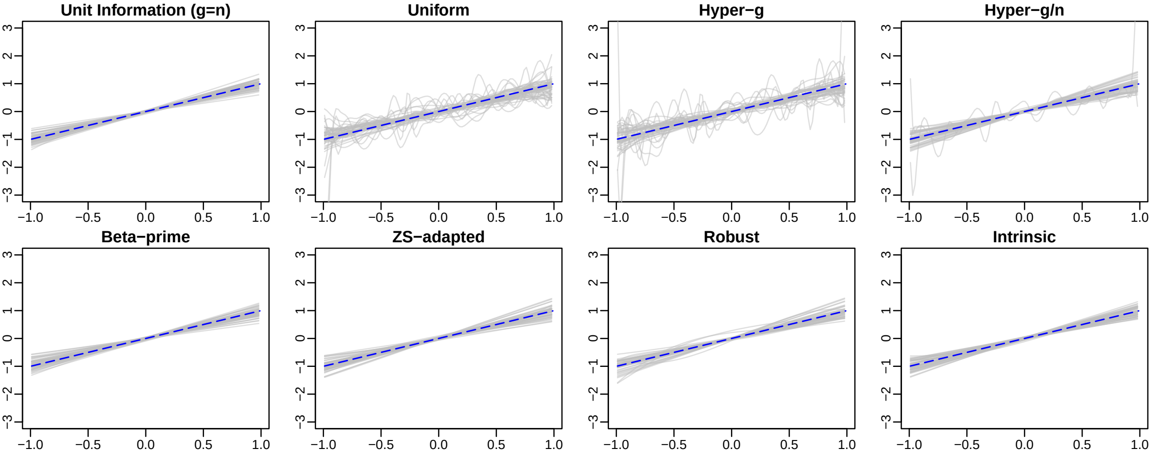

Figure: Pointwise posterior means of randomly chosen 50 replications (gray solid) and the true functions (blue dashed), $n = 500$

Comparison Among The Mixtures Of g-priors

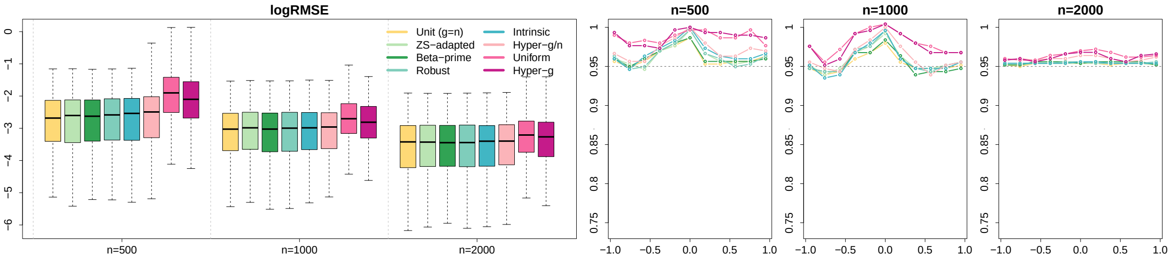

Figure: The log RMSE and the coverage probabilities for f_3 in the nonparametric logistic regression models with $n = 500, 1000, 2000$.

Comparison Among The Mixtures Of g-priors

Our conclusions are:

- Difference between priors on g become noticeable at smaller sample size.
- The unit information prior $(g = n)$ generally underperforms for nonlinear function estimation, preferring simplistic models.
- Among mixtures of g-priors, **robust** and **intrinsic** prior are ideal almost always; the beta-prime and ZS-adpated priors tend to oversmooth, whereas the uniform, hyper-g, and hyper-g/n undersmooth.
- The simulation results for Gaussian and Poisson regression also lead to similar conclusion.
- We set **robust** as our default choice as it is easier to sample; the corresponding tCCH posterior reduces to a truncated gamma for exponential family model and to a Gaussian hypergeometric distribution [Armero and Bayarri, 1994] for the Gaussian regression model.

Comparison With Other Methods

Apart from our BMS-based methods (Even-knot, VS-knot, Free-knot), another approach to spline estimation is using a **smoothing parameter** λ (P-spline);

$$\hat{f} = \arg\max_{f} l(f) - \frac{\lambda}{2} \int f''(x)^2 \, dx$$

Basis for $\hat{f}$ are determined a priori (e.g. equidistant knots).
B-spline basis allows for **Bayesian P-spline** representation with a regularizing prior;

$$\pi(\beta \mid \lambda) \propto \exp(-\beta^T S_\lambda \beta / 2)$$

Variations of **Bayesian P-spline** determine λ by

- Optimized $\hat{\lambda}$ (Empirical Bayes, GCV) [Wood, 2017], package Mgcv (**Mgcv-ps**: a non-adaptive spline, **Mgcv-ad**: an adaptive spline)
- Sampled λ (Fully Bayesian) [Brezger and Lang, 2006], package **R2BayesX**
- Numerically integrated (INLA) [Gressani and Lambert, 2021], package **Blapsr**

Comparison with Other Methods

- Throughout the simulations, we use the following uncentered functions $f_j^* : [-1, 1] \mapsto \mathbb{R}$, $j = 1, 2, 3$ as test functions:

$$
\begin{aligned}
f_1^*(x) &= 0.5(2x^5 + 3x^2 + \cos(3\pi x) - 1), \\
f_2^*(x) &= \frac{21(3x+1.5)^3}{8000} + \frac{21(3x-2.5)^2 e^{3x+1.5}}{400} \sin\left(\frac{(3x+1.5)^2\pi}{3.2}\right) \mathbf{1}_{(-0.5<x<0.85)}, \qquad (9) \\
f_3^*(x) &= x.
\end{aligned}
$$

- For each j, we sample $\eta_i = f_j^*(x_i) = \alpha + \sum_{j=1}^{3} f_j(x_i)$ **(multivariate)** where $x_i \sim$ Unif$(-1, 1)$ so that f_j is the centered version of f_j^* and α the induced intercept. The test dataset is generated from a nonlinear logistic regression model: $Y_i \sim$ Bernoulli $(e^{\eta_i}/(1 + e^{\eta_i}))$, $i = 1, \cdots, n$.
- For all methods, we configured the settings for a fair comparison: the number of evenly spaced knots for competitors was set to 30, and the same for the maximum number of knots for BMS-based methods. We compared RMSE and coverage probabilities of 95% pointwise credible interval of each method.

Comparison with Other Methods

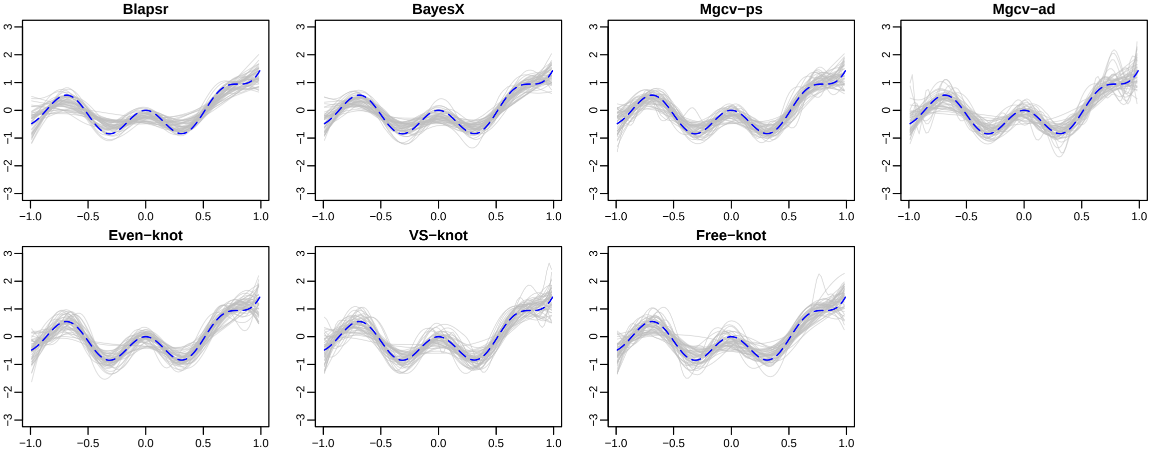

Figure: Pointwise posterior means of randomly chosen 50 replications (gray solid) and the true functions (blue dashed), $n = 1000$

Comparison with Other Methods

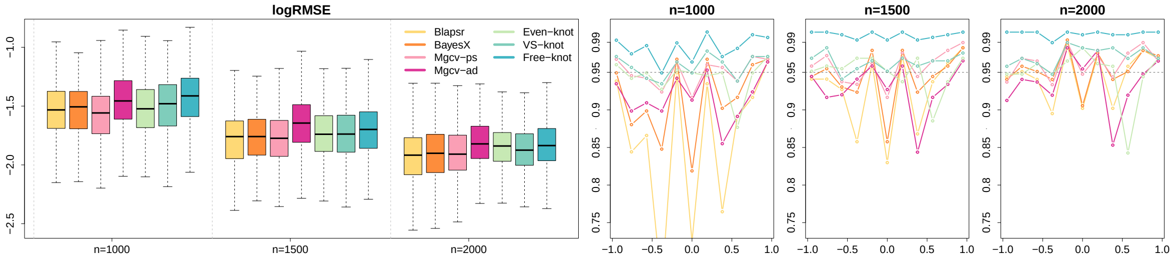

Figure: The log RMSE and the coverage probabilities for f_1 in the nonparametric logistic regression models with $n = 1000, 1500, 2000$.

Comparison with Other Methods

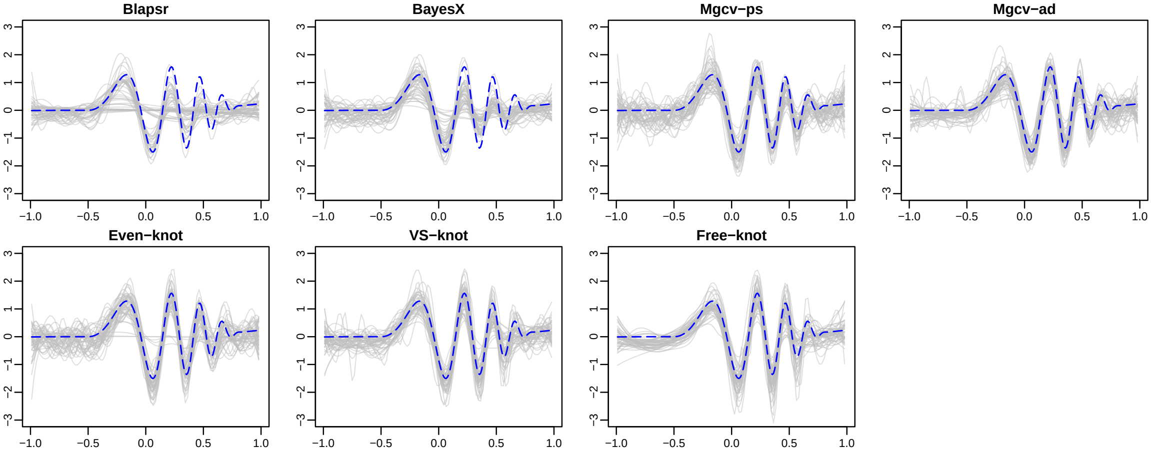

Figure: Pointwise posterior means of randomly chosen 50 replications (gray solid) and the true functions (blue dashed), $n = 1000$

Comparison with Other Methods

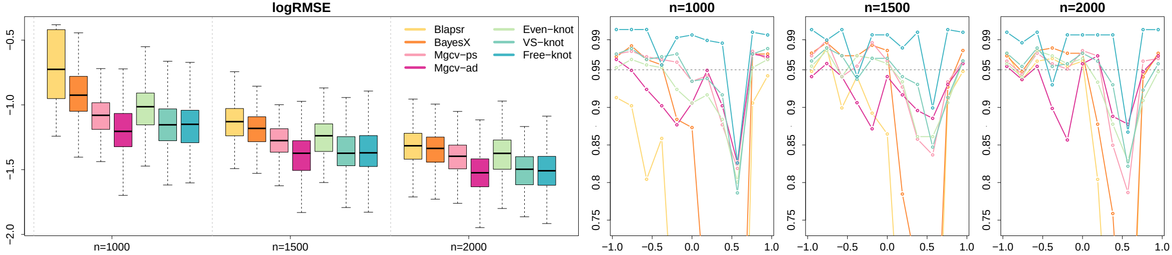

Figure: The log RMSE and the coverage probabilities for f_2 in the nonparametric logistic regression models with $n = 1000, 1500, 2000$.

Comparison with Other Methods

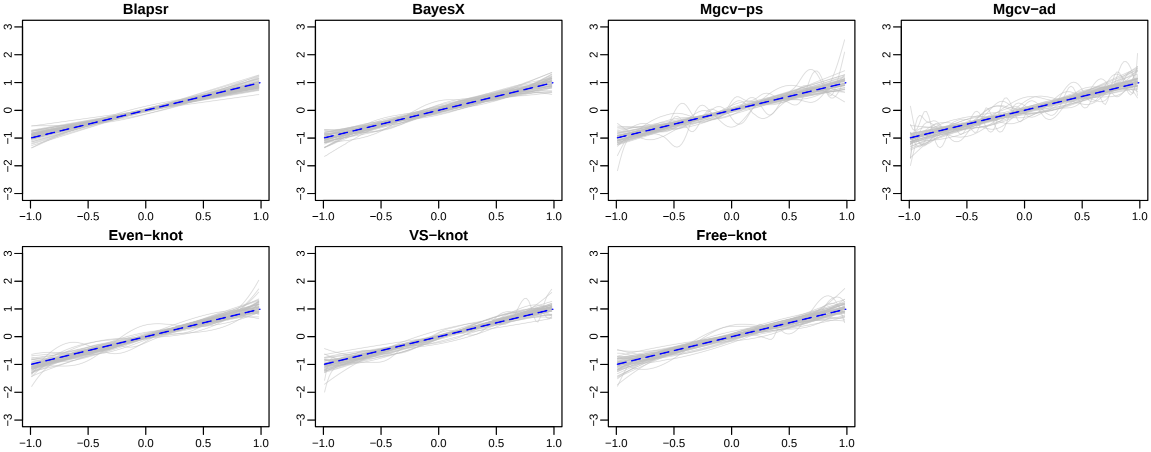

Figure: Pointwise posterior means of randomly chosen 50 replications (gray solid) and the true functions (blue dashed), $n = 1000$

Comparison with Other Methods

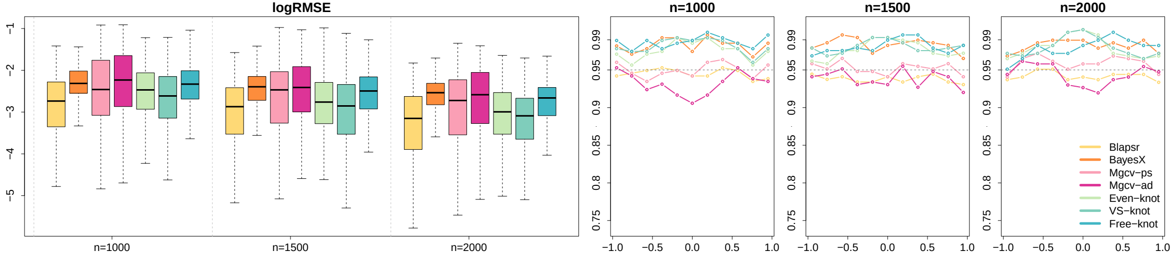

Figure: The log RMSE and the coverage probabilities for f_3 in the nonparametric logistic regression models with $n = 1000, 1500, 2000$.

Comparison with Other Methods

We conclude that $\cdots$

- **R2BayesX** and **Blapsr** often oversmooth with excessive penalization and unfit for locally varying functions, as expected.
- In general, **Mgcv** provides too wiggly estimates of the linear function, indicating undersmoothing for simple functions. **Mgcv-ad**, with local adaptation, works very well for spatially varying functions, but may lead to higher MSE and lower coverage for others.
- Among the BMS-based methods, **Even-knot** is fast and has performances comparable to the others except for the locally varying functions. **Free-knot** is similar in performance to **VS-knot** except for a linear function, but fare much worse in terms of sampling efficiency (a ratio of effective sample size to runtime) than **VS-knot**.

Comparison with Other Methods

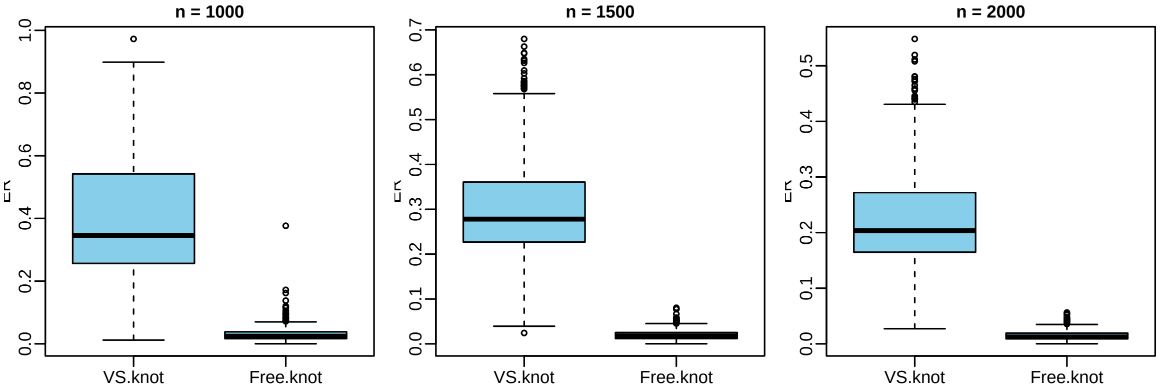

Figure: The efficiency ratio, the number of effective samples per one second of CPU runtime, in the nonparametric logistic regression models with $n = 1000, 1500, 2000$ for **VS-knot** and **Free-knot**.

V. Applications

Boston house price data

- The Boston housing dataset consists of housing information in the area of Boston for a total of $n = 506$ counties in the 1970s [Harrison Jr and Rubinfeld, 1978]. The variables in the dataset are descried in Table 2.
- Treating the log of the median house price as a response variable Y_i, we fit a Gaussian additive model,

$$\begin{aligned} Y_i = \alpha + \beta_1 chas_i + f_1(crim_i) + f_2(zn_i) + f_3(indus_i) + f_4(nox_i) + f_5(rm_i) + f_6(age_i) \\ + f_7(dis_i) + f_8(rad_i) + f_9(tax_i) + f_{10}(ptratio_i) + f_{11}(black_i) + f_{12}(lstat_i) + \epsilon_i, \qquad (10) \\ \epsilon_i \sim N(0, 1/\phi). \end{aligned}$$

- $chas$ is binary and assumed to have a linear effect. Each fixed dimensional parameter and nonparametric function is estimated by the VS-knot splines approach. For each nonparametric function, the number of knots M_j is reasonably chosen based on the observed predictor variables.

Boston house price data

Variable	Description
Y	Log of median value of owner-occupied homes in USD 1000's
$chas$	Charles River dummy variable ($= 1$ if tract bounds river; $= 0$ otherwise)
$crim$	Crime rate per capita by town
zn	Proportion of residential land zoned for lots over $25,000$ square feet
$indus$	Proportion of non-retail business acres per town
nox	Nitric oxides concentration
rm	Average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	Weighted distances to five Boston employment centers in log scale
rad	Index of accessibility to radial highways
tax	Full-value property-tax rate per USD $10,000$
$ptratio$	Pupil-teacher ratio by town
$black$	$1000(B - 0.63)^2$ where B is the proportion of African Americans by town
$lstat$	Percentage of lower status of population

Table: Description of the variables in the Boston housing dataset.

Boston house price data

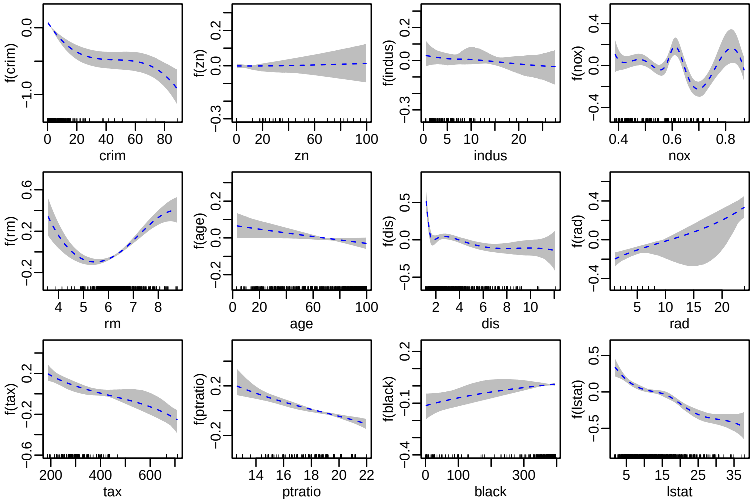

Figure: Pointwise posterior mean (blue dashed curve) and pointwise 95% credible band (gray shade) of the functions for the model in (10).

Boston house price data

Parameter	Mean	Median	95% lower limit	95% upper limit
α	3.0366	3.0366	3.0237	3.0499
β_1 $(chas)$	0.0287	0.0287	-0.0264	0.0829
$1/\sqrt{\phi}$	0.1418	0.1416	0.1328	0.1519

Table: Summary statistics of the posterior distribution for the model in (10).

Variable	$crim$	zn	$indus$	nox	rm	age
$\Pi(\lvert\xi_j\rvert = 0\lvert Y)$	0.00	0.88	0.78	0.00	0.00	0.89
Variable	dis	rad	tax	$ptratio$	$black$	$lstat$
$\Pi(\lvert\xi_j\rvert = 0\lvert Y)$	0.00	0.74	0.61	0.73	0.74	0.14

Table: Marginal posterior probabilities of linear effects.

Pima diabetes data

- The Pima diabetes dataset includes signs of diabetes and 7 potential risk factors of $n = 532$ Pima Indian women in Arizona [Smith et al., 1988]. The variables are summarized in Table 5.
- To model the sign of diabetes (0 or 1) as a response variable Y_i, we consider the following GAM with a logit link,

$$\begin{aligned} \log \frac{E(Y_i)}{1 - E(Y_i)} &= \alpha + f_1(pregnant_i) + f_2(glucose_i) + f_3(pressure_i) \\ &\quad + f_4(triceps_i) + f_5(mass_i) + f_6(pedigree_i) + f_7(age_i). \end{aligned} \tag{11}$$

- The observations with missing values are removed for analysis. Each nonparametric function is estimated by the VS-knot splines.

Pima diabetes data

Variable	Description
Y	Signs of diabetes according to WHO criteria (pos $= 1$, neg $= 0$)
$pregnant$	Number of times the subject was pregnant
$glucose$	Plasma glucose concentration in two hours in an oral glucose tolerance test $[mg/dl]$
$pressure$	Diastolic blood pressure $[mm/Hg]$
$triceps$	Triceps skin fold thickness $[mm/Hg]$
$mass$	Body Mass Index (BMI) $[kg/m^2]$
$pedigree$	Diabetes pedigree function [Smith et al., 1988]
age	Age [years]

Table: Description of the variables in the Pima diabetes data.

Pima diabetes data

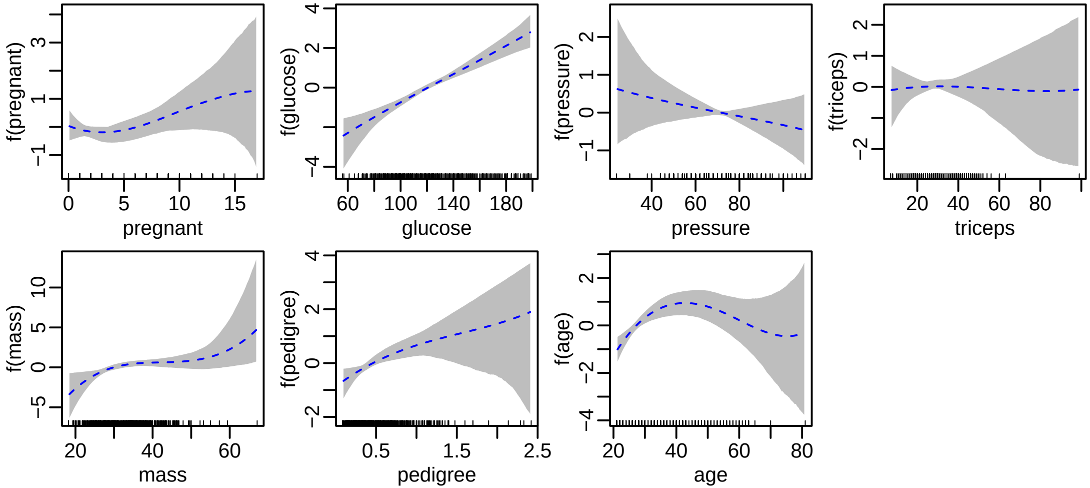

Figure: Pointwise posterior mean (blue dashed curve) and pointwise 95% credible band (gray shade) of the functions for the model in (11).

Pima diabetes data

Parameter	Mean	Median	95% lower limit	95% upper limit
α	-1.1567	-1.1556	-1.4039	-0.9066

Table: Summary statistics of the posterior distribution for the model in (11).

Variable	*pregnant*	*glucose*	*pressure*	*triceps*	*mass*	*pedigree*	*age*
$\Pi(\lvert\xi_j\rvert = 0 \mid Y)$	0.34	0.80	0.82	0.79	0.14	0.54	0.01

Table: Marginal posterior probabilities of linear effects.

Bibliography I

C Armero and MJ Bayarri. Prior assessments for prediction in queues. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 43(1):139–153, 1994.

Maria J Bayarri, James O Berger, Anabel Forte, and Gonzalo García-Donato. Criteria for bayesian model choice with application to variable selection. *The Annals of statistics*, 40(3):1550–1577, 2012.

Andreas Brezger and Stefan Lang. Generalized structured additive regression based on bayesian p-splines. *Computational Statistics & Data Analysis*, 50(4):967–991, 2006.

Paul Damlen, John Wakefield, and Stephen Walker. Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):331–344, 1999.

DGT Denison, BK Mallick, and AFM Smith. Automatic bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):333–350, 1998.

Ilaria DiMatteo, Christopher R Genovese, and Robert E Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, 2001.

Bibliography II

Robert G Edwards and Alan D Sokal. Generalization of the fortuin-kasteleyn-swendsen-wang representation and monte carlo algorithm. *Physical review D*, 38(6):2009, 1988.

Michael B Gordy. A generalization of generalized beta distributions. 1998.

Oswaldo Gressani and Philippe Lambert. Laplace approximations for fast bayesian inference in generalized additive models based on p-splines. *Computational Statistics & Data Analysis*, 154: 107088, 2021.

David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297 – 310, 1986. doi: 10.1214/ss/1177013604. URL https://doi.org/10.1214/ss/1177013604.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Leonhard Held, Daniel Sabanés Bové, and Isaac Gravestock. Approximate bayesian model selection with the deviance statistic. *Statistical Science*, pages 242–257, 2015.

Bibliography III

Pierre Humbert. Ix.—the confluent hypergeometric functions of two variables. *Proceedings of the Royal Society of Edinburgh*, 41:73–96, 1922.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90 (430):773–795, 1995.

Robert E Kass and Larry Wasserman. A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the american statistical association*, 90(431): 928–934, 1995.

Yingbo Li and Merlise A Clyde. Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association*, 113(524):1828–1845, 2018.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

Yuzo Maruyama and Edward I George. Fully bayes factors with a generalized g-prior. *The Annals of Statistics*, 39(5):2740–2765, 2011.

Bibliography IV

Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association, 1988.

Xinlei Wang and Edward I George. Adaptive bayesian criteria in variable selection for generalized linear models. *Statistica Sinica*, pages 667–690, 2007.

Andrew J Womack, Luis León-Novelo, and George Casella. Inference from intrinsic bayes' procedures under model selection and uncertainty. *Journal of the American Statistical Association*, 109(507): 1040–1053, 2014.

Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.