

Model Selection-Based Estimation for Generalized Additive Models Using Mixtures of g-priors: Towards Systematization

Gyeonghun Kang* and Seonghyun Jeong^{†,‡,§}

Abstract. We explore the estimation of generalized additive models using basis expansion in conjunction with Bayesian model selection. Although Bayesian model selection is useful for regression splines, it has traditionally been applied mainly to Gaussian regression owing to the availability of a tractable marginal likelihood. We extend this method to handle an exponential family of distributions by using the Laplace approximation of the likelihood. Although this approach works well with any Gaussian prior distribution, consensus has not been reached on the best prior for nonparametric regression with basis expansions. Our investigation indicates that the classical unit information prior may not be ideal for nonparametric regression. Instead, we find that mixtures of g-priors are more effective. We evaluate various mixtures of g-priors to assess their performance in estimating generalized additive models. Additionally, we compare several priors for knots to determine the most effective strategy. Our simulation studies demonstrate that model selection-based approaches outperform other Bayesian methods.

Keywords: Bayesian nonparametrics, exponential family models, mixtures of g-priors, nonparametric regression, regression splines.

MSC2020 subject classifications: Primary 62G08; secondary 62J12.

1 Introduction

Since its inception, the generalized additive model (GAM) has been pivotal in statistics and machine learning, garnering significant attention from both theorists and practitioners. The GAM represents an interpretable semiparametric approach that balances between parametric generalized linear models (GLMs) and fully nonparametric regression with multidimensional smoothing. Specifically, the GAM describes the relationship between multiple predictor variables and a (possibly non-Gaussian) response variable through an additive structure of univariate functions (Hastie and Tibshirani, 1986). This approach sacrifices the flexibility of multidimensional smoothing for a clear interpretation of each predictor variable's contribution to the mean as a univariate function.

Several estimation methods have been proposed for nonparametric regression and additive models, from both frequentist and Bayesian perspectives. Common Bayesian

*Department of Statistical Science, Duke University, Durham, North Carolina, USA

†Department of Statistics and Data Science, Yonsei University, Seoul, Korea

‡Department of Applied Statistics, Yonsei University, Seoul, Korea

§Corresponding author: sjeong@yonsei.ac.kr

techniques for estimating univariate smooth functions include Gaussian process priors (Williams and Rasmussen, 1995), Bayesian P-splines (Lang and Brezger, 2004), and basis expansion methods with model selection (Smith and Kohn, 1996; Denison et al., 1998a; DiMatteo et al., 2001). Among these approaches, basis expansion with Bayesian model selection (BMS), which we call the BMS-based approach to nonparametric regression, stands out for its theoretical benefits and empirical success (Smith and Kohn, 1996; Denison et al., 1998a; DiMatteo et al., 2001; Rivoirard and Rousseau, 2012; De Jonge and Van Zanten, 2012; Shen and Ghosal, 2015). BMS-based approaches determine suitable basis functions by comparing Bayes factors, thereby selecting more plausible basis terms in a data-driven manner. These methods are also useful for multidimensional smoothing, as seen in Bayesian multivariate adaptive regression splines (Denison et al., 1998b) and Bayesian additive regression trees (Chipman et al., 2010; Jeong and Rockova, 2023).

Despite their conceptual simplicity, BMS-based methods can be computationally challenging owing to the need for marginal likelihood calculations. This limitation has typically restricted the application of BMS to Gaussian regression within nonparametric regression contexts. For GLMs and GAMs, marginalization is often impractical even with conjugate priors on the coefficients (Chen and Ibrahim, 2003). The most feasible scenario often involves cases with available latent variable expressions, such as probit regression (e.g., Jeong et al., 2017; Sohn et al., 2023). When marginalization is not analytically tractable, BMS-based methods require numerical marginalization of the coefficients using Markov chain Monte Carlo (MCMC) algorithms, such as reversible jump MCMC (Green, 1995). These methods can be significantly less efficient than using Bayes factors unless a well-designed proposal distribution is available. This challenge has contributed to the early preference for P-spline-based Bayesian methods for estimating GAMs (e.g., Fahrmeir and Lang, 2001; Brezger and Lang, 2006).

A practical solution to this issue is to use an approximation of the likelihood, such as the Laplace approximation, which allows for the calculation of marginal likelihood with a Gaussian prior distribution on the coefficients (Li and Clyde, 2018). This approach enables the application of BMS-based methods to estimate GAMs with distributions from the exponential family. While the Laplace approximation has been occasionally used in BMS-based methods (e.g., DiMatteo et al., 2001), it is more widely accepted in the literature for improving computational efficiency in Bayesian P-splines for GAM estimation (Sabanés Bové et al., 2015; Gressani and Lambert, 2021).

When a Gaussian or Gaussian mixture prior is used for the coefficients, the Laplace approximation allows for a straightforward derivation of a closed-form expression for the marginal likelihood. However, the optimal prior distribution for basis determination remains unclear. Literature on variable selection indicates that mixture priors often outperform the classical Gaussian prior, known as Zellner’s g-prior, and its variants (Liang et al., 2008; Li and Clyde, 2018). Such mixture priors, also known as mixtures of g-priors, are preferred because of their desirable properties and ability to resolve issues associated with the g-prior (Liang et al., 2008). Various mixtures of g-priors have been proposed within the framework of linear regression (e.g., Zellner and Siow, 1980; Liang et al., 2008; Maruyama and George, 2011; Bayarri et al., 2012; Womack et al., 2014),

and some attempts have been made to extend them to GLMs (Sabanés Bové and Held, 2011; Held et al., 2015; Fouskakis et al., 2018). Recently, Li and Clyde (2018) provided a comprehensive framework for mixtures of g-priors for the GLM. However, the best-performing mixture prior for the BMS-based GAM estimation remains uncertain. To address this, understanding how mixtures of g-priors affect the penalization of nonparametric functions is essential. Even within Gaussian additive regression, determining the best mixture prior remains unresolved.

Another important consideration in BMS-based methods is selecting a prior distribution for the intrinsic basis terms. Since spline basis functions are often determined by the knot locations, this implies a prior on the knots. Various prior distributions have been proposed to balance computational efficiency and estimation quality (e.g., Smith and Kohn, 1996; Denison et al., 1998a; DiMatteo et al., 2001; Shen and Ghosal, 2015). These priors can be categorized based on their underlying principles; generally, more flexible priors offer better approximation but come with greater computational costs. The most suitable class of prior distribution for determining optimal spline knots remains unclear.

This study makes three key contributions. First, we systematize BMS-based approaches for GAM estimation by using the Laplace approximation and a unified framework for mixtures of g-priors, as proposed by Li and Clyde (2018). In doing so, we enhance computational efficiency by introducing a new form of natural cubic spline function specifically tailored for BMS-based methods. Second, among various mixtures of g-priors within a general class, we identify the default mixture prior for GAM estimation. We deepen our understanding of how mixtures of g-priors penalize the model during GAM estimation and evaluate the empirical performance of different mixture priors through extensive simulations. Our findings suggest that the traditional g-prior, also known as the unit information prior (Kass and Wasserman, 1995), may be less suitable. Instead, a mixture of g-priors is recommended. Finally, we categorize prior distributions for knots into three groups and assess which class is most effective for GAM estimation. Our investigation reveals that a prior distribution balancing flexibility and computational efficiency performs best. Specifically, while the most flexible prior, the free-knot spline (Denison et al., 1998a; DiMatteo et al., 2001), may be excessive in practice, a less flexible but computationally efficient approach based on variable selection (Smith and Kohn, 1996) yields better empirical results with fast mixing. We support these findings with various numerical results. The R package implementing the sampling algorithms for our GAM estimation is available on the first author’s GitHub page.¹

The remainder of this paper is organized as follows. Section 2 introduces the construction of GAMs using spline basis expansion with natural cubic splines. Section 3 discusses mixtures of g-priors for BMS within a unified framework and compares these priors for GAM estimation, interpreting them as penalty functions for nonparametric regression. Section 4 categorizes prior distributions for knots into three strategies and evaluates their effectiveness for BMS-based approaches. Section 5 presents comprehensive simulations and numerical studies to identify the best prior distribution and

¹<https://github.com/hun-learning94/gambms>

compare BMS-based methods with other approaches for GAM estimation. Section 6 applies the BMS-based method to the Pima diabetes dataset. Finally, Section 7 concludes the study with a discussion. Supplementary material (Kang and Jeong, 2024) includes proofs of propositions, additional simulation studies, and instructions for installing the R package.

2 Generalized additive models via basis expansion

For given predictor variables $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T \in \mathbb{R}^p$, suppose the response variable $Y_i \in \mathbb{R}$ follows a distribution from the exponential family. The density of Y_i is given by

$$y_i \mapsto p(y_i; \theta_i, \phi) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right), \quad i = 1, \dots, n, \quad (2.1)$$

where θ_i is the natural parameter modeled by x_i , ϕ is a scale parameter, and b and c are known functions. The dependence of θ_i on x_i is clarified below. Although we focus primarily on cases where the dispersion parameter ϕ is known, we also consider Gaussian regression with an unknown ϕ in Section S5 of the supplementary material. Assuming b is twice differentiable with $b''(\theta_i) > 0$, the expected value and variance of Y_i are $E(Y_i) = b'(\theta_i)$ and $Var(Y_i) = \phi b''(\theta_i)$, respectively. We use a monotonically increasing link function h to parameterize the natural parameter as $\theta_i = (h \circ b')^{-1}(\eta_i)$, where η_i is an additive predictor defined as

$$\eta_i = \alpha + \sum_{j=1}^p f_j(x_{ij}), \quad i = 1, \dots, n, \quad (2.2)$$

with a global mean α and univariate functions $f_j : \mathbb{R} \rightarrow \mathbb{R}$, $j = 1, \dots, p$. To ensure identifiability, we assume that the functions f_j satisfy the restriction $\sum_{i=1}^n f_j(x_{ij}) = 0$, $j = 1, \dots, p$.

The key aspect of the model specification is determining how to characterize the nonparametric functions f_j . In this study, the functions f_j are parameterized using a spline basis representation. Specifically, f_j are expressed as linear combinations of K_j basis functions b_{j1}, \dots, b_{jK_j} ; that is, with coefficients $\beta_{jk} \in \mathbb{R}$,

$$f_j(\cdot) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(\cdot), \quad j = 1, \dots, p.$$

To satisfy the identifiability condition $\sum_{i=1}^n f_j(x_{ij}) = 0$, we assume that each basis function satisfies $\sum_{i=1}^n b_{jk}(x_{ij}) = 0$, $j = 1, \dots, p$. This can be achieved by centering unrestricted basis functions b_{jk}^* as

$$b_{jk}(\cdot) = b_{jk}^*(\cdot) - \frac{1}{n} \sum_{i=1}^n b_{jk}^*(x_{ij}), \quad j = 1, \dots, p, \quad k = 1, \dots, K_j. \quad (2.3)$$

Let $B_j \in \mathbb{R}^{n \times K_j}$ be the matrix whose (i, k) th component is $b_{jk}(x_{ij})$. The centering procedure is achieved by the projection $B_j = (I_n - n^{-1}1_n 1_n^T)B_j^*$ with the unrestricted basis matrix B_j^* defined with b_{jk}^* for its (i, k) th component. We define $B = [B_1, \dots, B_p] \in \mathbb{R}^{n \times J}$ and a vector of full coefficients $\beta = (\beta_{11}, \dots, \beta_{1K_1}, \dots, \beta_{p1}, \dots, \beta_{pK_p})^T \in \mathbb{R}^J$, where $J = \sum_{j=1}^p K_j$. The vector of additive predictors $\eta = (\eta_1, \dots, \eta_n)^T$ can then be written as $\eta = \alpha 1_n + B\beta$.

Various classes of basis functions can be used to estimate smooth functions. In this study, we employ natural cubic spline basis functions to avoid erratic behavior near the boundaries. This approach is equivalent to using any piecewise polynomial basis function (including B-splines) with appropriate natural boundary conditions, provided that the prior distribution remains invariant under linear bijections of the design matrix. For boundary knots $\{t^L, t^U\}$ and a set of M interior knots $\{t_1, \dots, t_M\}$ satisfying $-\infty < t^L < t_1 < \dots < t_M < t^U < \infty$, we define the natural cubic spline basis functions $N_k : \mathbb{R} \rightarrow \mathbb{R}$, $k = 1, \dots, M + 1$, as follows:

$$\begin{aligned} N_1(u) &= u, \\ N_{k+1}(u) &= N(u; t^L, t^U, t_k) \\ &\equiv \frac{(u - t_k)_+^3 - (u - t^U)_+^3}{t^U - t_k} - \frac{(u - t^L)_+^3 - (u - t^U)_+^3}{t^U - t^L}, \quad k = 1, \dots, M. \end{aligned} \tag{2.4}$$

Combined with the constant term $N_0(u) = 1$, the basis functions in (2.4) generate piecewise cubic functions. These functions are linear beyond the boundary knots $\{t^L, t^U\}$, enhancing stability near the boundaries owing to the constraints imposed at $\{t^L, t^U\}$. The constant term is excluded from (2.4), as it is redundant given the intercept term. Our findings, consistent with well-known observations, show that natural cubic splines significantly reduce estimation bias near boundaries compared to cubic splines without natural conditions.

Although the basis construction in (2.4) is based on a truncated power series, our definition differs slightly from the truncated power natural cubic splines typically used in the literature, such as those in Equations (5.4) and (5.5) of Hastie et al. (2009). The basis terms in (2.4) span the same piecewise cubic polynomial space with natural boundary conditions, as demonstrated. However, our definition in (2.4) has an additional advantageous property: inserting a new knot-point $t_* \in (t^L, t^U)$ simply adds a new basis term $N(\cdot; t^L, t^U, t_*)$ to the set $\mathcal{N} = \{N_k, k = 0, 1, \dots, M + 1\}$ without altering the existing basis terms in \mathcal{N} . Similarly, removing a knot-point simply deletes an existing basis term in \mathcal{N} . This feature may not be present in other natural cubic spline basis functions, such as the natural cubic B-spline basis or those in Equations (5.4) and (5.5) of Hastie et al. (2009), where a single basis term might depend on more than two knots, and adding or removing a knot-point could alter other basis terms. This characteristic makes the basis terms in (2.4) more attractive for model selection-based approaches (see Sections 4.2 and 4.3) because it allows faster computation by reducing the time spent expanding the design matrix at each iteration. For related simulation results, see Section S7 of the supplementary material. To the best of our knowledge, this is the first study to use the form of natural cubic splines in (2.4). These properties are formalized as follows.

Proposition 1. *The set $\mathcal{N} = \{N_k, k = 0, 1, \dots, M + 1\}$ is a basis for the cubic spline space with natural boundary conditions.*

Proposition 2. *The addition of a new interior knot-point $t_* \in (t^L, t^U)$ introduces the corresponding basis term $N(\cdot; t^L, t^U, t_*)$ into \mathcal{N} . Similarly, the elimination of an existing interior knot-point $t_k \in t$ eliminates the corresponding basis term $N(\cdot; t^L, t^U, t_k)$ in \mathcal{N} .*

Proofs are provided in Section S2 of the supplementary material. We select our basis terms b_{jk}^* using the natural cubic spline basis functions defined in (2.4). Specifically, for each j , we set the boundary knots to $\xi_j^L = \min_{1 \leq i \leq n} x_{ij}$ and $\xi_j^U = \max_{1 \leq i \leq n} x_{ij}$ based on the observed design points. With a given set of knots $\xi_j = \{\xi_{j1}, \dots, \xi_{jL_j}\}$ where $\xi_j^L < \xi_{j1} < \dots < \xi_{jL_j} < \xi_j^U$, the uncentered basis terms are chosen as

$$b_{j1}^*(\cdot) = N_1(\cdot), \quad b_{j,k+1}^*(\cdot) = N(\cdot; \xi_j^L, \xi_j^U, \xi_{jk}), \quad k = 1, \dots, L_j. \quad (2.5)$$

The class of spline functions is highly dependent on the placement of knots $\xi = \{\xi_1, \dots, \xi_p\}$. Therefore, choosing suitable knot locations is crucial for accurately capturing both local and global functional characteristics while avoiding overfitting. From a Bayesian perspective, a natural approach is to let the data select the most appropriate knots ξ from a predetermined set Ξ using BMS. This approach is well-established in the literature (e.g., Smith and Kohn, 1996; Denison et al., 1998a; DiMatteo et al., 2001; Rivoirard and Rousseau, 2012; De Jonge and Van Zanten, 2012; Shen and Ghosal, 2015; Jeong and Park, 2016; Jeong et al., 2017). A set Ξ can be a countable or uncountable collection of knots. A richer Ξ allows for more flexible estimation of regression spline functions but may result in computational inefficiency. Within the Bayesian framework, specifying Ξ via a predetermined law is akin to assigning a prior distribution to ξ over an infinite-dimensional space with restricted support Ξ . The key to success is placing a prior on ξ with appropriately restricted support Ξ . Several options for specifying a prior for ξ are discussed in Section 4.

An additional advantage of the formulation in (2.5) is its ability to easily characterize a fully linear relationship. Specifically, if ξ_j is empty, the basis consists only of the linear term b_{j1}^* . This is particularly useful when a predictor variable is binary or assumed to have a linear effect. In such cases, we can assign a point mass prior to empty ξ_j . As a result, generalized additive partial linear models (GAPLMs) with both parametric and nonparametric additive terms (Wang et al., 2011) are naturally accommodated by our construction without modification. Additionally, while not explored in this study, variable selection could be incorporated by introducing additional latent variables for the linear basis term b_{j1}^* . A related idea is discussed in Jeong et al. (2022).

One major advantage of BMS-based approaches to nonparametric regression is that they provide model-averaged estimates rather than relying on specific knot locations. Our goal is to examine the model-averaged estimates of a functional $\mathcal{L} : (\alpha, f_1, \dots, f_p) \mapsto \mathcal{L}(\alpha, f_1, \dots, f_p)$, which is parameterized by coefficients α and β . For example, we may be interested in a pointwise evaluation of the additive predictor $\alpha + \sum_{j=1}^p f_j(x_j)$ or the univariate function $f_j(x_j)$, $j = 1, \dots, p$, at a given point $x = (x_1, \dots, x_p)^T$. The

model-averaged posterior of a functional is given by

$$\pi(\mathcal{L}(\alpha, f_1, \dots, f_p) | Y) = \int_{\Xi} \pi(\mathcal{L}(\alpha, f_1, \dots, f_p) | \xi, Y) d\Pi(\xi | Y). \quad (2.6)$$

A key aspect of our Bayesian procedure is assigning a prior distribution for model selection and exploring the posterior distribution of ξ , $\Pi(\xi | Y)$. To highlight the dependency on ξ , we use the notation $B_\xi = B$, $\beta_\xi = \beta$, $J_\xi = J$, and $\eta_\xi = \alpha 1_n + B_\xi \beta_\xi$. Note that $J_\xi = p + \sum_{j=1}^p |\xi_j|$, where $|\xi_j|$ represents the number of knots ξ_j , $j = 1, \dots, p$.

3 Mixtures of g-priors for generalized additive models

Our main objective is to explore the posterior distribution of a functional $\mathcal{L}(\alpha, f_1, \dots, f_p)$. To obtain a model-averaged estimate, we need to numerically evaluate the integral in (2.6), which involves exploring the posterior distribution $\Pi(\alpha, \beta_\xi, \xi | Y)$. Therefore, we need to specify a prior distribution $\Pi(\alpha, \beta_\xi, \xi)$ over the parameter space. The possible priors for ξ , $\Pi(\xi)$, are discussed in Section 4. A critical aspect is determining a prior for the knot-specific coefficients β_ξ , that is, $\Pi(\beta_\xi | \xi)$. This study employs mixtures of g-priors for this purpose. In this section, we explain the use of mixtures of g-priors in BMS-based approaches to GAMs and discuss the resulting posteriors. Additionally, we provide a toy example to illustrate how mixed priors penalize GAMs.

3.1 Mixtures of g-priors for exponential family models

We specify the prior distribution as $\Pi(\alpha, \beta_\xi | \xi) = \Pi(\alpha)\Pi(\beta_\xi | \xi)$. In line with common practice, we assign an improper uniform prior to the intercept parameter α , that is,

$$\pi(\alpha) \propto 1. \quad (3.1)$$

This improper prior has been justified in the literature (Berger et al., 1998; Bayarri et al., 2012). Next, we discuss $\Pi(\beta_\xi | \xi)$. For model selection in linear regression, Zellner's g-prior is often preferred owing to its computational efficiency and invariance to linear transformations (Zellner, 1986). In our spline setup, this invariance is particularly valuable because it ensures that the procedure remains unaffected by specific choices of basis functions, as long as the target spline space is correctly generated. Therefore, the invariance property of the g-prior supports the spline basis system defined in (2.5). However, the computational advantage of the g-prior is typically diminished in GAMs because Gaussian priors are not conjugate to non-Gaussian models, making it impossible to obtain a closed-form expression for the marginal likelihood $p(Y | \xi)$. This complicates the computation of the posterior distribution in (2.6) owing to the intractability of the marginal likelihood. To address this issue, we consider approximating the likelihood using the Laplace approximation with a suitable variant of the g-prior.

Let $\theta = (h \circ b')^{-1}$, and define $\mathcal{J}_n(\hat{\eta}_\xi) = \text{diag}(-Y_i \theta''(\hat{\eta}_{\xi,i}) + (b \circ \theta)''(\hat{\eta}_{\xi,i}), i = 1, \dots, n)$ as the observed information matrix of η_ξ evaluated at $\hat{\eta}_\xi$ (the Hessian matrix of the negative log-likelihood), where $\hat{\eta}_\xi = (\hat{\eta}_{\xi,1}, \dots, \hat{\eta}_{\xi,n})^T = \hat{\alpha}_\xi 1_n + B_\xi \hat{\beta}_\xi$ with the maximum

likelihood estimators $\hat{\alpha}_\xi$ and $\hat{\beta}_\xi$ (assuming they exist). We focus on cases where $\mathcal{J}_n(\hat{\eta}_\xi)$ is positive definite, which is generally true except in extreme situations like complete separation in logistic regression (Li and Clyde, 2018). Among the variants of the g-prior for exponential family models, we use the form proposed by Li and Clyde (2018),

$$\beta_\xi \mid g, \xi \sim N(0, g(\tilde{B}_\xi^T \mathcal{J}_n(\hat{\eta}_\xi) \tilde{B}_\xi)^{-1}), \quad (3.2)$$

where $g > 0$ serves as a dispersion factor that controls the influence of the prior, and $\tilde{B}_\xi = [I_n - \text{tr}(\mathcal{J}_n(\hat{\eta}_\xi))^{-1} \mathbf{1}_n \mathbf{1}_n^T \mathcal{J}_n(\hat{\eta}_\xi)] B_\xi$ is the matrix consisting of the columns of B_ξ centered by the weighted average with the diagonal elements of $\mathcal{J}_n(\hat{\eta}_\xi)$. The prior in (3.2) requires that $\tilde{B}_\xi^T \mathcal{J}_n(\hat{\eta}_\xi) \tilde{B}_\xi$ be invertible. This condition is satisfied if and only if B_ξ has full-column rank (observe that $\mathcal{J}_n(\hat{\eta}_\xi)$ is positive definite and $\text{rank}(B_\xi) = \text{rank}(\tilde{B}_\xi)$, where $\text{rank}(\cdot)$ is the rank of a matrix). Thus, a full-column rank condition will be imposed on $\Pi(\xi)$ in Section 4. Although the prior in (3.2) could be extended using a generalized inverse, we do not pursue this approach here (for further discussion, see Section 2.5 of Li and Clyde (2018)).

In addition to the prior in (3.2), many other variants of the g-prior exist for exponential family models (e.g., Hansen and Yu, 2003; Wang and George, 2007; Gupta and Ibrahim, 2009; Sabanés Bové and Held, 2011; Held et al., 2015). We note that the prior in (3.2) depends on the observation vector Y , which means it does not strictly adhere to the pure Bayesian philosophy. Some methods address this issue by using the expected information matrix instead of $\mathcal{J}_n(\hat{\eta}_\xi)$, while substituting $\eta_\xi = \alpha \mathbf{1}_n$ based on the null model (Sabanés Bové and Held, 2011; Held et al., 2015; Castellanos et al., 2021; García-Donato et al., 2023). However, within this framework, the marginal likelihood is not available in closed form unless α is fixed and a specific prior on g is used (Sabanés Bové and Held, 2011; Held et al., 2015). In contrast, the prior in (3.2) provides a convenient expression for the approximate marginal likelihood, enabling relatively fast computation. Moreover, our prior captures the large-sample covariance structures and local geometry better than other variants of the g-prior (Li and Clyde, 2018).

By integrating the second-order Taylor expansion of the likelihood with the priors specified in (3.1) and (3.2), we obtain

$$p(Y \mid g, \xi) \approx p(Y \mid \hat{\eta}_\xi) \text{tr}(\mathcal{J}_n(\hat{\eta}_\xi))^{-1/2} (g+1)^{-J_\xi/2} \exp\left(-\frac{Q_\xi}{2(g+1)}\right), \quad (3.3)$$

where $p(Y \mid \hat{\eta}_\xi)$ represents the likelihood evaluated at $\hat{\eta}_\xi$ for a given ξ and $Q_\xi = \hat{\beta}_\xi^T \tilde{B}_\xi^T \mathcal{J}_n(\hat{\eta}_\xi) \tilde{B}_\xi \hat{\beta}_\xi$ is the Wald statistic; see Section S3 of the supplementary material for the derivation of (3.3). The expression in (3.3) shows that when g is treated as a fixed hyperparameter, the marginal likelihood becomes highly sensitive to its value. Determining an appropriate choice for g has been widely discussed in the literature. The most common approach is to set $g = n$, known as the unit information prior (Kass and Wasserman, 1995). This concept is also frequently used in the literature on nonparametric regression using BMS (e.g. Gustafson, 2000; DiMatteo et al., 2001; Kohn et al., 2001). From a Bayesian perspective, the unit information prior can be viewed as a point mass prior at $g = n$, expressed as $\Pi(g) = \delta_n(g)$, where δ_b denotes the Dirac

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

Table 1: Distributions belonging to the tCCH family.

measure at b . However, research has shown that using a suitable prior distribution for g , known as a mixture of g-priors, enhances empirical performance and addresses paradoxes in BMS (Liang et al., 2008; Li and Clyde, 2018). To unify various mixtures of g-priors, we adopt a general family that encompasses various mixture distributions. Specifically, following Li and Clyde (2018), we assign the truncated compound confluent hypergeometric (tCCH) distribution to $(g + 1)^{-1}$ (Gordy, 1998), that is,

$$\frac{1}{g+1} \sim \text{tCCH}\left(\frac{a}{2}, \frac{b}{2}, r, \frac{s}{2}, \nu, \kappa\right), \quad a, b, \kappa > 0, \quad r, s \in \mathbb{R}, \quad \nu \geq 1. \quad (3.4)$$

The tCCH distribution is a type of generalized beta distribution characterized by five parameters, which allow it to exhibit multi-modal or long-tailed density. Parameters a and b behave similarly to those in a beta distribution, while parameters r , s , and κ control the skewness of the density. Parameter ν determines the support of the distribution. For a detailed discussion, including the density function and moments of the tCCH distribution, see Section S1 of the supplementary material.

Table 1 presents several distributions from the tCCH family, including the uniform prior (on $(g+1)^{-1}$), the hyper-g and hyper-g/n priors (Liang et al., 2008), the beta-prime prior (Maruyama and George, 2011), the Zellner Siow (ZS)-adapted prior (Held et al., 2015), the robust prior (Bayarri et al., 2012), and the intrinsic prior (Womack et al., 2014). Note that the beta-prime prior is only proper if $J_\xi < n - 1$, so this constraint needs to be incorporated into $\Pi(\xi)$ when using the beta-prime prior. According to Li and Clyde (2018), prior distributions can be classified into two categories based on their concentration: $g = O(1)$ and $g = O(n)$. (This notation can be misleading, as it refers to the concentration order of the distribution rather than the actual value of g ; Maruyama and George (2011) uses the same notation.) Figure 1 illustrates the concentration behavior of each prior distribution on g .

We define the confluent hypergeometric function of two variables (Gordy, 1998) as $\Phi_1(\alpha, \beta, \gamma, x, y) = B(\alpha, \gamma - \alpha)^{-1} \int_0^1 u^{\alpha-1} (1-u)^{\gamma-\alpha-1} (1-yu)^{-\beta} e^{xu} du$, for $\gamma > \alpha > 0$,

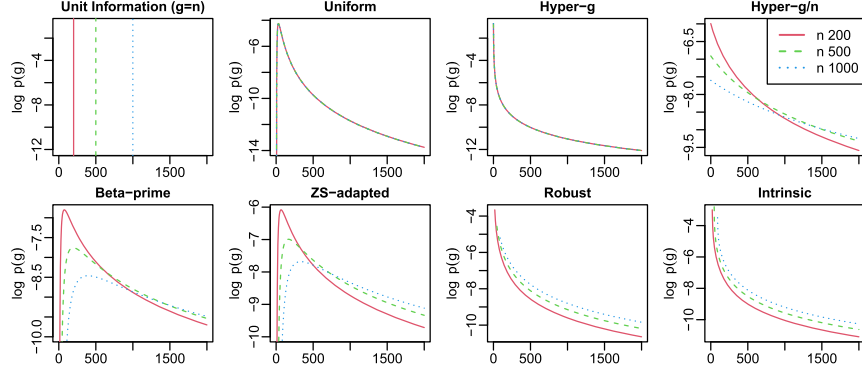


Figure 1: Distributions belonging to the tCCH family for $n = 200, 500, 1000$, with $J_\xi = 10$ if required.

$\beta > 0$, $x \in \mathbb{R}$, and $y < 1$.² The resulting marginal likelihood is expressed as

$$\begin{aligned}
 p(Y | \xi) &= p(Y | \hat{\eta}_\xi) \text{tr}(\mathcal{J}_n(\hat{\eta}_\xi))^{-1/2} \nu^{-J_\xi/2} \exp\left(-\frac{Q_\xi}{2\nu}\right) \frac{B((a + J_\xi)/2, b/2)}{B(a/2, b/2)} \\
 &\times \Phi_1\left(\frac{b}{2}, r, \frac{a + b + J_\xi}{2}, \frac{s + Q_\xi}{2\nu}, 1 - \kappa\right) \Big/ \Phi_1\left(\frac{b}{2}, r, \frac{a + b}{2}, \frac{s}{2\nu}, 1 - \kappa\right),
 \end{aligned} \tag{3.5}$$

where $B(\cdot, \cdot)$ denotes the beta function. The derivation of (3.5) is detailed in Section S3 of the supplementary material. Generally, Φ_1 cannot be evaluated analytically and requires numerical approximation. For this purpose, we utilize the Gaussian-Kronrod quadrature routine available in the Boost C++ library.

The approximate posterior for $((g + 1)^{-1}, \alpha, \beta_\xi)$ conditional on ξ is given by

$$\begin{aligned}
 \frac{1}{g + 1} | Y, \xi &\sim \text{tCCH}\left(\frac{a + J_\xi}{2}, \frac{b}{2}, r, \frac{s + Q_\xi}{2}, \nu, \kappa\right), \\
 \beta_\xi | Y, g, \xi &\sim \text{N}\left(\frac{g}{g + 1} \hat{\beta}_\xi, \frac{g}{g + 1} (\tilde{B}_\xi^T J(\hat{\eta}_\xi) \tilde{B}_\xi)^{-1}\right), \\
 \alpha | Y, g, \beta_\xi, \xi &\sim \text{N}\left(\hat{\alpha}_\xi - \text{tr}(\mathcal{J}_n(\hat{\eta}_\xi))^{-1} \mathbf{1}_n^T \mathcal{J}_n(\hat{\eta}_\xi) B_\xi (\beta_\xi - \hat{\beta}_\xi), \text{tr}(\mathcal{J}_n(\hat{\eta}_\xi))^{-1}\right).
 \end{aligned} \tag{3.6}$$

The derivation of (3.6) is detailed in Section S3 of the supplementary material. This expression is also applicable for the unit information prior by substituting the first line with the point mass posterior $\Pi(g | Y, \xi) = \delta_n(g)$. Sampling from tCCH distributions can be performed using MCMC, but exact sampling is possible with certain prior specifications. Specifically, if the uniform prior, hyper-g prior, ZS-adapted prior, or robust prior is used, the first line of (3.6) simplifies to a truncated gamma distribution, making exact sampling straightforward. For the remaining priors, slice sampling with data

²These parameter ranges ensure that Φ_1 is finite, positive, and real; see Theorem 1 of Gordy (1998).

augmentation can be employed. Details on the sampling procedures are provided in Section S4 of the supplementary material. The joint posterior $\Pi(\alpha, \beta_\xi, \xi, g | Y)$ is fully specified by the posterior in (3.6) and the marginal posterior of ξ , $\Pi(\xi | Y)$. The latter is obtained by specifying a prior $\Pi(\xi)$ as described in Section 4 and using the approximate marginal likelihood $p(Y | \xi)$ in (3.5) (or $p(Y | g, \xi)$ in (3.3) for the unit information prior). The posterior distribution of a functional in (2.6) can then be evaluated either by directly marginalizing ξ or by using MCMC for Monte Carlo integration of ξ , depending on the prior specified for ξ in Section 4.

3.2 Behavior of the Bayes factor

The choice of g is crucial for achieving appropriate sparsity in model selection with the g-prior (Kass and Raftery, 1995). A large value of g tends to favor sparse models, while a small value of g supports more complex models. This choice is particularly important in our additive model setup, as it directly influences the smoothness of the additive functions. In the literature on nonparametric regression with basis expansion, many studies use the unit information prior, which corresponds to setting $g = n$ (e.g. Gustafson, 2000; DiMatteo et al., 2001; Kohn et al., 2001). However, as noted earlier, a mixture of g-priors can offer improved empirical performance in BMS (Liang et al., 2008; Li and Clyde, 2018). While attempts have been made to assign a prior to g in nonparametric regression (Jeong and Park, 2016; Jeong et al., 2017; Francom et al., 2018; Francom and Sansó, 2020; Jeong et al., 2022), a thorough investigation into how these approaches differ from the unit information prior is still lacking. In this section, we explore how mixtures of g-priors compare to the unit information prior and discuss why the unit information prior might not be the optimal choice for estimating GAMs.

Our investigation utilizes Bayes factors. For two sets of knots $\xi_{(1)}$ and $\xi_{(2)}$, the Bayes factor of $\xi_{(1)}$ to $\xi_{(2)}$ is defined as $BF[\xi_{(1)}; \xi_{(2)}] = p(Y | \xi_{(1)})/p(Y | \xi_{(2)})$. For exponential family models with a known ϕ , the marginal likelihood $p(Y | \xi)$ is given by (3.3) with $g = n$ for the unit information prior and by (3.5) for mixtures of g-priors induced by tCCH priors on $(g + 1)^{-1}$. To understand how the Bayes factor penalizes model complexity, we consider two knots $\xi_{(1)}$ and $\xi_{(2)}$ such that $J_{\xi_{(1)}} = J_{\xi_{(2)}} + 1$ and $\hat{\eta}_{\xi_{(1)}} = \hat{\eta}_{\xi_{(2)}}$. In other words, both knots contribute equally to the model fit, but $\xi_{(1)}$ has one additional redundant knot-point compared to $\xi_{(2)}$. Accordingly, the Bayes factor satisfies $BF[\xi_{(1)}; \xi_{(2)}] < 1$, indicating that the larger model $\xi_{(1)}$ is never preferable over the smaller model $\xi_{(2)}$ owing to the same model fit. The Bayes factor $BF[\xi_{(1)}; \xi_{(2)}]$ quantifies the relative preference for the larger model $\xi_{(1)}$ over the smaller model $\xi_{(2)}$. For example, if $BF[\xi_{(1)}; \xi_{(2)}] = 1/2$, the larger model $\xi_{(1)}$ is only half as preferred as the smaller model $\xi_{(2)}$ (or equivalently, the smaller model $\xi_{(2)}$ is twice as preferred). As $BF[\xi_{(1)}; \xi_{(2)}]$ approaches 1, the preference for the two models becomes equal.

We examine how the Bayes factor behaves with changes in $J_{\xi_{(1)}}$ and the goodness-of-fit. In Gaussian regression, the goodness-of-fit is naturally assessed using the coefficient of determination. For the exponential family models, the pseudo- R^2 , defined as $1 - \exp(-D/n)$ with the usual deviance statistic D , can be used alternatively (Cox and Snell, 1989; Magee, 1990), with the caveat that its maximum value may be less than 1 depending on the specific model (Nagelkerke, 1991). To relate the Bayes factor to

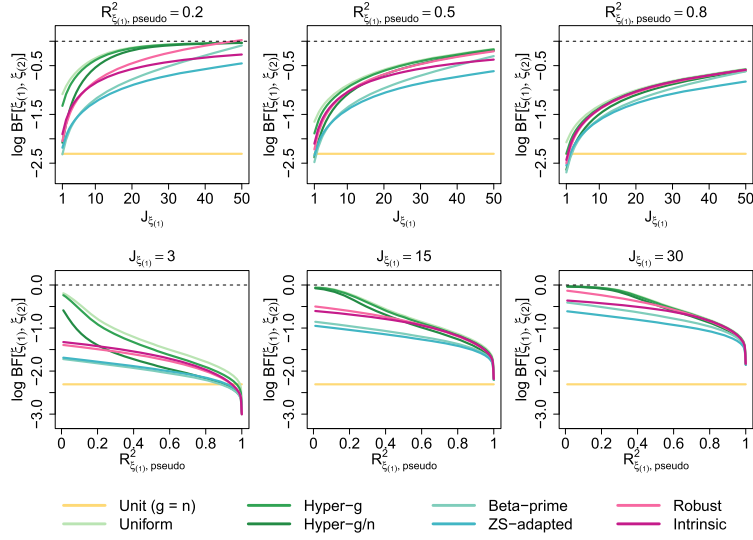


Figure 2: Change in $\log BF[\xi_{(1)}; \xi_{(2)}]$ as a function of $J_{\xi_{(1)}} (= J_{\xi_{(2)}} + 1)$ and $R_{\xi_{(1)}, \text{pseudo}}^2 (= R_{\xi_{(2)}, \text{pseudo}}^2)$ for $n = 1000$. The black dashed lines denote zero values, indicating equal preference for $\xi_{(1)}$ and $\xi_{(2)}$, i.e., $BF[\xi_{(1)}; \xi_{(2)}] = 1$.

the pseudo- R^2 , we use the fact that Q_ξ is asymptotically equivalent to the deviance D under mild conditions (Held et al., 2015; Li and Clyde, 2018). Therefore, we define $R_{\xi, \text{pseudo}}^2 = 1 - \exp(-Q_\xi/n)$ to gauge the goodness-of-fit for exponential family models.

Figure 2 presents a toy example with $n = 1000$, illustrating how $\log BF[\xi_{(1)}; \xi_{(2)}]$ changes with $J_{\xi_{(1)}}$ (where $J_{\xi_{(1)}} = J_{\xi_{(2)}} + 1$) and $R_{\xi_{(1)}, \text{pseudo}}^2$ (where $R_{\xi_{(1)}, \text{pseudo}}^2 = R_{\xi_{(2)}, \text{pseudo}}^2$). Similar patterns were observed with other values of n . The unit information prior consistently produces a constant Bayes factor, regardless of $J_{\xi_{(1)}}$ and $R_{\xi_{(1)}, \text{pseudo}}^2$. In contrast, the first row of Figure 2 shows that mixture priors cause $\log BF[\xi_{(1)}; \xi_{(2)}]$ to increase as the model size $J_{\xi_{(1)}}$ increases. This indicates that when comparing two small models (i.e., models with both small $J_{\xi_{(1)}}$ and $J_{\xi_{(2)}}$), mixtures of g-priors significantly penalize the larger model $\xi_{(1)}$ unless the marginal likelihood exhibits a notable improvement. Conversely, when comparing two large models (i.e., models with relatively large $J_{\xi_{(1)}}$ and $J_{\xi_{(2)}}$), the larger model $\xi_{(1)}$ is less likely to be penalized, even in the absence of a substantial benefit. This property of mixture priors can enhance GAM estimation, as it needs the comparison of large models generated through basis expansion to detect both local and global signals in the target functions that might be otherwise overlooked. The second row of Figure 2 shows that mixture priors cause $\log BF[\xi_{(1)}; \xi_{(2)}]$ to decrease as $R_{\xi_{(1)}, \text{pseudo}}^2$ increases. This aligns with intuition: with a sufficiently high goodness-of-fit, a more complex model may not be necessary, and a simpler model is often preferable unless it provides a significant improvement in the marginal likelihood. The unit information prior does not account for these characteristics in GAM estimation.

Figure 2 illustrates the differences between mixtures of g-priors. The beta-prime and ZS-adapted priors show similar behavior, with the smallest values of $\log BF[\xi_{(1)}; \xi_{(2)}]$, indicating the weakest inclination towards the larger model $\xi_{(1)}$. In contrast, the robust and intrinsic priors exhibit comparable decay patterns and result in larger values of $\log BF[\xi_{(1)}; \xi_{(2)}]$, reflecting a stronger relative preference for $\xi_{(1)}$ compared to the beta-prime and ZS-adapted priors. The two $O(1)$ -type priors (uniform and hyper-g) demonstrate a more pronounced preference for $\xi_{(1)}$ compared to the $O(n)$ -type priors. Interestingly, the hyper-g/n prior, although categorized as an $O(n)$ prior, behaves similarly to the $O(1)$ -type priors. Based on the discussion in the preceding paragraph, the latter three mixture priors may be mistakenly considered suitable for GAM estimation. However, when $R_{\xi_{(1)}, \text{pseudo}}^2$ is small, these priors tend to drive $\log BF[\xi_{(1)}; \xi_{(2)}]$ towards zero. This suggests that the preferences for the smaller and larger models may become undesirably similar, which may lead to overfitting. In contrast, other mixture priors appear to be less affected by this issue. The question of which mixture prior performs best for GAMs remains unresolved. Our numerical studies in Section 5 suggest that robust and intrinsic priors are the most effective. The following proposition provides a basic interpretation of where the differences among mixtures of g-priors may arise.

Proposition 3. *For the model in (2.1) and (2.2) with the priors in (3.1) and (3.2), consider two knots $\xi_{(1)}$ and $\xi_{(2)}$ such that $J_{\xi_{(1)}} = J_{\xi_{(2)}} + k$ and $\hat{\eta}_{\xi_{(1)}} = \hat{\eta}_{\xi_{(2)}}$, where k is a positive integer. For any positive integer k ,*

$$BF[\xi_{(1)}; \xi_{(2)}] = \begin{cases} (1+b)^{-k/2}, & \text{if } g = b, \\ E[(1+g)^{-k/2} \mid \xi_{(2)}, Y], & \text{if } g \text{ has a tCCH prior.} \end{cases}$$

The proof can be found in Section S2 of the supplementary material. This proposition implies that the Bayes factor $BF[\xi_{(1)}; \xi_{(2)}]$ represents the conditional posterior mean of $(1+g)^{-k/2}$, as induced by the unit information prior or tCCH priors. The proposition clarifies why the Bayes factor with the unit information prior remains constant. Differences in Bayes factors with mixture priors arise from variations in the posterior means of the shrinkage factor $(1+g)^{-k/2}$.

In conjunction with a specified prior for knots, $\Pi(\xi)$, the actual model comparison for determining the basis terms relies on the posterior odds $\Pi(\xi_{(1)} \mid Y)/\Pi(\xi_{(2)} \mid Y)$, rather than solely on the Bayes factor. Instead of employing mixtures of g-priors, one may consider using the unit information prior and adjusting the posterior odds with a suitable prior $\Pi(\xi)$. However, this approach is generally less favorable. This is because, for the posterior odds to reflect changes in goodness-of-fit with the unit information prior, $\Pi(\xi)$ would need to be excessively data-dependent. Thus, using a mixture of g-priors along with standard priors for knots is a more natural and practical choice.

4 Priors for knots

A prior $\Pi(\alpha, \beta_\xi \mid \xi)$ on the coefficients was specified in Section 3. To complete the Bayesian framework, we need to specify $\Pi(\xi)$ for the knots. Our prior for β_ξ requires

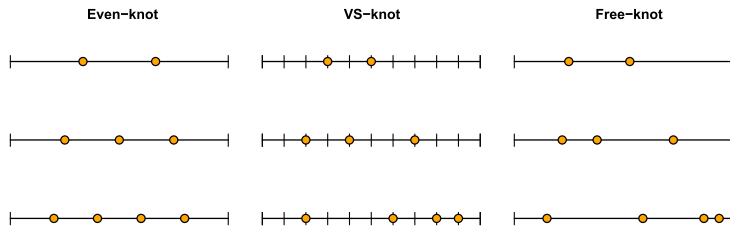


Figure 3: A graphical illustration of the three strategies for constructing Ξ discussed in Sections 4.1–4.3. For even-knot splines, the locations of knots are deterministically ascertained once $|\xi_j|$ is chosen. VS-knot splines select knot-points from a pre-determined set of locations. Free-knot splines are the most flexible and have no such limitation.

that B_ξ be of full-column rank (see (3.2) above). Therefore, we choose $\Pi(\xi)$ under the condition that B_ξ has full-column rank with a prior probability of one; that is, $\Pi(\text{rank}(B_\xi) = J_\xi) = 1$. This condition is typically satisfied by ensuring $J_\xi < n$, provided the knots and design points are well distributed.

Intuitively, ξ_j can be any set of singletons within the interval (ξ_j^L, ξ_j^U) , indicating that the intrinsic parameter space for ξ_j is infinite-dimensional. However, for computational reasons, a finite truncation to a restricted support may be beneficial. As previously mentioned, we denote Ξ as the induced support for $\Pi(\xi)$. The support Ξ restricts the function class generated by the natural cubic spline basis terms. A smaller space reduces model complexity but may fail to capture both local and global features of the target function. Thus, the choice of restricted support Ξ is crucial for balancing estimation quality and computational efficiency. Various strategies have been proposed for specifying Ξ for $\Pi(\xi)$. In this section, we discuss widely accepted methods for constructing Ξ , classifying them into three categories. These approaches are described in detail in Sections 4.1–4.3. Figure 3 provides a graphical summary of these strategies. A comparison of the empirical performance of these three approaches is presented in Section 5 based on a numerical study.

4.1 Even-knot splines: equidistant knots

The simplest yet powerful Bayesian adaptation arises from the assumption that the number of knots is not fixed but their locations are determined by an intrinsic law. The idea has been extensively considered in the literature and has been empirically and theoretically successful (e.g., Rivoirard and Rousseau, 2012; De Jonge and Van Zanten, 2012; Shen and Ghosal, 2015). We refer to this approach as *even-knot splines*. The name should be carefully understood because evenness may be assessed by an empirical measure rather than a geometric distance.

In this approach, a prior is assigned a number $|\xi_j|$ of knots ξ_j for $j = 1, \dots, p$. The specific locations of these knots are then determined based on a predefined rule. For example, for a given number $|\xi_j|$, the knots ξ_j may be equally spaced or chosen based

on the quantiles of the design points x_{ij} , $i = 1, \dots, n$. We prefer the latter approach for its stability. Using the quantiles also ensures that B_ξ has full-column rank, provided that $J_\xi < n$ and the design points x_{ij} , $i = 1, \dots, n$, are distinct. To avoid issues with duplicated quantile values for discretized design points, we recommend using only the unique quantile values. For computational efficiency, limiting each $|\xi_j|$ such that $|\xi_j| \leq M_j$ for a predetermined M_j , is computationally useful. The induced support is defined as

$$\Xi_{EK} = \left\{ \xi : \text{rank}(B_\xi) = J_\xi, |\xi_j| \leq M_j, \xi_{jk} = Q_{jk}, j = 1, \dots, p, k = 1, \dots, |\xi_j| \right\},$$

where Q_{jk} , $k = 1, \dots, |\xi_j|$, are the unique quantile values of x_{ij} , $i = 1, \dots, n$.³ Examples of knots in Ξ_{EK} are shown in Figure 3. With a density $q_j : \{0, 1, \dots, M_j\} \rightarrow (0, \infty)$ on $|\xi_j|$, the prior can be formally expressed as

$$\pi_{EK}(\xi) \propto \prod_{j=1}^p q_j(|\xi_j|), \quad \xi \in \Xi_{EK}. \quad (4.1)$$

Further discussion of the density q_j is presented in Section 4.4.

The key advantage of the prior in (4.1) is its low model complexity. Specifically, for a moderately large p , all possible models can be enumerated because $|\Xi_{EK}| \leq \prod_{j=1}^p (1 + M_j)$. This allows for MCMC-free posterior computation in relatively low-dimensional problems. If p is too large to enumerate all possibilities, the Metropolis-Hastings algorithm can be employed to explore the model space by proposing changes in $|\xi_j|$. In such cases, the computation can be streamlined by storing the value of the marginal likelihood $p(Y | \xi)$ for the current ξ and reusing it when the same ξ is revisited. We observe that this storage approach is effective unless p is extremely large.

Despite its advantages, the even-knot spline approach has a significant drawback due to its deterministic rules. Specifically, it cannot accommodate functions with spatially adaptive smoothness, such as Doppler functions. This limitation highlights the need for a more flexible construction, which is addressed in the following two subsections.

4.2 VS-knot splines: knot selection

The limitations of even-knot splines described in Section 4.1 can be mitigated by using a prior that induces a richer Ξ allowing for spatial adaptation. This can be achieved by allowing knot placement as well as the number of knots to be data-driven. A common approach is to set a large set of candidate basis functions and select the most important ones using Bayesian variable selection. This idea was introduced by Smith and Kohn (1996) and has been widely adopted in the literature on nonparametric regression (e.g., Kohn et al., 2001; Chan et al., 2006; Jeong and Park, 2016; Jeong et al., 2017; Park and Jeong, 2018; Jeong et al., 2022). We refer to this approach as *VS-knot splines*.

³These unique quantile values are obtained by removing duplicates from the usual quantiles of x_{ij} , $i = 1, \dots, n$, with equal probability. When ties are absent, they correspond to the usual quantiles.

Consider a set $\xi_j^c = \{\xi_{j1}^c, \dots, \xi_{jM_j}^c\}$ of knot candidates such that $\xi_j^L < \xi_{j1}^c < \dots < \xi_{jM_j}^c < \xi_j^U$ with large enough $M_j < n$. Similar to Section 4.1, the candidates ξ_j^c can be equidistant or determined using the unique values of the quantiles of x_{ij} , $i = 1, \dots, n$. We prefer the latter setup for its stability. The actual knots ξ_j are selected as a subset of ξ_j^c (including an empty set) using BMS. Consequently, the support consists of all possible subsets of $\{\xi_1^c, \dots, \xi_p^c\}$ with the restriction $\text{rank}(B_\xi) = J_\xi$, that is

$$\Xi_{VS} = \left\{ \xi : \text{rank}(B_\xi) = J_\xi, \xi_j \subset \xi_j^c, j = 1, \dots, p \right\}.$$

As in Section 4.1, we assign the density $q_j : \{0, 1, \dots, M_j\} \rightarrow (0, \infty)$ to $|\xi_j|$. We then assign equal weights to all knot locations conditional on $|\xi_j|$. The resulting prior is

$$\pi_{VS}(\xi) \propto \prod_{j=1}^p q_j(|\xi_j|) \binom{M_j}{|\xi_j|}^{-1}, \quad \xi \in \Xi_{VS}. \quad (4.2)$$

The VS-knot spline approach has proven effective in adapting to spatially inhomogeneous smoothness (e.g., Chan et al., 2006; Jeong and Park, 2016; Jeong et al., 2017). The cardinality $|\Xi_{VS}| \leq 2^{\sum_{j=1}^p M_j}$ indicates that enumerating all possible models is usually impractical, highlighting the usefulness of MCMC methods for exploring model spaces. Standard Gibbs sampling and Metropolis-Hastings algorithms are well-suited for this setup (Dellaportas et al., 2002). Sampling efficiency can be enhanced using block updates (Kohn et al., 2001; Jeong et al., 2022) or adaptive sampling (Nott and Kohn, 2005; Ji and Schmidler, 2013). Additionally, since Ξ_{VS} is finite-dimensional, storing the marginal likelihood, as discussed in Section 4.1, appears feasible. However, our experience shows that this approach is only effective when p is very small, due to memory constraints (e.g., $p \leq 2$). Consequently, we do not pursue this direction.

We emphasize that the basis system in (2.5) is particularly useful for the VS-knot spline approach. According to Proposition 2, knot selection naturally translates into basis selection. This property simplifies computation using the basis system in (2.4) and (2.5): one can generate a full basis matrix $B_j^c \in \mathbb{R}^{n \times (M_j+1)}$ whose (i, k) th component is $b_{jk}(x_{ij})$ constructed with the knot candidates $\xi_j^c = (\xi_{j1}^c, \dots, \xi_{jM_j}^c)$, and then choose important columns of B_j^c , while always including the first column for the linear term. As previously noted, this approach cannot be applied to other natural cubic spline basis functions, such as natural cubic B-splines or those in (5.4) and (5.5) of Hastie et al. (2009). For these basis functions, the basis term $b_{j,k+1}^*$ may be specified with more than one knot-point for some k . Consequently, inserting or deleting a knot-point may alter multiple basis terms, leading to a conflict between knot selection and basis selection.

4.3 Free-knot splines

The VS-knot spline strategy discussed in Section 4.2 selects important knot locations from a set of predetermined candidates. As a result, the knots are not equally spaced, which allows for spatially varying degrees of smoothness. Despite this flexibility, further relaxation of the restriction imposed by the discrete set of knot candidates remains a

topic of interest. This can be achieved with a fully nonparametric approach by allowing knots to be any singleton set within the specified range, provided that the induced B_ξ is of full-column rank. This approach is known as *free-knot splines* (Denison et al., 1998a; DiMatteo et al., 2001).

As in Section 4.1, capping each $|\xi_j|$ so that $|\xi_j| \leq M_j$ for a predetermined M_j can be computationally beneficial. The resulting support for ξ is

$$\Xi_{FK} = \left\{ \xi : \text{rank}(B_\xi) = J_\xi, |\xi_j| \leq M_j, \xi_j^L < \xi_{j1} < \cdots < \xi_{j|\xi_j|} < \xi_j^U, j = 1, \dots, p \right\}.$$

Clearly, the set Ξ_{FK} is uncountable. The prior is specified similarly to the one in (4.2). However, because the mapping $|\xi_j| \mapsto \xi_j$ is a surjection rather than a bijection, the conditional prior density of ξ_j given $|\xi_j|$, denoted by $\tilde{q}_j(\cdot | |\xi_j|)$, must be defined on the corresponding support. Following DiMatteo et al. (2001), \tilde{q}_j is chosen based on a uniform prior on the $|\xi_j|$ -simplex by scaling (ξ_j^L, ξ_j^U) to $(0, 1)$. With density $q_j : \{0, 1, \dots, M_j\} \rightarrow (0, \infty)$, the prior on ξ_j is formally expressed as:

$$\pi_{FK}(\xi) \propto \prod_{j=1}^p q_j(|\xi_j|) \tilde{q}_j(\xi_j | |\xi_j|), \quad \xi \in \Xi_{FK}. \quad (4.3)$$

While the original approach in DiMatteo et al. (2001) requires that at least one knot be always included, our free-knot spline prior in (4.3) extends it by allowing the possibility of an empty knot, which can account for a completely linear effect. To explore the posterior distribution, reversible jump MCMC with birth, death, and relocation proposals can be used (DiMatteo et al., 2001). This approach is generally more computationally demanding than methods for VS-knot splines. Despite the increased flexibility of the free-knot spline prior compared to the one in (4.2), our experience indicates that this flexibility does not significantly improve performance in most practical cases. The inherent inefficiency of reversible-jump MCMC further underscores the importance of avoiding unnecessary use of free-knot splines. Our simulation study in Section 5 shows that while performance measures for free-knot splines are comparable to those for VS-knot splines, the sampling efficiency (measured as the ratio of effective sample size to runtime) is notably lower for free-knot splines.

Similar to the VS-knot spline approach, the basis construction in (2.4) and (2.5) is useful for free-knot splines. According to Proposition 2, adding or removing a knot-point corresponds to adding or removing the corresponding basis term. Consequently, reversible-jump MCMC can be implemented by modifying the matrix columns without needing to reconstruct the entire basis term.

4.4 Prior distribution on $|\xi_j|$

The priors described in Sections 4.1–4.3 require specifying the density q_j for $|\xi_j|$. Previous studies have shown that to achieve optimal properties in nonparametric regression, priors used in BMS-based methods must have appropriately decaying tail properties (e.g., Shen and Ghosal, 2015). Priors with guaranteed tail properties include Poisson

and geometric distributions, with suitable truncation as needed. To leverage both the theoretical benefits and practical performance, we select our default prior as a mixture of a point mass at $|\xi_j| = 0$ and a truncated geometric distribution for $|\xi_j| > 0$. Specifically, the density is given by

$$q_j(u) = \begin{cases} \lambda_j, & u = 0, \\ (1 - \lambda_j)(1 - \varpi_j)^u / \sum_{\ell=1}^{M_j} (1 - \varpi_j)^\ell, & u = 1, \dots, M_j, \end{cases} \quad (4.4)$$

where the hyperparameter $\lambda_j \in [0, 1]$ represents the prior belief regarding a linear effect, while $\varpi_j \in [0, 1]$ governs the tail behavior. A reasonable default choice for λ_j is $1/2$. Selecting a value close to zero for ϖ_j makes the second part of the prior in (4.4) closely resemble a discrete uniform distribution while still ensuring the desired tail property for optimality. However, we find that using a moderately small value for ϖ_j improves the stability in estimating knot specifications. Consequently, we set $\varpi_j = 0.2$ as the default value.

The density q_j in (4.4) is also useful in GAPLMs, where some predictor variables are expected to have linear effects (e.g., binary variables). This is achieved by fixing specific f_j to include only the linear basis term N_1 . Accordingly, for predictor variables with linear effects, we set $\lambda_j = 1$ for the linear additive components and $\lambda_j = 1/2$ for the nonparametric additive components.

5 Numerical study

The primary goal of this study is to investigate the behavior of mixtures of g-priors in BMS-based approaches for estimating GAMs. While Section 3.2 provides some foundational insights, the optimal mixture prior for GAMs remains unclear. Additionally, we evaluate three strategies for specifying priors for knots, as discussed in Section 4, and compare them with other function estimation methods. This section introduces the simulation study designed to address these objectives.

5.1 Comparison among the mixtures of g-priors

We first conduct a simulation study to examine the differences in performance between mixtures of g-priors for estimating GAMs. For the synthetic functions, we consider the following four uncentered functions $f_j^* : [-1, 1] \rightarrow \mathbb{R}$, $j = 1, 2, 3, 4$:

$$\begin{aligned} f_1^*(x) &= 0.5(2x^5 + 3x^2 + \cos(3\pi x) - 1), \\ f_2^*(x) &= \frac{21(3x + 1.5)^3}{8000} + \frac{21(3x - 2.5)^2}{400e^{-3x-1.5}} \sin\left(\frac{17\pi(3x + 1.5)^2}{32}\right) \mathbb{1}_{(-0.5, 0.85)}(x), \\ f_3^*(x) &= x, \\ f_4^*(x) &= 0, \end{aligned} \quad (5.1)$$

where $\mathbb{1}_A$ is the indicator function of a set A . Specifically, f_1^* is a nonlinear function that is not a polynomial, f_2^* is a nonlinear function with locally varying smoothness,

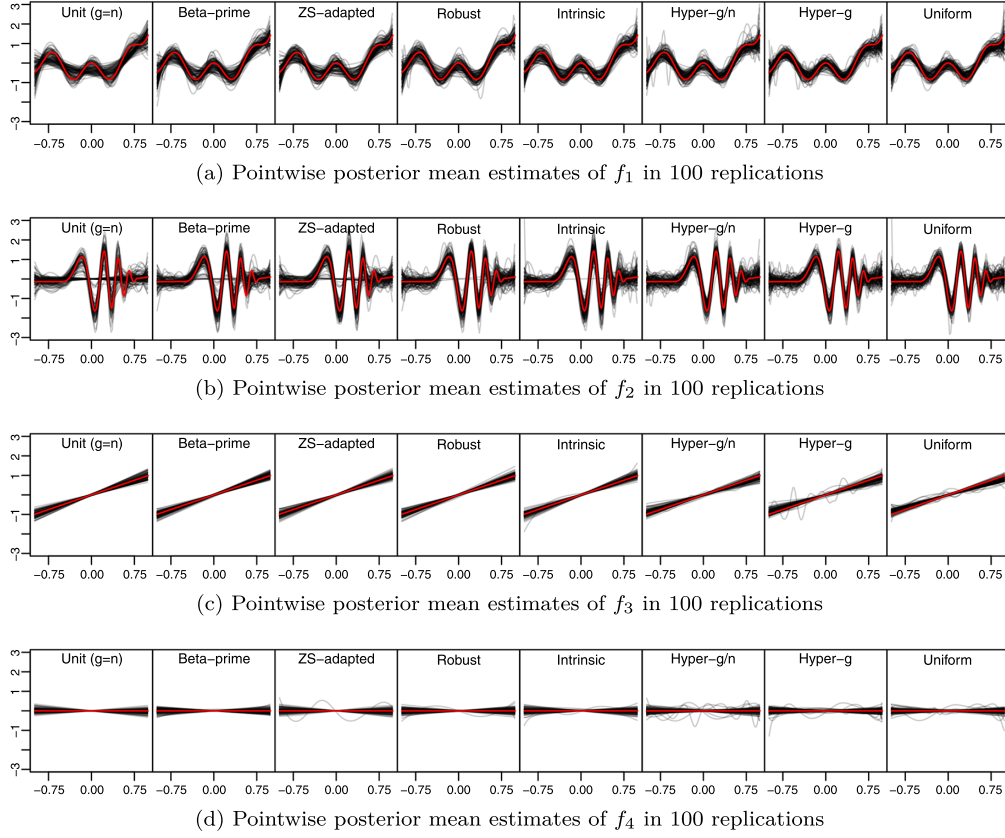


Figure 4: Pointwise posterior means (gray) of f_1 , f_2 , f_3 and f_4 in the nonparametric logistic regression model with $n = 1000$, obtained from randomly chosen 100 replicated datasets, along with the true function (red).

f_3^* is a linear function, and f_4^* is a constant function. The two nonlinear functions f_1^* and f_2^* are adapted from Gressani and Lambert (2021) and Francom and Sansó (2020), respectively. These functions are illustrated in Figure 4 with appropriate centering.

The simulation datasets are generated using the exponential family model with the additive predictor $\eta_i = \sum_{j=1}^4 f_j^*(x_{ij}) = \alpha + \sum_{j=1}^4 f_j(x_{ij})$, where x_{ij} are drawn independently from $\text{Unif}(-1, 1)$, f_j represents the centered version of f_j^* , and α is the induced intercept. In this section, we present the simulation results for the nonlinear logistic regression model, where $Y_i \sim \text{Bernoulli}(e^{\eta_i}/(1 + e^{\eta_i}))$. Section S6 of the supplementary material includes a simulation study for Poisson regression $Y_i \sim \text{Poi}(e^{\eta_i})$ and Gaussian regression $Y_i \sim N(\eta_i, \sigma^2)$.

As noted in Section 5.2, the VS-knot spline approach performs reasonably well compared to the other strategies for choosing Ξ described in Section 4. Therefore, we focus

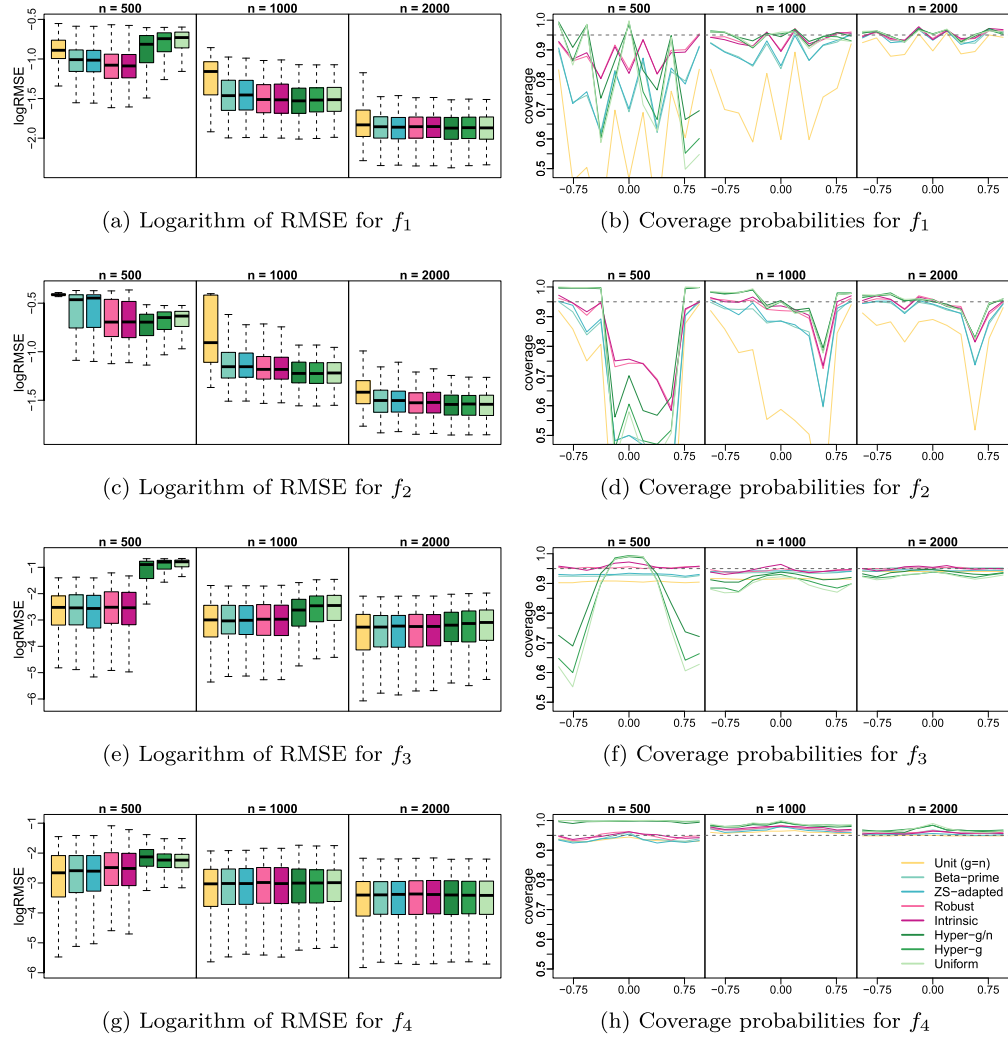


Figure 5: Logarithm of RMSE and coverage probabilities for f_1 , f_2 , f_3 and f_4 in the nonparametric logistic regression models with $n = 500, 1000, 2000$, obtained from 500 replicated datasets. Outliers are excluded to improve visualization.

specifically on VS-knot splines in this section. For each value of $n = 500, 1000, 2000$, we generate 500 data replications and estimate f_j using VS-knot splines with $M_j = 30$ knot candidates. We employ the unit information prior and the mixture priors summarized in Table 1, with the prior in (4.4) with $\varpi = 0.2$ and $\lambda = 1/2$. For each prior distribution, we run a Markov chain of length 10,000 to explore the posterior distribution, ensuring convergence after an appropriate burn-in period. We then calculate the root mean squared error (RMSE) and the coverage probabilities of 95% pointwise credible bands.

Figures 4 and 5 display the simulation results. As discussed in Section 3.2, the unit information prior behaves quite differently from the mixture priors. It generally underperforms in nonlinear function estimation and exhibits clear signs of underfitting. This indicates that using the unit information prior for function estimation may be unsuitable. The main challenge is determining the most appropriate mixture prior for function estimation. Although differences between mixtures of g-priors become less pronounced with larger sample sizes, intrinsic and robust priors consistently outperform other priors in finite samples. The beta-prime and ZS-adapted priors tend to exhibit slight underfitting, while the uniform, hyper-g, and hyper-g/n priors tend to exhibit overfitting. This aligns with the expectations discussed in Section 3.2. Across all functions, the robust and intrinsic priors achieve moderate RMSE and coverage properties, with the intrinsic prior being slightly more accurate for smaller samples. The simulation results for Poisson and Gaussian regressions in the supplementary material support this conclusion. We recommend using the intrinsic or robust prior as the default choice for g .

5.2 Comparison with other methods

We now compare BMS-based methods for GAMs with other approaches to GAM estimation. We consider the three strategies described in Section 4: even-knot splines, VS-knot splines, and free-knot splines, alongside several competitors available in R packages: `R2BayesX` (Umlauf et al., 2015), `Blapsr` (Gressani and Lambert, 2021), `mgcv` (Wood, 2017), and `bsamGP` (Jo et al., 2019). Based on the results in Section 5.1, the three BMS-based approaches use the intrinsic prior. Among the competitors, `mgcv` is the only frequentist method, while the others are Bayesian. Specifically, `R2BayesX` and `Blapsr` are based on Bayesian P-splines (Lang and Brezger, 2004). `R2BayesX` offers conventional MCMC estimates, whereas `Blapsr` provides an option for the Laplace approximation, which can improve computational efficiency when the number of additive components is small. In contrast, `bsamGP` uses a second-order Gaussian process to estimate nonparametric functions in GAMs.

To ensure a fair comparison, we carefully select simulation specifications. For the BMS-based methods (i.e., even-knot, VS-knot, and free-knot splines), the maximum number of knots M_j is consistently set to 30 for each $j = 1, 2, 3, 4$. Similarly, for the competitors relying on penalized splines (i.e., `R2BayesX`, `Blapsr`, and `mgcv`), we use $M_j = 30$, ensuring comparable least-penalized models across both BMS-based and penalized spline approaches. The `mgcv` package offers an option for locally adaptive smooth functions. We explore both the standard version with a single smoothness parameter (`mgcv-ps`) and a variant with local adaptation (`mgcv-ad`). For `bsamGP`, the number of cosine basis functions in the spectral representation of the Gaussian process priors is set equal to M_j for each f_j . The simulation settings follow those in Section 5.1, using the functions specified in (5.1). This section presents the simulation results for nonparametric logistic regression, with results for Poisson and Gaussian regression available in Section S6 of the supplementary material. For each Bayesian method relying on MCMC, we generate a Markov chain of length 10,000 to explore the posterior distribution, ensuring convergence after a suitable burn-in period. We then calculate the RMSE and 95% pointwise credible bands for selected points for each method.

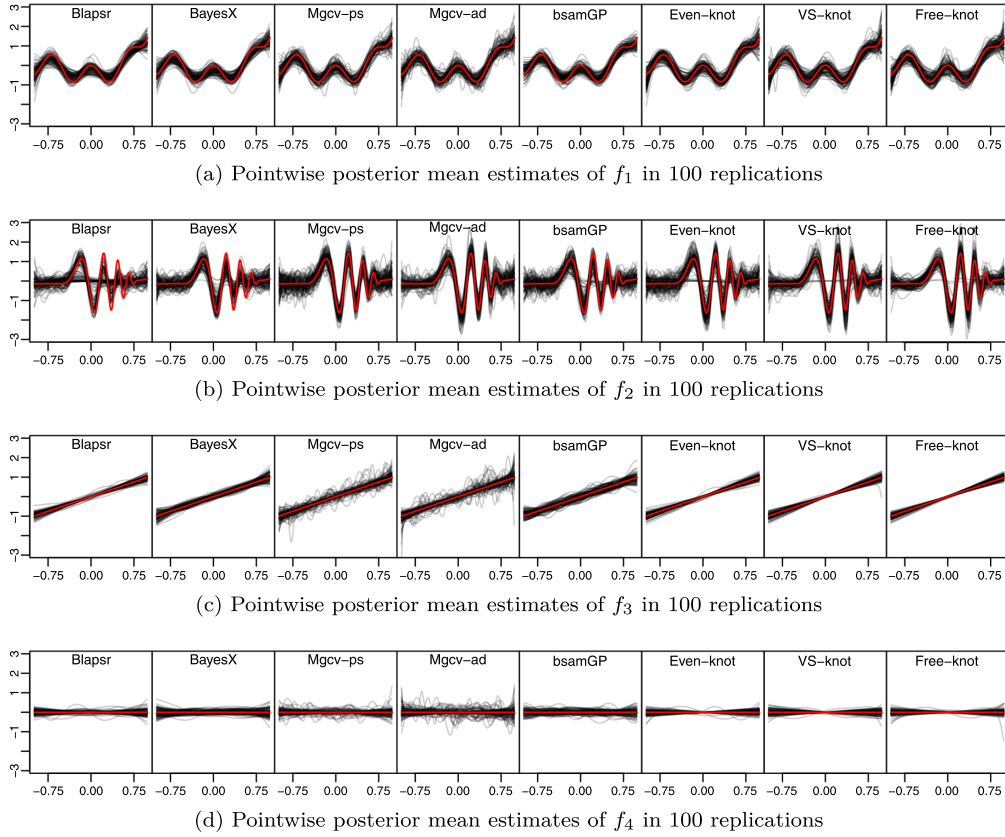


Figure 6: Pointwise posterior means (gray) of f_1 , f_2 , f_3 and f_4 in the nonparametric logistic regression model with $n = 1000$, obtained from randomly chosen 100 replicated datasets, along with the true function (red).

Figures 6 and 7 summarize the simulation results for the nonlinear logistic regression models. Observations reveal that `R2BayesX` and `Blapsr` tend to oversmooth the target functions owing to excessive penalization. In contrast, `mgcv` produces highly oscillatory estimates for the linear and constant functions, reflecting a tendency to overfit simpler functions. Both `R2BayesX` and `Blapsr` struggle with locally varying smoothness, as penalized splines are not inherently designed for such adaptability without significant modifications (Crainiceanu et al., 2007; Jullion and Lambert, 2007; Scheipl and Kneib, 2009). While `mgcv` with local adaptation performs well in estimating the locally varying smoothness of f_2 , the performance for f_1 , f_3 , and f_4 suggests that adaptive estimation using `mgcv` may lead to higher RMSEs and incorrect coverage probabilities. A major drawback of `mgcv` is the challenge of accurately specifying whether adaptive estimation will achieve optimal performance, given the unknown characteristics of the target function.

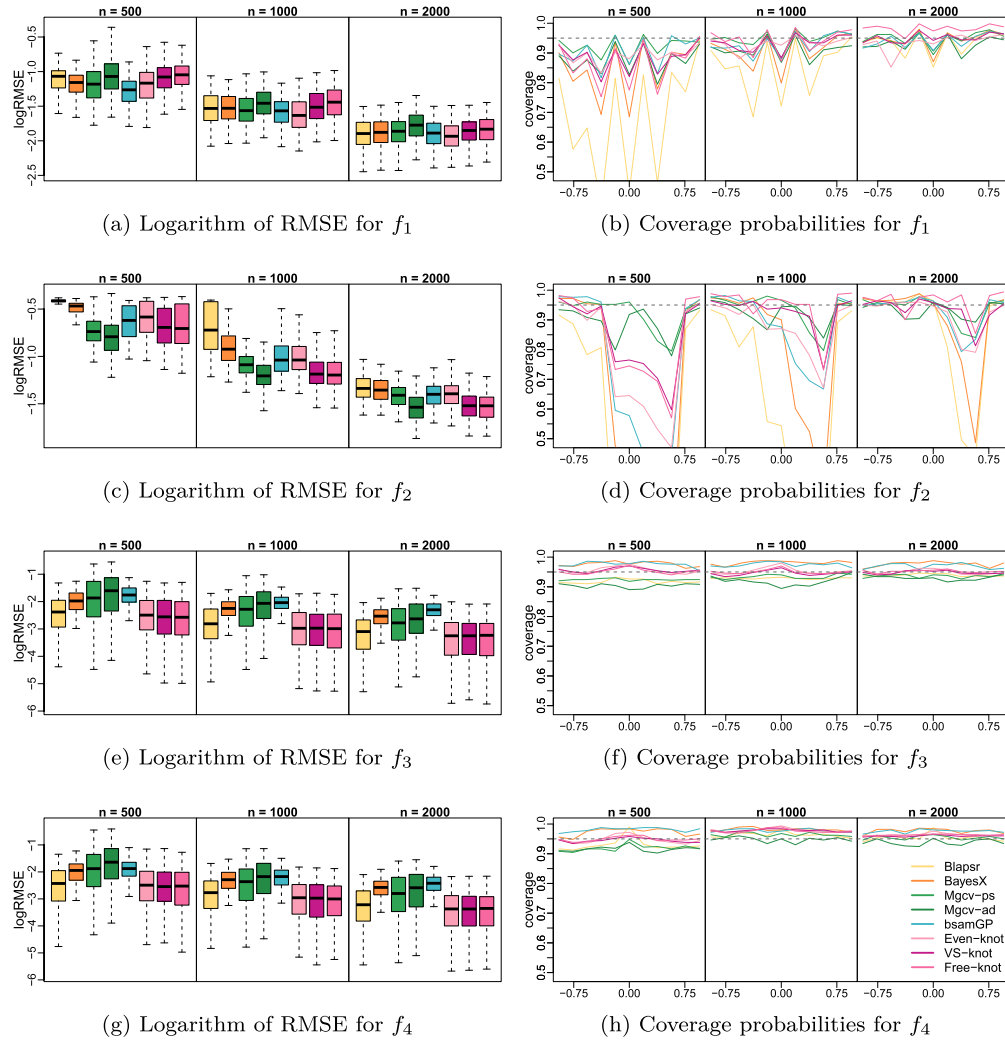


Figure 7: Logarithm of RMSE and coverage probabilities for f_1 , f_2 , f_3 and f_4 in the nonparametric logistic regression models with $n = 500, 1000, 2000$, obtained from 500 replicated datasets. Outliers are excluded to improve visualization.

Among the BMS-based approaches, even-knot splines exhibit limitations in adapting to the locally varying smoothness of f_2 owing to their construction with equidistant knots. In contrast, both the VS-knot and free-knot splines effectively identify the local features of f_2 . The results show that the RMSEs for these two adaptive estimation methods are comparable, although free-knot splines tend to slightly overestimate the coverage probabilities. Similar to the comparison between `mgcv-ps` and `mgcv-ad`, even-knot splines perform better than the VS-knot and free-knot splines in estimating f_1 .

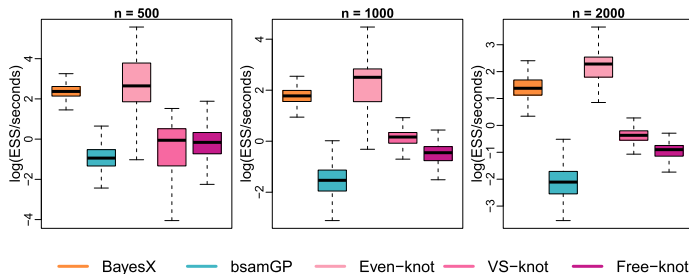


Figure 8: Logarithm of the effective sample sizes of the joint posterior per second of runtime, in the nonparametric logistic regression models with $n = 500, 1000, 2000$, obtained from 500 replicated datasets.

However, the BMS-based methods are comparable in estimating f_3 and f_4 , indicating that BMS-based methods are generally less sensitive to whether adaptive estimation is employed. Given that VS-knot splines typically outperform other methods and effectively handle local adaptation, we recommend using them as the default option. Nonetheless, even-knot splines are faster than other BMS-based methods and eliminate the need for MCMC when p is relatively small.

We also evaluated the computational efficiency of the Bayesian methods based on MCMC, excluding `mgcv` (which follows a frequentist approach) and `Blapsr` (which does not use MCMC). We measured the effective sample size of the posterior per second of runtime, after accounting for appropriate burn-in periods. Figure 8 displays the efficiency measures across 500 replicates. Contrary to the common belief that BMS-based methods might be inefficient, they perform comparably to other Bayesian methods. In particular, even-knot splines are the most efficient, owing to their fast mixing despite their slower overall processing. Although VS-knot splines are slower than even-knot splines, the trade-off is justified by their capability for local adaptation.

6 Application to Pima diabetes data

In this section, we analyze the Pima dataset using the VS-knot spline approach with the intrinsic prior. The Pima diabetes dataset consists of signs of diabetes and seven potential risk factors for $n = 532$ Pima Indian women in Arizona (Smith et al., 1988). We explore the relationship between the signs of diabetes and these risk factors using a GAM. The response variable Y_i indicates the presence of diabetes (0: negative, 1: positive). For each individual i , the predictor variables (risk factors) are $pregnant_i$ (number of times the subject was pregnant), $glucose_i$ (plasma glucose concentration in two hours in an oral glucose tolerance test, mg/dl), $pressure_i$ (diastolic blood pressure, mm/Hg), $triceps_i$ (triceps skin fold thickness, mm/Hg), $mass_i$ (body mass index, BMI), $pedigree_i$ (diabetes pedigree function), and age_i (age).

To examine the relationship between Y_i and the risk factors, we consider the following

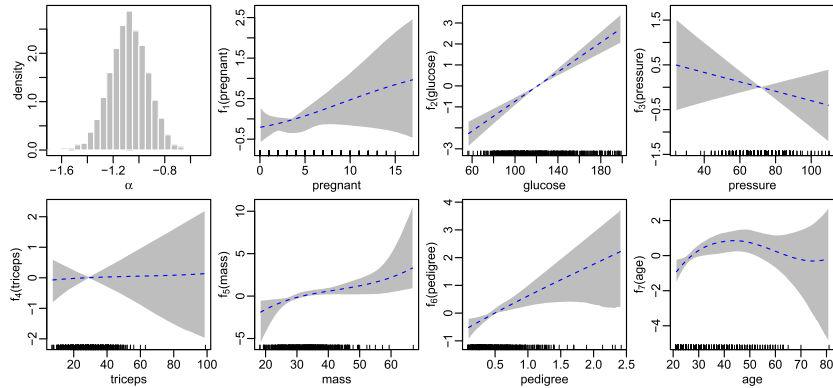


Figure 9: The posterior distribution of α and the pointwise posterior means (blue dashed curve) and pointwise 95% credible bands (gray shade) of the functions f_j , $j = 1, \dots, 7$, for the model in (6.1).

GAM with a logit link,

$$\log \frac{E(Y_i)}{1 - E(Y_i)} = \alpha + f_1(\text{pregnant}_i) + f_2(\text{glucose}_i) + f_3(\text{pressure}_i) + f_4(\text{triceps}_i) + f_5(\text{mass}_i) + f_6(\text{pedigree}_i) + f_7(\text{age}_i). \quad (6.1)$$

The individuals with missing values are removed from the analysis. For VS-knot splines, a set of knot candidates $\xi_j^c = \{\xi_{j1}^c, \dots, \xi_{jM_j}^c\}$ is determined from the unique values of the quantiles for each predictor variable. Some predictors are discrete in the observed data (e.g., pregnant_i and age_i). For each j , we set M_j as the number of unique values among the 30 quantile values with equal weights, meaning that $M_j < 30$ for some j .

The results summarized in Figure 9 largely align with intuition. Many variables show near-linear effects, while a few, such as mass_i and age_i , exhibit clear nonlinear effects. The effect of triceps_i appears to be negligible, suggesting that it may be worth considering the exclusion of this variable from the analysis.

7 Discussion

This study examined BMS-based estimation methods for GAMs using the Laplace approximation with mixtures of g-priors. We establish a default prior by analyzing the behavior of the Bayes factor and presenting numerical results. Additionally, this study consolidates existing ideas on priors for knots and mixtures of g-priors.

A significant limitation of the BMS-based approach is that the Laplace approximation depends on the maximum likelihood estimator, which requires extensive computations. Given that the VS-knot spline approach proves effective, one potential solution is to use shrinkage priors for exponential family models in combination with a reasonable

MCMC sampling algorithm (e.g., Schmidt and Makalic, 2020). Another avenue is to explore computationally less expensive likelihood approximations (e.g., Rossell et al., 2021). Additionally, investigating higher-order approximations could offer improved approximation performance (Shun and McCullagh, 1995).

Funding

This research was supported by the Yonsei University Research Fund of 2021-22-0032 and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (2022R1C1C1006735, RS-2023-00217705).

Supplementary Material

Supplementary Material of ‘Model Selection-Based Estimation for Generalized Additive Models Using Mixtures of g-priors: Towards Systematization’ (DOI: [10.1214/24-BA1466SUPP](https://doi.org/10.1214/24-BA1466SUPP); .pdf). It includes proofs of propositions, additional simulation studies, and instructions for installing the R package.

References

- Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). “Criteria for Bayesian model choice with application to variable selection.” *The Annals of Statistics*, 40(3): 1550–1577. [MR4586699](https://doi.org/10.1016/j.csda.2023.107766). doi: <https://doi.org/10.1016/j.csda.2023.107766>. 2, 7, 9
- Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998). “Bayes factors and marginal distributions in invariant situations.” *Sankhyā: The Indian Journal of Statistics, Series A*, 307–321. 7
- Brezger, A. and Lang, S. (2006). “Generalized structured additive regression based on Bayesian P-splines.” *Computational Statistics & Data Analysis*, 50(4): 967–991. [MR2893854](https://doi.org/10.1214/11-AOS885). doi: <https://doi.org/10.1214/11-AOS885>. 2
- Castellanos, M. E., García-Donato, G., and Cabras, S. (2021). “A model selection approach for variable selection with censored data.” *Bayesian Analysis*, 16(1): 271 – 300. [MR2408684](https://doi.org/10.1214/21-BA1466). 8
- Chan, D., Kohn, R., Nott, D., and Kirby, C. (2006). “Locally adaptive semiparametric estimation of the mean and variance functions in regression models.” *Journal of Computational and Graphical Statistics*, 15(4): 915–936. 15, 16
- Chen, M.-H. and Ibrahim, J. G. (2003). “Conjugate priors for generalized linear models.” *Statistica Sinica*, 461–476. [MR3265679](https://doi.org/10.1080/01621459.2014.880348). doi: <https://doi.org/10.1080/01621459.2014.880348>. 2
- Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). “BART: Bayesian additive regression trees.” *The Annals of Applied Statistics*, 4(1): 266–298. [MR3726911](https://doi.org/10.1214/10-AOS503). 2

- Cox, D. and Snell, E. (1989). *The Analysis of Binary Data*, volume 32. CRC Press. [MR0881437](#). 11
- Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Joshi, A., and Goodner, B. (2007). “Spatially adaptive Bayesian penalized splines with heteroscedastic errors.” *Journal of Computational and Graphical Statistics*, 16(2): 265–288. 22
- De Jonge, R. and Van Zanten, J. (2012). “Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors.” *Electronic Journal of Statistics*, 6: 1984–2001. 2, 6, 14
- Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). “On Bayesian model and variable selection using MCMC.” *Statistics and Computing*, 12(1): 27–36. 16
- Denison, D., Mallick, B., and Smith, A. (1998a). “Automatic Bayesian curve fitting.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2): 333–350. 2, 3, 6, 17
- Denison, D. G., Mallick, B. K., and Smith, A. F. (1998b). “Bayesian MARS.” *Statistics and Computing*, 8(4): 337–346. 2
- DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). “Bayesian curve-fitting with free-knot splines.” *Biometrika*, 88(4): 1055–1071. 2, 3, 6, 8, 11, 17
- Fahrmeir, L. and Lang, S. (2001). “Bayesian inference for generalized additive mixed models based on Markov random field priors.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2): 201–220. 2
- Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2018). “Power-expected-posterior priors for generalized linear models.” *Bayesian Analysis*, 13(3): 721–748. 3
- Francom, D. and Sansó, B. (2020). “BASS: An R package for fitting and performing sensitivity analysis of Bayesian adaptive spline surfaces.” *Journal of Statistical Software*, 94(8): 1–36. 11, 19
- Francom, D., Sansó, B., Kupresanin, A., and Johannesson, G. (2018). “Sensitivity analysis and emulation for functional data using Bayesian adaptive splines.” *Statistica Sinica*, 791–816. 11
- García-Donato, G., Cabras, S., and Castellanos, M. E. (2023). “Model uncertainty quantification in Cox regression.” *Biometrics*, 79(3): 1726–1736. 8
- Gordy, M. B. (1998). “A generalization of generalized beta distributions.” Division of Research and Statistics, Division of Monetary Affairs, Federal Reserve Boards. 9, 10
- Green, P. J. (1995). “Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination.” *Biometrika*, 82(4): 711–732. 2
- Gressani, O. and Lambert, P. (2021). “Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines.” *Computational Statistics & Data Analysis*, 154: 107088. 2, 19, 21

- Gupta, M. and Ibrahim, J. G. (2009). “An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data.” *Statistica Sinica*, 19(4): 1641–1663. 8
- Gustafson, P. (2000). “Bayesian regression modeling with interactions and smooth effects.” *Journal of the American Statistical Association*, 95(451): 795–806. 8, 11
- Hansen, M. H. and Yu, B. (2003). “Minimum description length model selection criteria for generalized linear models.” *Lecture Notes-Monograph Series*, 145–163. 8
- Hastie, T. and Tibshirani, R. (1986). “Generalized additive models.” *Statistical Science*, 297–318. 1
- Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition. 5, 16
- Held, L., Sabanés Bové, D., and Gravestock, I. (2015). “Approximate Bayesian model selection with the deviance statistic.” *Statistical Science*, 242–257. 3, 8, 9, 12
- Jeong, S., Park, M., and Park, T. (2017). “Analysis of binary longitudinal data with time-varying effects.” *Computational Statistics & Data Analysis*, 112: 145–153. 2, 6, 11, 15, 16
- Jeong, S. and Park, T. (2016). “Bayesian semiparametric inference on functional relationships in linear mixed models.” *Bayesian Analysis*, 11(4): 1137–1163. 6, 11, 15, 16
- Jeong, S., Park, T., and van Dyk, D. A. (2022). “Bayesian model selection in additive partial linear models via locally adaptive splines.” *Journal of Computational and Graphical Statistics*, 31(2): 324–336. 6, 11, 15, 16
- Jeong, S. and Rockova, V. (2023). “The art of BART: Minimax optimality over nonhomogeneous smoothness in high dimension.” *Journal of Machine Learning Research*, 24(337): 1–65. 2
- Ji, C. and Schmidler, S. C. (2013). “Adaptive Markov Chain Monte Carlo for Bayesian variable selection.” *Journal of Computational and Graphical Statistics*, 22(3): 708–728. 16
- Jo, S., Choi, T., Park, B., and Lenk, P. (2019). “bsamGP: An R Package for Bayesian Spectral Analysis Models Using Gaussian Process Priors.” *Journal of Statistical Software*, 90(10): 1–41. 21
- Jullion, A. and Lambert, P. (2007). “Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models.” *Computational Statistics & Data Analysis*, 51(5): 2542–2558. 22
- Kang, G. and Jeong, S. (2024). “Supplementary Material for “Model Selection-Based Estimation for Generalized Additive Models Using Mixtures of g-priors: Towards Systematization”.” *Bayesian Analysis*. doi: <https://doi.org/10.1214/24-BA1466SUPP>. 4

- Kass, R. E. and Raftery, A. E. (1995). “Bayes factors.” *Journal of the American Statistical Association*, 90(430): 773–795. [11](#)
- Kass, R. E. and Wasserman, L. (1995). “A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion.” *Journal of the American Statistical Association*, 90(431): 928–934. [3](#), [8](#)
- Kohn, R., Smith, M., and Chan, D. (2001). “Nonparametric regression using linear combinations of basis functions.” *Statistics and Computing*, 11(4): 313–322. [8](#), [11](#), [15](#), [16](#)
- Lang, S. and Brezger, A. (2004). “Bayesian P-splines.” *Journal of Computational and Graphical Statistics*, 13(1): 183–212. [2](#), [21](#)
- Li, Y. and Clyde, M. A. (2018). “Mixtures of g-priors in generalized linear models.” *Journal of the American Statistical Association*, 113(524): 1828–1845. [2](#), [3](#), [8](#), [9](#), [11](#), [12](#)
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). “Mixtures of g-priors for Bayesian variable selection.” *Journal of the American Statistical Association*, 103(481): 410–423. [2](#), [9](#), [11](#)
- Magee, L. (1990). “ R^2 measures based on Wald and likelihood ratio joint significance tests.” *The American Statistician*, 44(3): 250–253. [11](#)
- Maruyama, Y. and George, E. I. (2011). “Fully Bayes factors with a generalized g-prior.” *The Annals of Statistics*, 39(5): 2740–2765. [2](#), [9](#)
- Nagelkerke, N. J. (1991). “A note on a general definition of the coefficient of determination.” *Biometrika*, 78(3): 691–692. [11](#)
- Nott, D. J. and Kohn, R. (2005). “Adaptive sampling for Bayesian variable selection.” *Biometrika*, 92(4): 747–763. [16](#)
- Park, T. and Jeong, S. (2018). “Analysis of Poisson varying-coefficient models with autoregression.” *Statistics*, 52(1): 34–49. [15](#)
- Rivoirard, V. and Rousseau, J. (2012). “Posterior concentration rates for infinite dimensional exponential families.” *Bayesian Analysis*, 7(2): 311–334. [2](#), [6](#), [14](#)
- Rossell, D., Abril, O., and Bhattacharya, A. (2021). “Approximate Laplace approximations for scalable model selection.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4): 853–879. [26](#)
- Sabanés Bové, D. and Held, L. (2011). “Hyper-g priors for generalized linear models.” *Bayesian Analysis*, 6(3): 387–410. [3](#), [8](#)
- Sabanés Bové, D., Held, L., and Kauermann, G. (2015). “Objective Bayesian model selection in generalized additive models with penalized splines.” *Journal of Computational and Graphical Statistics*, 24(2): 394–415. [2](#)
- Scheipl, F. and Kneib, T. (2009). “Locally adaptive Bayesian P-splines with a Normal-Exponential-Gamma prior.” *Computational Statistics & Data Analysis*, 53(10): 3533–3552. [22](#)

- Schmidt, D. F. and Makalic, E. (2020). “Bayesian generalized horseshoe estimation of generalized linear models.” In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 598–613. Springer. 26
- Shen, W. and Ghosal, S. (2015). “Adaptive Bayesian procedures using random series priors.” *Scandinavian Journal of Statistics*, 42(4): 1194–1213. 2, 3, 6, 14, 17
- Shun, Z. and McCullagh, P. (1995). “Laplace approximation of high dimensional integrals.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(4): 749–760. 26
- Smith, J. W., Everhart, J. E., Dickson, W., Knowler, W. C., and Johannes, R. S. (1988). “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus.” In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261. American Medical Informatics Association. 24
- Smith, M. and Kohn, R. (1996). “Nonparametric regression using Bayesian variable selection.” *Journal of Econometrics*, 75(2): 317–343. 2, 3, 6, 15
- Sohn, J., Jeong, S., Cho, Y. M., and Park, T. (2023). “Functional clustering methods for binary longitudinal data with temporal heterogeneity.” *Computational Statistics & Data Analysis*, 185: 107766. 2
- Umlauf, N., Adler, D., Kneib, T., Lang, S., and Zeileis, A. (2015). “Structured Additive Regression Models: An R Interface to BayesX.” *Journal of Statistical Software*, 63(21): 1–46. 21
- Wang, L., Liu, X., Liang, H., and Carroll, R. J. (2011). “Estimation and variable selection for generalized additive partial linear models.” *Annals of Statistics*, 39(4): 1827. 6
- Wang, X. and George, E. I. (2007). “Adaptive Bayesian criteria in variable selection for generalized linear models.” *Statistica Sinica*, 667–690. 8
- Williams, C. and Rasmussen, C. (1995). “Gaussian processes for regression.” *Advances in Neural Information Processing Systems*, 8. 2
- Womack, A. J., León-Novelo, L., and Casella, G. (2014). “Inference from intrinsic Bayes’ procedures under model selection and uncertainty.” *Journal of the American Statistical Association*, 109(507): 1040–1053. 2, 9
- Wood, S. N. (2017). *Generalized Additive Models: an Introduction with R*. CRC press. 21
- Zellner, A. (1986). “On assessing prior distributions and Bayesian regression analysis with g-prior distributions.” *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243. 7
- Zellner, A. and Siow, A. (1980). “Posterior odds ratios for selected regression hypotheses.” *Trabajos de Estadística Y de Investigación Operativa*, 31(1): 585–603. 2