Generalized additive models using Bayesian model selection with mixtures of g-priors

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

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

[2]Department of Applied Statistics, Yonsei University

Main Contributions

- Reviewed and extended estimation methods for GAM via BMS using g-priors.
- Proposed a simple slice sampler algorithm to draw from a class of generalized beta distributions, including truncated compound hypergeometric distribution (tCCH).
- Explained different model selection behaviors of mixtures of g-priors in terms of Bayes Factor of adding redundant variable.

Bayesian Model Selection (BMS)

Generalized Additive Model (GAM) via BMS

- GAM is an extension of a linear model assuming, for $X_i \in R^p, i = 1, \cdots, n$,

$$h(E[y_i \mid X_i]) = \eta_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}), \qquad \sum_{i=1}^{n} f_j(x_{ij}) = 0$$

- f_j has a spline basis representation $f_j(\cdot) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(\cdot)$, where the basis $b_{jk}(\cdot)$ is linear for $k = 1$ and the others are determined by the knots $\xi_j = \{\xi_{j1}, \cdots, \xi_{jL_j}\}$.
- We use natural cubic spline basis for $b_{jk}(\cdot)$ so that for $K_j = 0$ it is reduced to linear.
- $\eta = (\eta_1, \cdots, \eta_n)^T$ is written as $\eta = \alpha 1_n + B\beta$ for $B = [B_1, \cdots, B_p] \in R^{n \times J}, J = \sum_{j=1}^{p} K_j$, $(B_j)_{i,k} = b_{jk}(x_{ij})$, where B is column-wise centered for identifiability.
- We are interested in $L: (\alpha, f_1, \cdots, f_p) \mapsto L(\alpha, f_1, \cdots, f_p)$ (e.g. pointwise estimate, CI, …)

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

i.e., via Bayesian Model Selection. To this end, our Bayesian model formulation is

$$\Pi(\alpha, \beta_\xi, \xi) = \Pi(\alpha)\Pi(\beta_\xi \mid \xi)\Pi(\xi)$$
$$\pi(\xi \mid Y) \propto \pi(\xi) p(Y \mid \xi)$$
$$p(Y \mid \xi) = \int\int p(Y \mid \alpha, \beta_\xi) d\Pi(\alpha) d\Pi(\beta_\xi \mid \xi)$$

Mixtures of g-priors for BMS

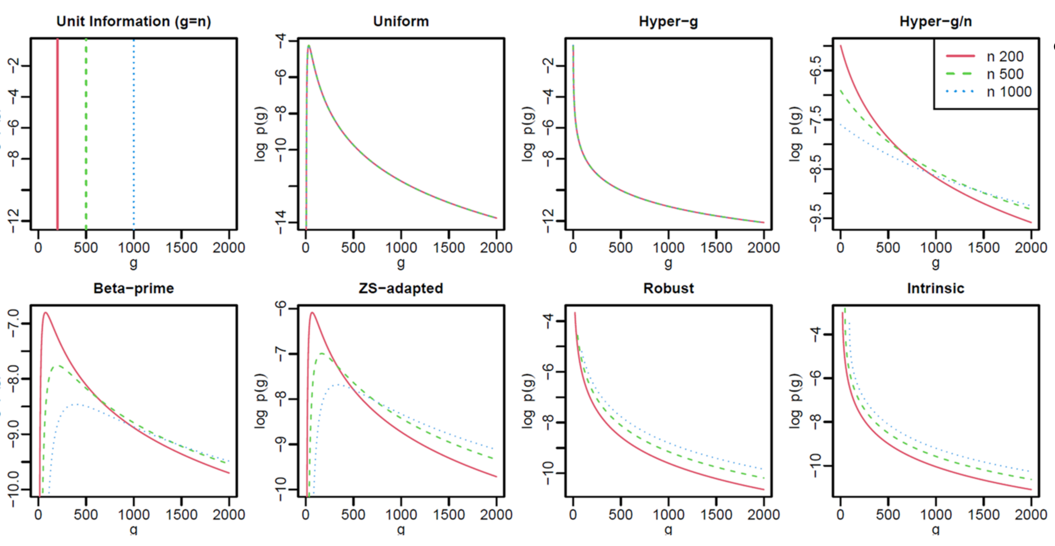
$$\frac{1}{g+1} \sim tCCH\left(\frac{a}{2}, \frac{b}{2}.r, \frac{s}{2}, \nu, \kappa\right)$$
$$\beta_\xi \mid g, \xi \sim N\left(0, g\left[\tilde{B}_\xi^T J_n(\hat{\eta}_\xi) \tilde{B}_\xi\right]^{-1}\right)$$

$$J_n(\hat{\eta}_\xi) = diag(-y_i \theta''(\hat{\eta}_{\xi,i}) + (b \circ \theta)''(\hat{\eta}_{\xi,i}), i = 1, \cdots, n)$$
$$\tilde{B}_\xi = \left[I_n - tr\left(J_n(\hat{\eta}_\xi)\right)^{-1} 1_n 1_n^T J_n(\hat{\eta}_\xi)\right] B_\xi$$

Mixtures of g prior

- For Laplace approximation of $p(Y \mid \xi)$, we use the variant of g prior for generalized linear model (Li and Clyde, 2018)

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



- Various mixtures of g-priors are classified into two groups according to the prior concentration:

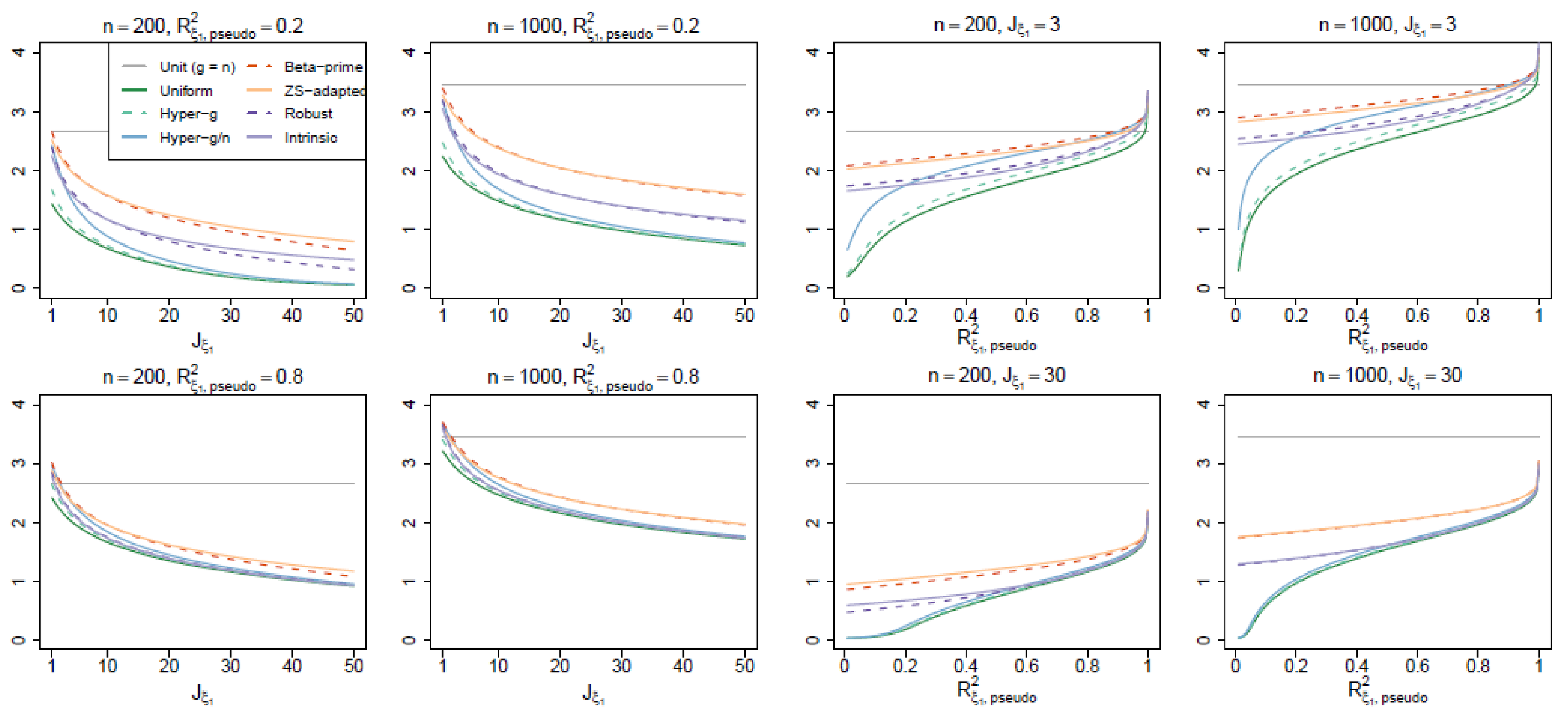
$g = O(1)$	$g = O(n)$
Uniform, Hyper-g	Hyper-g/n Robust, Intrinsic, Beta-prime, ZS-adapted,

Mixtures of g-priors as penalty functions

- The Bayes factor of two knots ξ_1, ξ_2 where $J_{\xi_1} = J_{\xi_2} + k$ ($k \in N^+$) but $\hat{\eta}_{\xi_1} = \hat{\eta}_{\xi_2}$ is

$$BF[\xi_1; \xi_2] = \begin{cases} (1+b)^{-k/2}, & \text{if } g \sim \delta_b(g), \\ E[(1+g)^{-k/2} \mid \xi_2, Y], & \text{if } g \sim tCCH(a/2, b/2, r, s/2, \nu, \kappa) \end{cases}$$

which is plotted below for $k = 1$: $R^2_{\xi,pseudo} = 1 - e^{-Q_\xi/n}$, *where* $Q_\xi = \hat{\beta}_\xi^T \tilde{B}_\xi^T J_n(\hat{\eta}_\xi) \tilde{B}_\xi \hat{\beta}_\xi$ *is the Wald statistics.*

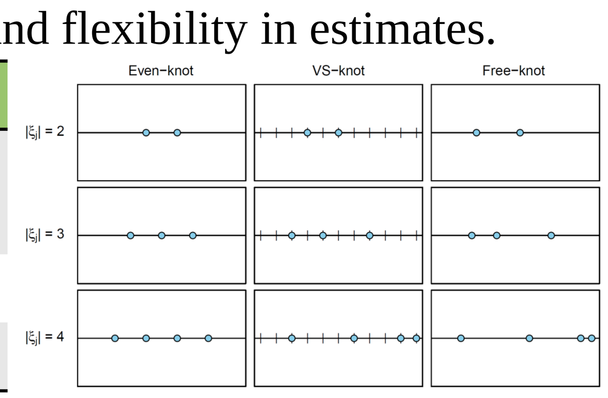


- $BF[\xi_1; \xi_2]$ is the penalty against the model ξ_1 by allowing k redundant variables to no avail ($\hat{\eta}_{\xi_1} = \hat{\eta}_{\xi_2}$). Whereas Unit information prior ($g = n$) yields a constant penalty regardless of J_{ξ_1} and the goodness-of-fit,
  1. mixtures of g-priors favor sparser models when comparing small models, but move towards more complex models when comparing large models, a trait desirable in capturing weak signals in the data.
  2. The $O(1)$ priors have the weakest penalty profiles among the mixtures of g-priors. Simulations showed that compared to the $O(n)$, the $O(1)$ priors tend to overfit to noise in the data.

Priors for knots

- A tradeoff exists between computational expedience and flexibility in estimates.

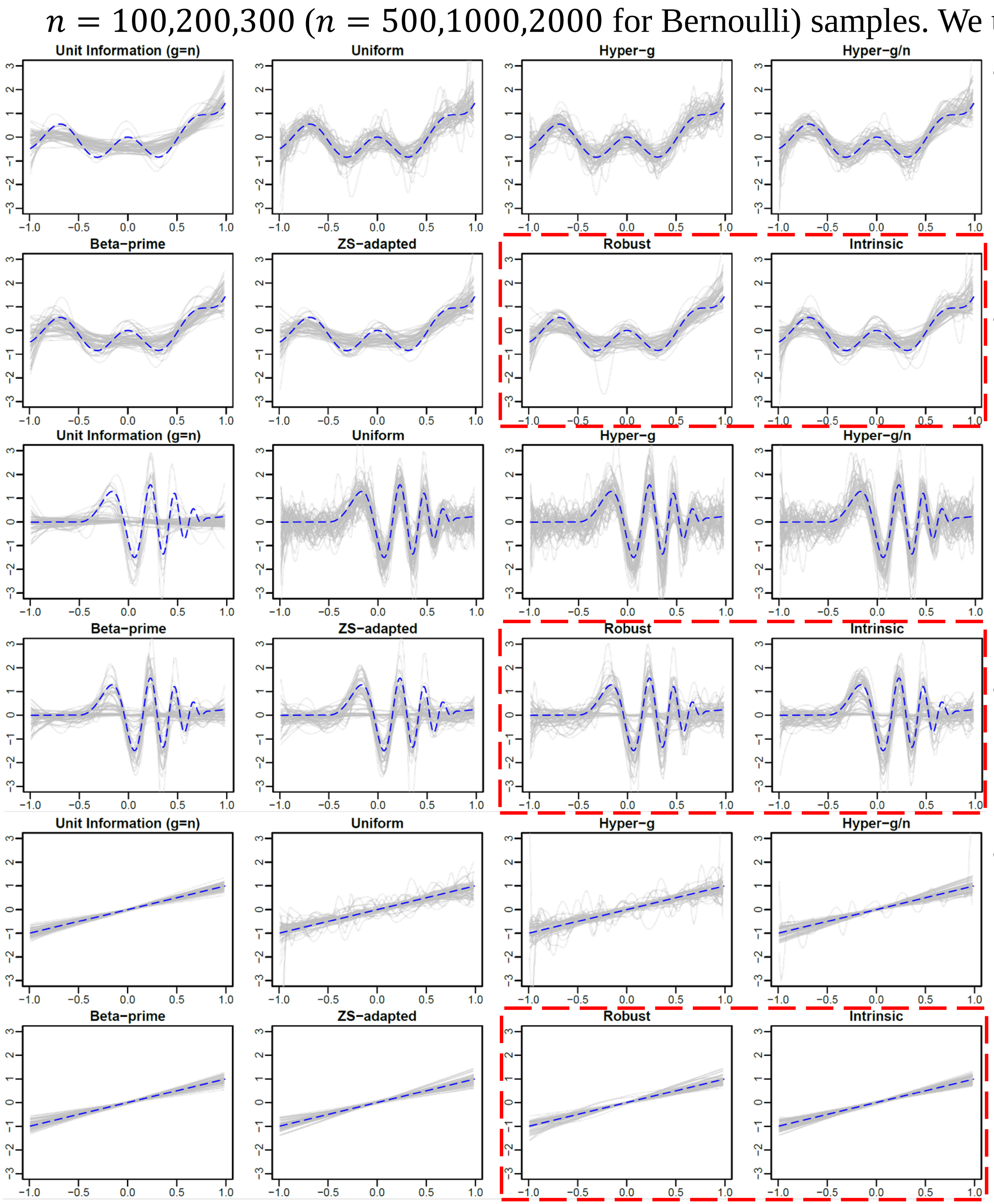
	Knot Locations	Posterior Sampling
Even-knot	Equidistant points	Direct enumeration Metropolis Hastings
VS-knot	Among grid points	Gibbs Sampling
Free-knot	Anywhere	Metropolis Hastings



Simulation Results

Comparison among the mixtures of g-priors

- For $f_j, j = 1,2,3$.test functions, we generated the univariate model $\eta_i = \alpha + f_j(x_i)$ where $x_j \sim Unif(-1,1)$ and the response follows Bernoulli, Poisson and Gaussian, for $n = 100,200,300$ ($n = 500,1000,2000$ for Bernoulli) samples. We used VS-knot.

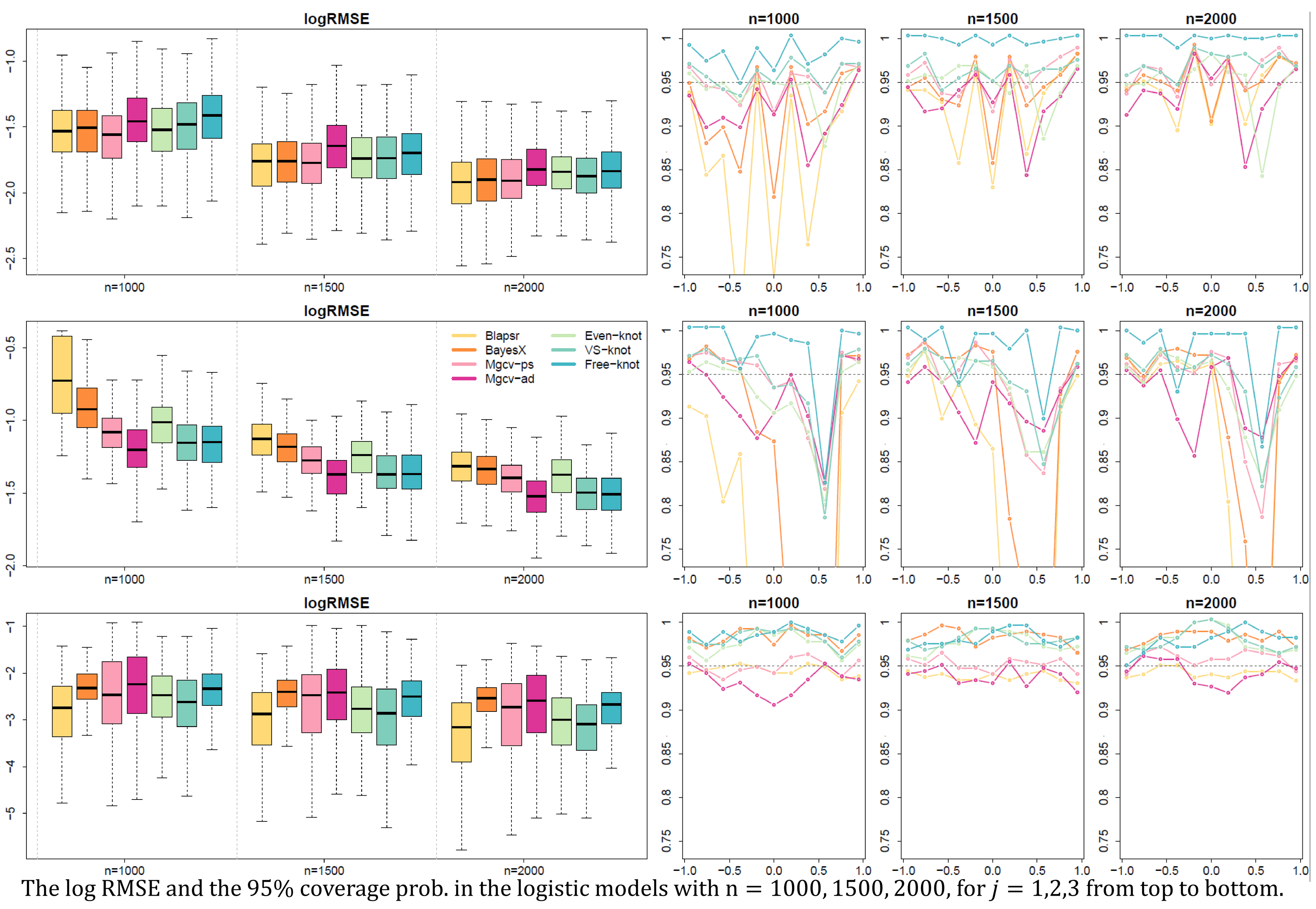


The result for n = 500 Bernoulli univariate model fitted via VS-knot. The true function is plotted in blue dashed and the posterior means for 50 of 500 replications in grey lines.

- Unit information prior generally underperforms, opting for simplistic models.
- Beta-prime and ZS-adapted prior exhibit under-smoothing, whereas Uniform, Hyper-g, and Hyper-g/n prior tend to over-smooth.
- The difference between the priors is blurred for a larger n.
- We recommend Robust and Intrinsic prior, and especially Robust prior for it is easier than Intrinsic to sample from the posterior.

Comparison with other methods

- The competitors, all based on the idea of Bayesian P-splines, include **R2BayesX**, **Blapsr**, and Mgcv with locally adaptive (**Mgcv-ad**) and non-adaptive estimation (**Mgcv-ps**).
- For $f_j, j = 1,2,3$.test functions, we generated the univariate model $\eta_i = \alpha + \sum_{j=1}^{3} f_j(x_i)$ where $x_j \sim Unif(-1,1)$ and the response follows Bernoulli, Poisson and Gaussian, for $n = 100,200,300$ ($n = 1000,1500,2000$ for Bernoulli) samples.
- **R2BayesX** and **Blapsr** often oversmooth the targets with excessive penalization, while **Mgcv** provides too wiggly estimates of the linear function, implying undersmoothing. **VS-knot** outperforms in most cases in terms of logRMSE and coverage probability.

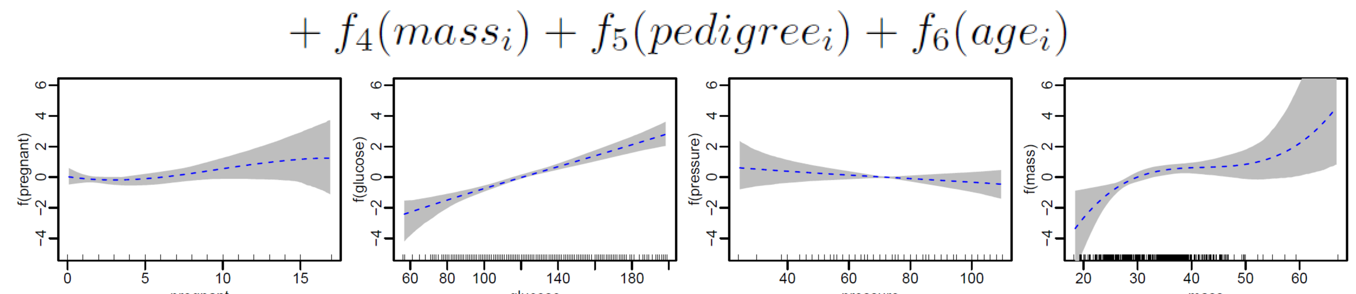


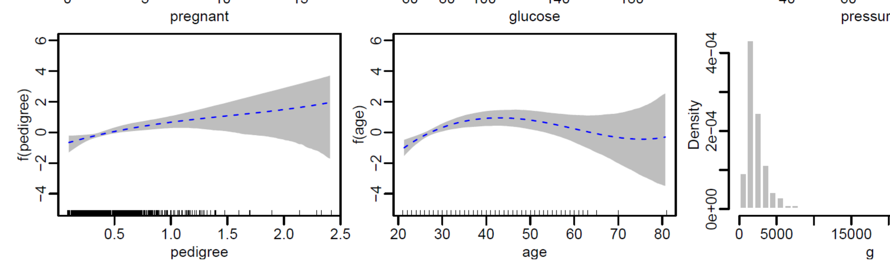
The log RMSE and the 95% coverage prob. in the logistic models with n = 1000, 1500, 2000, for $j = 1,2,3$ from top to bottom.

Real Data Applications (Pima Diabetes Data)

- Signs of diabetes (binary) of $n = 532$ women in Pima Indian population, Arizona.
- Our BMS-based methods provide posterior probability that a function is indeed linear.

$$\text{logit}(p_i) = \alpha + f_1(pregnant_i) + f_2(glucose_i) + f_3(pressure_i) + f_4(mass_i) + f_5(pedigree_i) + f_6(age_i)$$





Parameter	Mean	sd	Mean/sd	Lower 95%	Upper 95%
α	−1.1538	0.1271	−9.0782	−1.3996	−0.9014

Parameter estimation for the intercept term.

pregnant	*glucose*	*pressure*	*mass*	*pedigree*	*age*
0.37	0.81	0.84	0.13	0.54	0.0077

Marginal posterior probability of $J_{\xi_j} = 0$.

Li, Y., & Clyde, M. A. (2018). Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association, 113*(524), 1828-1845.