Model selection-based estimation for generalized additive models with mixtures of g-priors: Towards a systematization

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

[1]Department of Statistics and Data Science, Yonsei University, Seoul, Korea
[2]Department of Applied Statistics, Yonsei University, Seoul, Korea

December 29, 2022

Abstract

We consider estimation of generalized additive models using basis expansions with Bayesian model selection. Although Bayesian model selection is an intuitively appealing tool for regression splines due to the flexible knot placement and model-averaged function estimates, its use has traditionally been limited to Gaussian additive regression as posterior search of the model space requires a tractable form of the marginal model likelihood. We introduce an extension of the method to distributions belonging to the exponential family through the Laplace approximation to the likelihood. Furthermore, there is currently no broad agreement on the best prior distribution to use for nonparametric regression via model selection. We observe that the classical unit information prior for variable selection may not be suitable for nonparametric regression through basis expansions. Instead, our study reveals that mixtures of g-priors are the right choice. A large family of mixtures of g-priors is considered for a detailed examination of how various mixture priors perform in estimation of generalized additive models. Simulation studies and real data applications demonstrate the validity of the model selection-based approach. We provide an R package for the proposed method.

Keywords: Adaptive regression; exponential family models; mixtures of g-priors; nonparametric regression; splines.

Contents

*Corresponding author: sjeong@yonsei.ac.kr

1 Introduction

Since its inception, the generalized additive model (GAM) has played a significant role in statistics and machine learning, and has received a great deal of attention from many theorists and practitioners (e.g., Yee and Mitchell, 1991; Yee and Wild, 1996; Guisan et al., 2002; McLean et al., 2014; Wood et al., 2015). The GAM is seen as an interpretable semiparametric compromise between the parametric generalized linear model (GLM) and fully nonparametric regression with multidimensional smoothing. To be more explicit, the GAM explains the relationship between multiple predictor variables and a (possibly non-Gaussian) response variable by employing an additive structure of univariate functions (Hastie and Tibshirani, 1986). In other words, at the expense of the flexibility of multidimensional smoothing, the GAM provides a straightforward interpretation of the amount each predictor variable contributes to the mean response as a univariate function.

A variety of estimation procedures have been proposed for nonparametric regression and additive models from both the frequentist and Bayesian perspectives. Focusing on the Bayesian philosophy, typical techniques for univariate smooth function estimation 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., 1998; DiMatteo et al., 2001). As a branch of Bayesian estimation methods, basis expansion with Bayesian model selection (BMS) enjoys appealing theoretical properties and successful empirical performances (Smith and Kohn, 1996; Denison et al., 1998; DiMatteo et al., 2001; Rivoirard and Rousseau, 2012; De Jonge and Van Zanten, 2012; Shen and Ghosal, 2015). Specifically, the BMS-based approaches determine intrinsic basis terms by comparing the Bayes factors based on BMS, which translates into choosing more plausible basis terms among a possible set of candidates in a data-driven manner. Therefore, the BMS-based methods can be computationally successful only when the calculation of the marginal likelihood, obtained by marginalizing the coefficients out, is readily accessible, which has led to the use of BMS for nonparametric regression typically being limited to Gaussian nonparametric regression. For GLMs and GAMs, marginalization is often unachievable

unless a conjugate prior is used for the coefficients, but such a conjugate prior does not provide a convenient form of the marginal likelihood (Chen and Ibrahim, 2003). The sole accessible application may be nonparametric probit regression (e.g., Jeong et al., 2017; Sohn et al., 2022), which has a convenient expression of Gaussian regression using latent variables (Albert and Chib, 1993). If marginalization is not analytically tractable, the BMS-based methods require numerical marginalization of the coefficients through Markov chain Monte Carlo (MCMC) algorithms such as the reversible jump MCMC (Green, 1995), which is often far more inefficient than using the Bayes factors unless a sound proposal distribution is available (Al-Awadhi et al., 2004). This difficulty, in part, leads to the prevalence of the P-spline-based Bayesian methods for estimation in GAMs during the early stages of their development (e.g., Fahrmeir and Lang, 2001; Brezger and Lang, 2006), as marginalization is not required for the Bayesian P-splines.

One natural solution to this issue is considering an approximation to the likelihood, such as the Laplace approximation, so that the marginal likelihood can be obtained with a Gaussian prior distribution on the coefficients (Li and Clyde, 2018). This idea allows us to employ the BMS-based approaches for estimation in GAMs with distributions belonging to the exponential family of distributions, and has occasionally been used for the BMS-based approaches (e.g., DiMatteo et al., 2001). The approximation idea has also been widely accepted in the literature of GAM estimation with the Bayesian P-splines to facilitate the computation (Sabanés Bové et al., 2015; Wood, 2017; Gressani and Lambert, 2021). Using the Laplace approximation for the BMS-based methods is conceptually simple. However, to the best of our knowledge, it has not been methodically systematized under the general framework of non-Gaussian GAMs. Furthermore, although the Laplace approximation is straightforward as soon as a Gaussian or Gaussian mixture prior is used for the coefficients, it still remains unclear which prior distribution performs the best for our purpose of basis determination. In the literature of variable selection, it has been continually reported that mixture priors outperform the classical Gaussian prior called 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 shown to have many desirable properties and resolve the paradoxes of the g-prior (Liang et al., 2008). Various mixtures of g-priors have been proposed under the framework of linear regression (e.g., Zellner and Siow, 1980; Cui and George, 2008; Liang et al., 2008; Maruyama and George, 2011; Bayarri et al., 2012; Ley and Steel, 2012; Womack et al., 2014), but there have been only a few attempts to extend those to the GLM (Sabanés Bové and Held, 2011; Held et al., 2015; Fouskakis et al., 2018). Recently, Li and Clyde (2018) unifies and extends mixtures of g-priors for the GLM using the truncated compound confluent hypergeometric distributions (Gordy, 1998), which encompasses the currently existing mixture prior distributions. However, the mixture prior that performs the best for the BMS-based GAM estimation is still beyond the veil. To draw back the curtain, one must understand how mixtures of g-priors behave in order to penalize the nonparametric functions. The best mixture prior for nonparametric regression is unclear even for Gaussian additive regression.

In this paper, we methodize the use of the Laplace approximation for GAM estimation via BMS with mixtures of g-priors. For the construction of mixtures of g-priors, we follow the general recipe systematized by Li and Clyde (2018). To determine a default mixture prior, an understanding of mixtures of g-priors is provided as a penalty to the model in order to see how mixture priors behave for estimating GAMs. The empirical performance of mixture priors is also

investigated through an extensive simulation study. Our experience indicates that the traditional g-prior utilizing the sample size directly, also known as the unit information prior (Kass and Wasserman, 1995), may not be suitable, and a mixture of g-priors should be used instead. We support our claim with a verity of numerical results and real data analysis.

The remainder of this article is structured as follows. Section 2 introduces our formulation of GAM and mixtures of g-priors for Bayesian model selection. In Section 3, we describe three different strategies of choosing prior distributions for the BMS-based approaches for estimation in GAMs. Section 4.2 describes the case of Gaussian additive regression with unknown dispersion as a special case of GAMs. Section 4.3 suggests a framework to interpret the behaviors of mixtures of g-priors as penalty functions for GAMs. We also compare the proposed method against other competitors for GAM estimation rooted on the Bayesian philosophy in Section 5. Two real datasets are analyzed in Section 6. Lastly, Section 7 concludes with a discussion.

2 Generalized additive models via basis expansion

For given predictor variables $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \in \mathbb{R}^p$, suppose that a response variable $Y_i \in \mathbb{R}$ has a distribution belonging to the exponential family of distributions, i.e., the density of Y_i is

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

where θ_i is a natural parameter modeled by x_i, ϕ is a scale parameter, and b, c are known functions. The dependency of θ_i on x_i will be clarified below. The dispersion parameter ϕ is typically assumed to be known, but can be treated as an unknown parameter as in the case of Gaussian regression. Assuming that b is twice differentiable and $b''(\theta_i) > 0$, the expected value and the variance of Y_i is given by $E(Y_i) = b'(\theta_i)$ and $Var(Y_i) = \phi b''(\theta_i)$, respectively. We choose a monotonically increasing link function h that parameterizes the natural parameter as $\theta_i = (h \circ b')^{-1}(\eta_i)$, where η_i is an additive predictor expressed as

$$\eta_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}), \quad i = 1, \ldots, n, \tag{2}$$

with a global mean parameter α and univariate functions $f_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, p$. To ensure the identifiability, the functions f_j are assumed to satisfy $\sum_{i=1}^{n} f_j(x_{ij}) = 0$, $j = 1, \ldots, p$.

For the complete model specification, the most important part is determining how to characterize the nonparametric functions f_j. Throughout this paper, the functions f_j are parameterized by spline basis representation; that is, f_j is expressed as a linear combination of K_j basis functions $b_{j1}, \ldots, b_{jK_j}$, i.e., with coefficients $\beta_{jk} \in \mathbb{R}$,

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

For the identifiability condition $\sum_{i=1}^{n} f_j(x_{ij}) = 0$ to be satisfied, we assume that each basis function satisfies $\sum_{i=1}^{n} b_{jk}(x_{ij}) = 0$, $j = 1, \ldots, p$. This can be easily achieved by centering an

unrestricted basis term b^*_{jk} as

$$b_{jk}(\cdot) = b^*_{jk}(\cdot) - \frac{1}{n}\sum_{i=1}^{n} b^*_{jk}(x_{ij}), \quad j = 1, \ldots, p, \quad k = 1, \ldots, K_j. \tag{3}$$

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

There are many classes of basis functions that can be used for smooth function estimation. Typical examples include B-splines (De Boor, 1978), wavelets (Antoniadis, 1997), radial basis functions (Buhmann, 2003), and Fourier basis functions (Katznelson, 2004). In this paper, we deploy natural cubic spline basis functions to prevent the erratic behavior at the boundaries. This is equivalent to using any type of piecewise polynomial basis functions (including B-splines) with suitable natural boundary conditions provided that a prior distribution is invariant to linear transformations of a design matrix (more precisely, invariant to isomorphisms). Specifically, for boundary knots $\{t^L, t^U\}$ and a set of M interior knots $\{t_1, \ldots, t_M\}$ satisfying $-\infty < t^L < t_1 < \cdots < t_M < t^U < \infty$, we define natural cubic spline basis functions $N_k : \mathbb{R} \to \mathbb{R}$, $k = 1, \ldots, M+1$, as

$$\begin{aligned} N_1(u) &= u, \\ N_{k+1}(u) &= N(u; t^L, t^U, t_k) \\ &\equiv \frac{(u - t_k)^3_+ - (u - t^U)^3_+}{t^U - t_k} - \frac{(u - t^L)^3_+ - (u - t^U)^3_+}{t^U - t^L}, \quad k = 1, \ldots, M. \end{aligned} \tag{4}$$

Together with the constant term $N_0(u) = 1$, the basis functions in (4) generate piecewise cubic functions with the restriction that the spline function is linear beyond the boundary knots $\{t^L, t^U\}$, resulting in increased stability near the boundaries due to the induced constraints at $\{t^L, t^U\}$. The constant term is excluded in (4) because it is redundant due to the intercept term. In comparison to cubic splines without the natural conditions, our experience shows (as is well known) that the use of natural cubic splines substantially reduces estimation bias near the boundaries.

Although the basis construction in (4) is based on the truncated power series, one may see that our definition is slightly different from the form of the truncated power natural cubic splines commonly used in the literature, e.g., equations (5.4) and (5.5) in Hastie et al. (2009). It can be shown that the basis terms in (4) span the identical piecewise cubic polynomial space with natural boundary conditions, i.e., there is an isomorphism between the basis terms in (4) and the other natural cubic spline basis functions. However, our definition in (4) has a very nice property that inserting a new knot-point $t_* \in (t^L, t^U)$ simply leads to adding a new basis term defined as $N(\cdot; t^L, t^U, t_*)$ into the set $\mathcal{N} = \{N_k, k = 0, 1, \ldots, M+1\}$ without altering the current basis terms in $\mathcal{N}$. This property may not be satisfied with other natural cubic spline basis functions, such as the natural cubic B-spline basis and the basis in equations (5.4) and (5.5) of Hastie et al. (2009), because a single basis term may depend on more than two knots and inserting a knot-point may alter other basis terms. This fact makes our basis terms in (4) far more attractive for model

selection-based approaches (see Sections 3.2 and 3.3). To our knowledge, this work is the first to utilize the form of natural cubic splines in (4). The properties are formalized in the following propositions.

Proposition 1. *The set $\mathcal{N}$ is a basis for the cubic spline space with the natural boundary conditions.*

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

Proofs of Propositions 1 and 2 are provided in the supplementary material. We choose our basis terms b^*_{jk} using the natural cubic spline basis functions in (4). Specifically, using the observed values of the predictors variables, say $\{x_{ij}, i = 1, \ldots, n, j = 1, \ldots, p\}$, we set the boundary knots as $\xi^L_j = \min_{1 \le i \le n} x_{ij}$ and $\xi^U_j = \max_{1 \le i \le n} x_{ij}$ for each j. Then, with a given set of knots $\xi_j = \{\xi_{j1}, \ldots, \xi_{jL_j}\}$ satisfying $\xi^L_j < \xi_{j1} < \cdots < \xi_{jL_j} < \xi^U_j$, the unrestricted basis terms are chosen as

$$b^*_{j1}(\cdot) = N_1(\cdot), \quad b^*_{j,k+1}(\cdot) = N(\cdot; \xi^L_j, \xi^U_j, \xi_{jk}), \qquad k = 1, \ldots, L_j. \tag{5}$$

The class of spline functions is highly dependent on specification of knots $\xi = \{\xi_1, \ldots, \xi_p\}$. It is thus essential to choose a suitable knot placement to capture the local and global functional characteristics while avoiding overfitting. From the Bayesian point of view, a convenient approach is to let the data choose the most appropriate knots ξ from a predetermined set Ξ via BMS. The idea has become widely accepted in the literature (e.g., Smith and Kohn, 1996; Denison et al., 1998; 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, i.e., $\xi \in \Xi$. A richer Ξ provides flexible estimation for the regression spline functions, but may cause computational inefficiency. Under the Bayesian framework, specifying Ξ by a predetermined law can be viewed as assigning a prior distribution on ξ over the infinite-dimensional space for all possible knot locations with restricted support Ξ. The key to success is how to put a prior on ξ with a suitably restricted support Ξ. A few possibilities for a prior on ξ will be discussed in Section 3.

One additional advantage of the formulation in (5) is that the fully linear relationship is also easily characterized by our specification. More precisely, if ξ_j is empty, the basis consists only of the linear term b^*_{j1}. This situation is especially beneficial when a predictor variable is binary or is assumed to have a linear effect for any reason. In this scenario, we may simply set the corresponding ξ_j to be an empty set, which is viewed as assigning 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 subsumed by our construction without any modifications.

One of the main advantages of the BMS-based approaches to nonparametric regression is that they provide model-averaged estimates rather than resorting to a specific knot location. Our goal is to examine model-averaged estimates of a functional $\mathcal{L} : (\alpha, f_1, \ldots, f_p) \mapsto \mathcal{L}(\alpha, f_1, \ldots, f_p)$ of

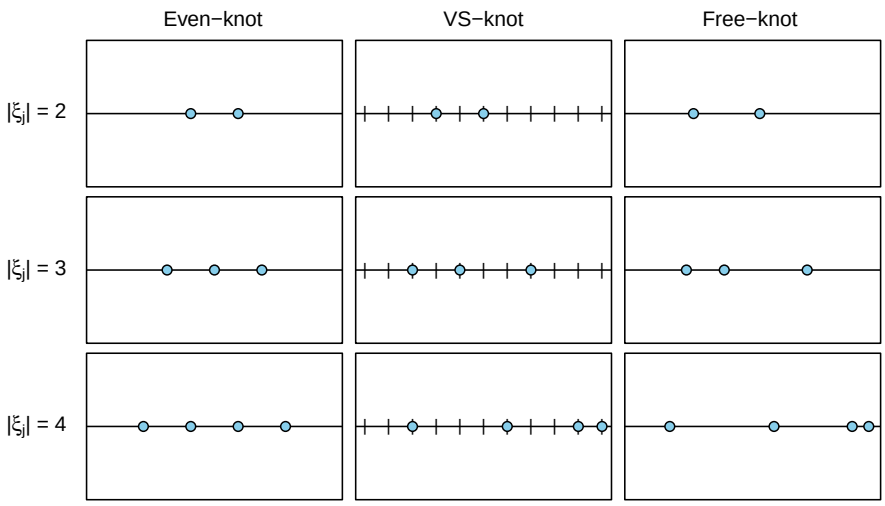


Figure 1: A graphical illustration of the three strategies for constructing Ξ discussed in Sections 3.1–3.3. For the even-knot splines, the locations of knots are determined deterministically once $|\xi_j|$ is chosen. The VS-knot splines select knot-points from a pre-determined set of locations. The free-knot splines are the most flexible and have no such limitation.

interest, which is parameterized by the 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 functions $f_j(x_j)$, $j = 1, \ldots, p$, at some $x = (x_1, \ldots, x_p)^T$. The model-averaged posterior of a functional is given by

$$\pi\big(\mathcal{L}(\alpha, f_1, \ldots, f_p) \mid Y\big) = \int_{\Xi} \pi\big(\mathcal{L}(\alpha, f_1, \ldots, f_p) \mid \xi, Y\big)\, d\Pi(\xi \mid Y). \tag{6}$$

A key to our Bayesian procedure is how to assign a prior distribution for model selection and how to explore the posterior distribution of ξ, i.e., $\Pi(\xi \mid Y)$. In what follows, we write $B_\xi = B$, $\beta_\xi = \beta$, $J_\xi = J$, and $\eta_\xi = \alpha 1_n + B_\xi \beta_\xi$ to emphasize the dependency on ξ. Observe that $J_\xi = p + \sum_{j=1}^{p} |\xi_j|$, where $|\xi_j|$ is the number of knots ξ_j, $j = 1, \ldots, p$.

3 Priors for knots

Our main objective is to explore the posterior distribution of a functional $\mathcal{L}(\alpha, f_1, \ldots, f_p)$. To obtain a model-averaged estimate, we need to (numerically) evaluate the integration in (6), which requires exploring the posterior distribution $\Pi(\alpha, \beta_\xi, \xi \mid Y)$ (or $\Pi(\alpha, \beta_\xi, \xi, \phi, \mid Y)$ if ϕ is unknown). To this end, we need to specify a prior distribution $\Pi(\alpha, \beta_\xi, \xi)$ (or $\Pi(\alpha, \beta_\xi, \xi, \phi)$) jointly over the parameter space. In this section, we will first look at a few different options for specifying $\Pi(\xi)$ for knots. Priors for the remaining parameters will be specified in Section 4.

We will see that our prior for β_ξ requires B_ξ to be of full-column rank (see (11) below). We thus choose $\Pi(\xi)$ with the condition that B_ξ is of full-column rank with prior probability one, i.e., $\Pi(\text{rank}(B_\xi) = J_\xi) = 1$. Intuitively, ξ_j can consist of any singletons lying on the interval (ξ_j^L, ξ_j^U), meaning that that the intrinsic parameter space for ξ_j is infinite-dimensional. However, finite

truncation to restricted support may be helpful for computational reasons. As mentioned earlier, we denote by Ξ the induced support of a prior on ξ. The support Ξ restricts the function class generated by our natural cubic spline basis terms. A smaller space reduces model complexity but may fail to capture local and global feature of the target function. This means that the restricted support Ξ balances estimation quality and computational efficiency, and it is crucial to choose a prior with suitable Ξ. There have been various ideas of specifying Ξ for $\Pi(\xi)$. In this section, we gather the existing strategies for constructing Ξ that are widely accepted in the literature and classify them into three categories. The three approaches are described in Sections 3.1–3.3 in detail. Figure 1 provides a graphical summary of the strategies.

3.1 Even-knot splines: equidistant knots

The simplest but 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 shown to be successful empirically and theoretically (e.g., Rivoirard and Rousseau, 2012; De Jonge and Van Zanten, 2012; Shen and Ghosal, 2015). We refer to this approach as the *even-knot splines*. The name should be carefully understood, as evenness may be assessed by the empirical measure rather than a geometric distance.

To be specific, a prior is assigned on the number $|\xi_j|$ of knots ξ_j for $j = 1, \ldots, p$, and the remaining specification on ξ is completed automatically by a given rule. For example, with a given number $|\xi_j|$, the knots ξ_j may be equally spaced or chosen as the quantiles of $\{x_{ij}, i = 1, \ldots, n\}$. We prefer the latter, as it guarantees full-column rank of B_ξ as soon as $J_\xi < n$ under mild conditions on the predictor variables. For computational reasons, it is also useful to put a cap on each $|\xi_j|$ such that $|\xi_j| \le M_j$ for a predetermined $M_j < n$, though it is not necessary. The induced support is then defined as

$$\Xi_{EK} = \left\{ \xi : J_\xi < n,\, |\xi_j| \le M_j,\, \xi_{jk} = Q_j\left(\frac{k}{|\xi_j| + 1}\right),\, j = 1, \ldots, p,\, k = 1, \ldots, |\xi_j| \right\},$$

where Q_j is the quantile function of $\{x_{ij}, i = 1, \ldots, n\}$, for $j = 1, \ldots, p$. Examples of knots belonging to Ξ_{EK} are illustrated in Figure 1. With an unnormalized density function $q_j : \{0, 1, \ldots, M_j\} \to (0, \infty)$ for a discrete nonnegative random variable, the prior can be formally expressed as

$$\pi_{EK}(\xi) \propto \mathbb{1}(J_\xi < n) \prod_{j=1}^{p} q_j(|\xi_j|). \tag{7}$$

Further discussion on the density q_j will be provided in Section 3.4.

The key benefit of the prior in (7) is that it has low model complexity. That is, we can enumerate all possible models for moderately large p because $|\Xi_{EK}| \le \prod_{j=1}^{p}(1 + M_j)$. This enables MCMC-free posterior computation for relatively low-dimensional problems. If p is too large to list all possibilities, the Metropolis-Hastings algorithm can be useful to explore the model spaces with a proposal that increases or decreases $|\xi_j|$ at a time. Even in this situation, if p is moderately large, computation may be facilitated by saving the value of the marginal likelihood $p(Y \mid \xi)$ with current ξ and utilizing it whenever the same ξ is revisited. We observe that the storing idea works well unless p is too large.

Although the equidistant knots (in the sense of the empirical measure) substantially reduce the model complexity, the major downside also arises from its deterministic rule. To be more specific, due to the nature of the construction, it is impossible to deal with functions with spatially adaptive smoothness such as a Doppler function. This drawback motivates the need for more flexible constructions discussed in the next two subsections.

3.2 VS-knot splines: knot selection

The limitation of the even-knot splines in Section 3.1 can be relaxed by considering a prior inducing a richer Ξ that allows for spatial adaptation. This can be fulfilled by allowing knot placement to be also data-driven as well as the number of knots. A common strategy is setting a large set of basis functions and choosing important ones among the candidates via Bayesian variable selection. The idea was initiated by Smith and Kohn (1996) and has been widely used in the literature of 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., 2021). Because the approach is based on variable selection, we refer to it as the *VS-knot splines*.

Consider a set $\xi_j^c = \{\xi_{j1}^c, \ldots, \xi_{jM_j}^c\}$ of knot candidates such that $\xi_j^L < \xi_{j1}^c < \cdots < \xi_{jM_j}^c < \xi_j^U$ with large enough $M_j < n$. Similar to Section 3.1, ξ_j^c can be equidistant or determined by sample quantiles of $\{x_{ij}, i = 1, \ldots, n\}$, and we prefer the latter setup to ensure full-column rank of B_ξ. Then, the knots ξ_j are chosen as a subset of ξ_j^c (including an empty set) via BMS. Therefore, the support consists of all possible subsets of $\{\xi_1^c, \ldots, \xi_p^c\}$ with the restriction $J_\xi < n$, i.e.,

$$\Xi_{VS} = \left\{\xi : J_\xi < n,\, \xi_j \subset \xi_j^c,\, j = 1, \ldots, p\right\}.$$

Similar to Section 3.1, it is reasonable to assign an unnormalized density $q_j : \{0, 1, \ldots, M_j\} \to (0, \infty)$ to $|\xi_j|$. And then we give the equal weights to all knot locations conditional on $|\xi_j|$. The resulting prior is given by

$$\pi_{VS}(\xi) \propto \mathbb{1}(J_\xi < n) \prod_{j=1}^{p} q_j(|\xi_j|) \binom{M_j}{|\xi_j|}^{-1}. \tag{8}$$

The prior has been shown to be successful in adapting to spatially inhomogeneous smoothness (e.g., Chan et al., 2006; Jeong and Park, 2016; Jeong et al., 2017). The cardinality $|\Xi_{VS}| \leq 2^{\sum_{j=1}^p M_j}$ shows that it is impractical to enumerate all possible models, and therefore MCMC will be useful to explore the model spaces. The standard Gibbs sampling and the Metropolis-Hastings algorithm can be easily applied to this setup (Dellaportas et al., 2002). Efficiency may be improved by block updates (Kohn et al., 2001; Jeong et al., 2021) or adaptive sampling (Nott and Kohn, 2005; Ji and Schmidler, 2013). Moreover, since Ξ_{VS} is finite-dimensional, the idea of storing the marginal likelihood in Section 3.1 may still be viable. However, our experience shows that this approach is effective only when p is very small because of the memory issue, e.g., $p \leq 2$, and therefore we do not purse this direction.

We emphasize that our basis system in (5) is particularly useful for the VS-spline approach. Due to Proposition 2, knot selection is naturally translated into basis selection with the linear term b_{j1}^* being always included. This property makes the computation straightforward with the basis system in (4) and (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 $\xi_j^c = (\xi_{j1}^c, \ldots, \xi_{jM_j}^c)$, and then choose important columns of B_j^c, while always including the first column for the linear term. As noted above, this is not possible with other natural cubic spline basis functions, such as B-splines or the one commonly used in the literature (Hastie et al., 2009). For these basis functions, a basis term $b_{j,k+1}^*$ is potentially specified with more than two knot-points for some k, and therefore inserting or deleting a knot-point may alter more than two basis terms, leading to conflict between knot selection and basis selection.

3.3 Free-knot splines

The strategy of the VS-splines in Section 3.2 chooses important knot locations among a set of predetermined candidates, thereby the resulting knots are not equally spaced so that they can account for spatially varying degree of smoothness. Despite this flexibility, there is a further desire to relax the restriction coming from the discrete set of knot candidates, with a fully nonparametric setup by allowing knots to be any singletons in the given range as soon as the induced B_ξ is of full-column rank. The idea is referred to as the *free-knot splines* (Denison et al., 1998; DiMatteo et al., 2001).

We still need the condition $J_\xi < n$, but the full-column rank should be incorporated in the prior more explicitly than the previous priors, as a restricted number of basis terms does not always guarantee the full rank with this unrestricted setup. For example, if there are too many knots in a narrow region, the resulting basis matrix may not be of full-column rank, and such a situation should be prevented by prior specification. Similar to Section 3.1, it can be useful to put a cap on each $|\xi_j|$ such that $|\xi_j| \leq M_j$ for a predetermined $M_j < n$. The resulting support of ξ is

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

It is clear that Ξ_{FK} is uncountable. The prior is specified in a manner similar to (8), but because the map $|\xi_j| \mapsto \xi_j$ is a surjection, not a bijection, a conditional prior density of ξ_j given $|\xi_j|$, denoted by $\tilde{q}_j(\cdot \mid |\xi_j|)$, should be specified on the corresponding support. Following DiMatteo et al. (2001), we choose $\tilde{q}_j$ induced by the uniform prior on the $|\xi_j|$-simplex by scaling (ξ_j^L, ξ_j^U) to $(0, 1)$. With an unnormalized density $q_j : \{0, 1, \ldots, M_j\} \to (0, \infty)$, the prior on ξ_j is formalized as

$$\pi_{FK}(\xi) \propto \mathbb{1}(J_\xi < n,\, \mathrm{rank}(B_\xi) = J_\xi) \prod_{j=1}^{p} q_j(|\xi_j|) \tilde{q}_j(\xi_j \mid |\xi_j|).$$

Our free-knot spline prior is slightly more general than the original construction by DiMatteo et al. (2001) in that they made the restriction that at least knot must be included, whereas we relaxed such limitation to account for a completely linear effect using an empty knot. The posterior distribution can be explored by the reversible jump MCMC with birth, death, and relocation proposals (DiMatteo et al., 2001) The computation is thus generally more demanding than the VS-knot splines. The accompanying benefit is that the prior is inherently more flexible than the one in (8) and has the better ability to approximate the target functions. However, our experience shows that there is no significant improvement for the free-knot splines in most

practical examples. Considering that the reversible jump MCMC is often inefficient, this fact points out that the inadvertent use of the free-knot splines should be avoided. Our simulation study shows that, while the performance measures for the free-knot splines are comparable to those for the VS-knot splines, the sampling efficiency (measured by the ratio of the effective sample size and the runtime) of the free-knot splines is significantly lower.

Similar to the VS-knot splines approach, the basis construction in (4) and (5) is useful for the free-knot splines. Due to Proposition 2, adding or removing a knot-point corresponds to adding or removing the corresponding basis term. Hence the reversible jump MCMC can be carried out by adding or removing a column of the matrix, without the need for reconstruction of the entire basis terms.

3.4 More about the prior distribution on $|\xi_j|$

The priors described in Sections 3.1–3.3 are commonly constructed of the unnormalized prior density q_j on $|\xi_j|$. Here we discuss more about q_j. To achieve the desired optimal properties in nonparametric regression, it has been shown that priors for the BMS-based methods must possess suitably decaying tail properties (e.g, Shen and Ghosal, 2015; Ročková and van der Pas, 2020; Jeong and Rockova, 2020). Such priors with the guaranteed tail property include Poisson and geometric distributions (with a suitable truncation if required by the setup) (Shen and Ghosal, 2015). Although they enjoy the theoretical flavor, one issue is that how to choose the right prior decay is unclear from the practical perspective. In the literature of variable selection, another common choice (believe to be weakly informative) is a discrete uniform distribution on $\{0, 1, \ldots, M_j\}$, resulting in the so-called hierarchical uniform prior on ξ_j (Kohn et al., 2001; Cripps et al., 2005; Scott and Berger, 2010). To be balanced between the theoretical flavor in nonparametric regression and the practical grounds in model selection, we set our default prior to be a truncated geometric distribution with small success probability ϖ, i.e., the unnormalized density is

$$q_j(u) = (1-\varpi)^u \varpi, \quad u = 0, 1, \ldots, M_j. \tag{9}$$

With sufficiently small ϖ, the prior in (9) mimics the discrete uniform distribution, while still preserving the desired tail property for the optimality.

The unnormalized density q_j can also be specified in a different manner to reduce the model complexity. As noted in Section 2, our model formulation subsumes GAPLMs, where a few predictor variables are assumed to have the linear effect (e.g., binary variables). Since this restriction is achieved by fixing some f_j to be composed only by the linear basis term N_1, we can choose a point mass at zero as q_j for such j, i.e., $\Pi(|\xi_j| = 0) = 1$, while using the prior in (9) for nonparametric additive components.

As an additional note, we comment on the case where model selection is also of interest in GAMs. Specifically, one may be interested not only in estimation but also in model selection for GAMs by examining the posterior probability that some predictor variables have the linear effect. This is easily accomplished by investigating $\Pi(|\xi_j| = 0 \mid Y)$ in the collected MCMC draws. However, we highlight that the prior in (9) is tailored for the estimation problem in GAMs and it should be carefully used for such a model selection purpose. More explicitly, the prior in (9)

induces $\Pi(|\xi_j| = 0) = \varpi/[1 - (1 - \varpi)^{M_j+1}] \approx \varpi$ for small ϖ, and therefore small ϖ yields a very informative prior for the linear effect caused by $|\xi_j| = 0$. One might choose ϖ based on a prior belief for the linearity $|\xi_j| = 0$, but this may induce too fast decay for $|\xi_j|$. Another option is considering a mixture-type prior on $|\xi_j|$ (Jeong et al., 2021). Since this paper focuses on estimation in GAMs, we do not discuss such extensions in detail and direct the reader to Jeong et al. (2021) for further discussion on model selection in GAMs.

4 Mixtures of g-priors for model selection

Prior distributions should be specified to conclude the joint posterior distribution $\Pi(\alpha, \beta_\xi, \xi \mid Y)$ (or $\Pi(\alpha, \beta_\xi, \xi, \phi \mid Y)$ if ϕ is unknown). In Section 3, we discussed possible priors for ξ, $\Pi(\xi)$. The full Bayesian framework is completed once priors for the remaining model parameters are specified. The most crucial part is determining a prior for the knot-specific coefficients β_ξ, i.e., $\Pi(\beta_\xi|\xi)$, for which this study employs mixtures of g-priors (Liang et al., 2008). Here we elucidate mixtures of g-priors for the BMS-based approaches to GAMs.

4.1 Generalized additive models with known dispersion

We first focus on GAMs in (1) and (2) with known dispersion parameter ϕ; that is, we want to specify a prior distribution $\Pi(\alpha, \beta_\xi \mid \xi) = \Pi(\alpha)\Pi(\beta_\xi \mid \xi)$. Following the convention, we assign an improper uniform prior on the common parameter α, i.e.,

$$\pi(\alpha) \propto 1. \tag{10}$$

Next we discuss $\Pi(\beta_\xi \mid \xi)$. For model selection in linear regression, Zellner's g-prior is often preferred due to its computational efficiency and invariance to linear transformations (Zellner, 1986). In our spline setup, invariance to transformations of a design matrix is particularly appealing, because it is ideal if the procedure is invariant to a specific choice of basis functions as long as a target function space is correctly generated. Unfortunately, the computational advantage of the g-prior is generally lost in GAMs since normal priors are not conjugate to non-normal likelihoods, resulting in the inability to achieve a closed-form expression of the marginal likelihood,

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

where $p(Y \mid \alpha, \beta_\xi)$ is the likelihood evaluated with α and β_ξ. The computation for the posterior distribution in (6) becomes more challenging without the calculation of the marginal likelihood. Therefore, we consider approximating the marginal likelihood using the Laplace approximation with a suitable variant of the g-prior. The Laplace approximation has occasionally been used for nonparametric regression with g-priors (e.g., DiMatteo et al., 2001).

For the function $\theta = (h \circ b')^{-1}$, let $J_n(\hat{\eta}_\xi) = \text{diag}(-Y_i\theta''(\hat{\eta}_{\xi,i}) + (b \circ \theta)''(\hat{\eta}_{\xi,i}), i = 1, \ldots, n)$ be the observed information matrix of η_ξ evaluated at $\hat{\eta}_\xi$ (defined as the Hessian matrix of the negative log-likelihood), where $\hat{\eta}_\xi = (\hat{\eta}_{\xi,1}, \ldots, \hat{\eta}_{\xi,n})^T = \hat{\alpha}_\xi 1_n + B_\xi \hat{\beta}_\xi$ with the maximum likelihood estimators $\hat{\alpha}_\xi$ and $\hat{\beta}_\xi$ (assuming that $\hat{\alpha}_\xi$ and $\hat{\beta}_\xi$ exist). We restrict our attention to the case where $J_n(\hat{\eta}_\xi)$ is positive definite. This is usually the case except for a few extreme situations such as

the complete separation in logistic regression (Li and Clyde, 2018). In this paper, we deploy the following variant of the g-prior proposed by Li and Clyde (2018),

$$\beta_\xi \mid g, \xi \sim \mathrm{N}\big(0, g(\tilde{B}_\xi^T J_n(\hat{\eta}_\xi)\tilde{B}_\xi)^{-1}\big), \tag{11}$$

where $\tilde{B}_\xi = [I_n - \mathrm{tr}(J_n(\hat{\eta}_\xi))^{-1}1_n 1_n^T J_n(\hat{\eta}_\xi)]B_\xi$ is the matrix consisting of the columns of B_ξ centered by the weighted average with the diagonal elements of $J_n(\hat{\eta}_\xi)$. The prior in (11) requires $\tilde{B}_\xi^T J_n(\hat{\eta}_\xi)\tilde{B}_\xi$ to be invertible. Assuming that $J_\xi < n$, this is satisfied if and only if B_ξ is of full-column rank (observe that $J_n(\hat{\eta}_\xi)$ is positive definite and $\mathrm{rank}(B_\xi) = \mathrm{rank}(\tilde{B}_\xi)$ if $J_n < n$). The priors for ξ discussed in Section 3 impose this restriction in their constructions. Although the prior in (11) can be extended by adopting a generalized inverse, we do not pursue this direction. We refer the reader to Section 2.5 of Li and Clyde (2018) for further discussion when B_ξ is not of full-column rank.

In addition to the prior in (11), there are many other variants of the g-prior for the exponential family models (e.g., Kass and Wasserman, 1995; Hansen and Yu, 2003; Gupta and Ibrahim, 2009; Sabanés Bové and Held, 2011; Held et al., 2015). Here we adopt the prior in (11) as it provides a convenient expression for the approximate marginal likelihood. Also, our prior may better capture the large sample covariance structures and the local geometry than other variants of the g-prior (Li and Clyde, 2018). It is worth mentioning that the prior in (11) is dependent on the observation vector Y, and thus should be accepted by the empirical Bayes philosophy. Notwithstanding this point, we suppress the dependency on Y in the expression of (11) for notational convenience.

Integrating the second order Taylor expansion for the likelihood with respect to (10) and (11), we obtain

$$p(Y \mid g, \xi) = p(Y \mid \hat{\eta}_\xi)\mathrm{tr}(J_n(\hat{\eta}_\xi))^{-1/2}(g+1)^{-J_\xi/2}\exp\left(-\frac{Q_\xi}{2(g+1)}\right), \tag{12}$$

where $p(Y \mid \hat{\eta}_\xi)$ is the likelihood evaluated at $\hat{\eta}_\xi$ with given ξ and $Q_\xi = \hat{\beta}_\xi^T \tilde{B}_\xi^T J_n(\hat{\eta}_\xi)\tilde{B}_\xi\hat{\beta}_\xi$ is the Wald statistic. The expression shows the marginal likelihood is highly sensitive to predetermined g. A suitable value for g has been extensively debated in the literature. The most common choice is letting $g = n$, called the unit information prior (Kass and Wasserman, 1995). This idea has also been widely adopted in the literature of nonparametric regression via BMS (e.g. Gustafson, 2000; DiMatteo et al., 2001; Kohn et al., 2001). From the Bayesian point of view, the unit information prior can be viewed as a point mass prior at $g = n$, i.e., $\Pi(g) = \delta_n(g)$ with the Dirac measure δ_b at b. However, it has been reported that putting a suitable prior distribution on g, called a mixture of g-priors, leads to improved empirical performance while addressing the paradoxes in BMS (Liang et al., 2008; Li and Clyde, 2018). To unify various mixtures of g-priors, we use a broad 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), i.e.,

$$\frac{1}{g+1} \sim \mathrm{tCCH}\left(\frac{a}{2}, \frac{b}{2}, r, \frac{s}{2}, \nu, \kappa\right), \quad a, b, \kappa > 0, \quad r, s \in \mathbb{R}, \quad \nu \geq 1. \tag{13}$$

See the supplementary material for the density of the tCCH distribution.

	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	$(n+1)/(J_\xi+1)$	1	$g = O(n)$
Intrinsic	1	1	1	0	$(n+J_\xi+1)/(J_\xi+1)$	$(n+J_\xi+1)/n$	$g = O(n)$

Table 1: Distributions belonging to the tCCH family.

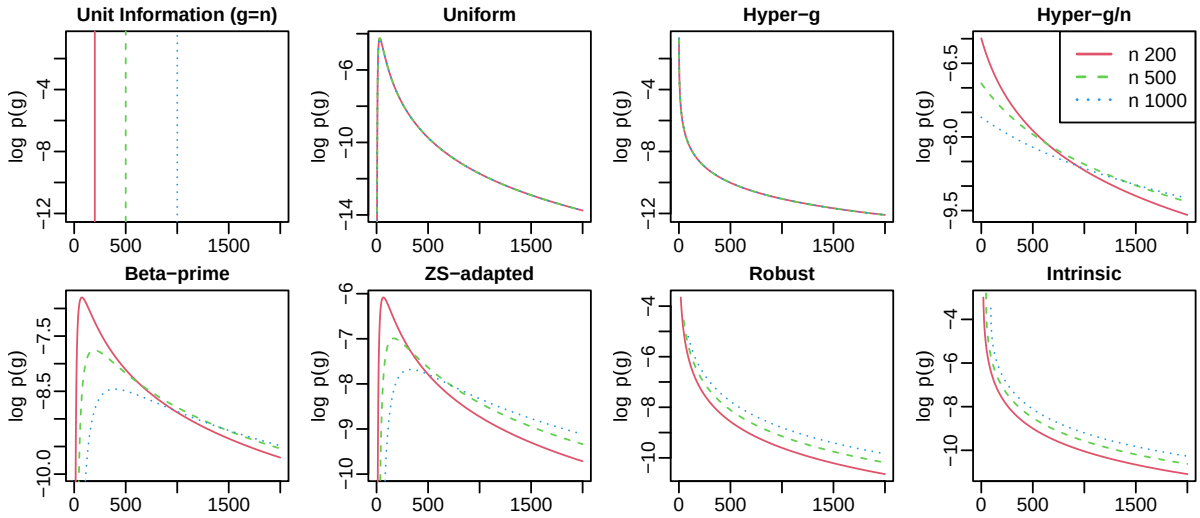


Figure 2: Distributions belonging to the tCCH family for $n = 200, 500, 1000$, with $J_\xi = 10$ if required.

Table 1 shows a few distributions belonging to the tCCH family: 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). According to Li and Clyde (2018), the prior distributions are classified into two groups based to the prior concentration: $g = O(1)$ and $g = O(n)$. (Notation can be misleading; it refers to the order of concentration of the distribution rather than the value of g. Maruyama and George (2011) also uses the same notation.) Figure 2 illustrates how the concentration of each prior distribution on g behaves.

The resulting marginal likelihood is expressed as

$$p(Y \mid \xi) = p(Y \mid \hat{\eta}_\xi)\mathrm{tr}(J_n(\hat{\eta}_\xi))^{-1/2}\nu^{-J_\xi/2}\exp\left(-\frac{Q_\xi}{2\nu}\right)\frac{B((a+J_\xi)/2, b/2)}{B(a/2,b/2)} \\ \times \Phi_1\left(\frac{b}{2}, r, \frac{a+b+J_\xi}{2}, \frac{s+Q_\xi}{2\nu}, 1-\kappa\right) \Big/ \Phi_1\left(\frac{b}{2}, r, \frac{a+b}{2}, \frac{s}{2\nu}, 1-\kappa\right), \tag{14}$$

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

eral the evaluation of Φ_1 cannot be performed analytically and must instead rely on numerical approximation. To calculate Φ_1, we use the Gaussian-Kronrod quadrature routine available in the Boost `C++` library.

With the tCCH prior on $(g+1)^{-1}$, it is not difficult to show that the approximate posterior for $((g+1)^{-1}, \alpha, \beta_\xi)$ conditional on ξ is given by

$$\begin{aligned} \frac{1}{g+1} \mid Y, \xi &\sim \text{tCCH}\left(\frac{a+J_\xi}{2}, \frac{b}{2}, r, \frac{s+Q_\xi}{2}, \nu, \kappa\right), \\ \alpha \mid Y, g, \xi &\sim \text{N}\big(\hat{\alpha}_\xi, \text{tr}(J_n(\hat{\eta}_\xi))^{-1}\big), \\ \beta_\xi \mid Y, g, \xi &\sim \text{N}\left(\frac{g}{g+1}\hat{\beta}_\xi, \frac{g}{g+1}(\tilde{B}_\xi^T J(\hat{\eta}_\xi)\tilde{B}_\xi)^{-1}\right). \end{aligned} \tag{15}$$

The expression is also valid for the unit information prior by replacing the first line with the point mass posterior, i.e., $\Pi(g|Y,\xi) = \delta_n(g)$. The joint posterior $\Pi(\alpha, \beta_\xi, \xi, g \mid Y)$ is fully specified with the posterior in (15) and the marginal posterior of ξ, i.e., $\Pi(\xi \mid Y)$. The latter one is fulfilled by specifying a prior $\Pi(\xi)$ as in Section 3 and deploying the expression for the approximate marginal likelihood $p(Y \mid \xi)$ in (14) (or $p(Y \mid g, \xi)$ in (12) for the unit information prior).

4.2 Gaussian additive regression with unknown precision

Thus far we have focused on GAMs with known dispersion parameter ϕ for the exponential family models. Now we consider a more classical setup with a Gaussian assumption on the distribution of Y_i, while treating ϕ as an unknown parameter. For Gaussian additive regression, a response variable Y_i is expressed as

$$Y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \phi^{-1}), \quad i = 1, \ldots, n, \tag{16}$$

where the precision parameter ϕ is typically unknown. Although model (16) also belongs to the GAM framework, things are different due to the unknown precision ϕ. Let $\eta = (\eta_1, \ldots, \eta_n)^T$ be the vector of mean response, i.e., $\eta_i = E(Y_i)$. We parameterize η as $\eta = \alpha 1_n + B_\xi \beta_\xi$ with α, B_ξ, and β_ξ defined in Section 2. Following the convention, an improper prior is put on (α, ϕ), i.e.,

$$\pi(\alpha, \phi) \propto 1/\phi.$$

Using the fact that the information matrix of a Gaussian distribution is the identity matrix, one can easily check that the prior in (11) is reduced to the usual g-prior distribution (Zellner, 1986),

$$\beta_\xi \mid \phi, g, \xi \sim N\big(0, g\phi^{-1}(B_\xi^T B_\xi)^{-1}\big). \tag{17}$$

(For the Gaussian case, we obtain $\tilde{B}_\xi = B_\xi$ since the columns of B_ξ are centered.)

Combining the marginal likelihood with one of the priors on ξ discussed in Section 3, we can obtain the marginal posterior of ξ, $\Pi(\xi \mid Y)$. The calculation of the marginal likelihood is complicated because ϕ needs to be integrated out along with g. First, it is well known that

$$p(Y \mid g, \xi) = p(Y \mid \varnothing) \frac{(1+g)^{(n-J_\xi-1)/2}}{[1+g(1-R_\xi^2)]^{(n-1)/2}},$$

where $p(Y \mid \varnothing) = n^{-1/2}(2\pi)^{-(n-1)/2}\Gamma((n-1)/2)(\|Y - n^{-1}1_n 1_n^T Y\|^2/2)^{-(n-1)/2}$ is the marginal likelihood with the intercept-only model and $R^2_\xi = \|B_\xi(B_\xi^T B_\xi)^{-1}B_\xi^T Y\|^2/\|Y - n^{-1}1_n 1_n^T Y\|^2$ is the coefficient of determination with ξ. For the unit information prior $\Pi(g) = \delta_n(g)$, the marginal likelihood $p(Y \mid \xi)$ is readily available from the above expression. Assigning the tCCH prior in (13) to $(g+1)^{-1}$, Li and Clyde (2018) showed that

$$\begin{aligned} p(Y \mid \xi) = {} & \frac{p(Y \mid \varnothing)}{\nu^{J_\xi/2}[1-(1-\nu^{-1})R^2_\xi]^{(n-1)/2}} \frac{B((a+J_\xi)/2, b/2)}{B(a/2,b/2)} \\ & \times \Phi_1\left(\frac{b}{2}, \frac{n-1}{2}, \frac{a+b+J_\xi}{2}, \frac{s}{2\nu}, \frac{R^2_\xi}{\nu-(\nu-1)R^2_\xi}\right) \Big/ {}_1F_1\left(\frac{b}{2}, \frac{a+b}{2}, \frac{s}{2\nu}\right), \end{aligned} \tag{18}$$

if $r = 0$ (or $\kappa = 1$ equivalently), and

$$\begin{aligned} p(Y \mid \xi) = {} & \frac{p(Y \mid \varnothing)\kappa^{(a+J_\xi-2r)/2}}{\nu^{J_\xi/2}(1-R^2_\xi)^{(n-1)/2}} \frac{B((a+J_\xi)/2, b/2)}{B(a/2,b/2)} \left[{}_2F_1\left(r, \frac{b}{2}; \frac{a+b}{2}; 1-\kappa\right)\right]^{-1} \\ & \times F_1\left(\frac{a+J_\xi}{2}; \frac{a+b+J_\xi+1-n-2r}{2}, \frac{n-1}{2}; \frac{a+b+J_\xi}{2}; 1-\kappa, 1-\kappa-\frac{R^2_\xi\kappa}{(1-R^2_\xi)v}\right), \end{aligned} \tag{19}$$

if $s = 0$, where ${}_1F_1(\alpha,\gamma,x) = \Phi_1(\alpha,0,\gamma,x,0)$ is the confluent hypergeometric function, ${}_2F_1(\beta,\alpha;\gamma;y) = \Phi_1(\alpha,\beta,\gamma,0,y)$ is the Gaussian hypergeometric function, and F_1 is the the Appell hypergeometric function defined as $F_1(\alpha;\beta,\beta';\gamma;x,y) = B(\gamma-\alpha,\alpha)^{-1}\int_0^1 u^{\alpha-1}(1-u)^{\gamma-\alpha-1}(1-xu)^{-\beta}(1-yu)^{-\beta'}du$. The prior distributions listed in Table 1 belong to either one of the above two cases. The expressions may further be simplified depending on the hyperparameters of the tCCH prior, but numerical evaluation of the transcendental functions is mostly required. The only exception is the beta-prime prior, which provides a closed-form expression for the marginal likelihood without a hypergeometric-type transcendental function. See Supplementary material for expressions for the marginal likelihood with each mixture of g-prior.

For the Gaussian case, the conditional posterior $\Pi((g+1)^{-1} \mid Y, \xi)$ is no longer a conjugate update of the tCCH prior. Still, it is simplified with some hyperparameter specification of the tCCH prior and sampling from $\Pi((g+1)^{-1} \mid Y, \xi)$ is easily carried out by introducing auxiliary variables. In particular, the beta-prime prior provides an exact sampling scheme from a beta distribution. We refer the reader to Supplementary material for more details on sampling procedures. The remaining specification of the joint posterior can easily be derived by direct calculations as

$$\begin{aligned} \phi \mid Y, g, \xi &\sim \text{Gamma}\left(\frac{n-1}{2}, \frac{\|Y - n^{-1}1_n 1_n^T Y\|^2[1+g(1-R^2_\xi)]}{2(1+g)}\right), \\ \alpha \mid Y, \phi, g, \xi &\sim \text{N}\left(n^{-1}1_n^T Y, n^{-1}\phi^{-1}\right), \\ \beta_\xi \mid Y, \phi, g, \xi &\sim \text{N}\left(\frac{g}{g+1}\hat{\beta}_\xi, \frac{g\phi^{-1}}{g+1}(B_\xi^T B_\xi)^{-1}\right). \end{aligned}$$

4.3 Complexity penalty by mixtures of g-priors

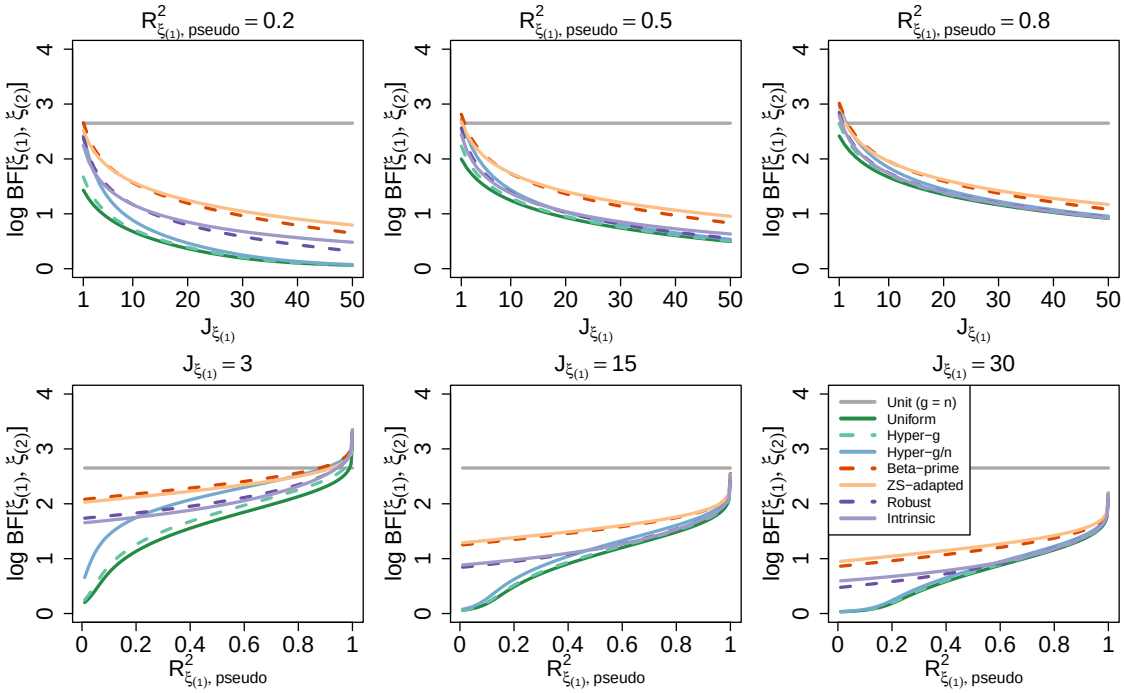
The influence of g has been found to be crucial for reasonable sparsity in model selection with the g-prior (Kass and Raftery, 1995); a large g favors sparse models, while a small g advocates

complex models. In particular, choosing a suitable g is extremely important in our additive model setup, as it directly controls the smoothness of the additive functions. In the literature of nonparametric regression with basis expansion, many previous works rely on the unit information prior induced by the choice $g = n$ (e.g. Gustafson, 2000; DiMatteo et al., 2001; Kohn et al., 2001). However, as noted above, a mixture of g-priors leads to improved empirical performance in BMS (Liang et al., 2008; Li and Clyde, 2018). There have been attempts to put a prior on g for nonparametric regression (Jeong and Park, 2016; Jeong et al., 2017; Francom et al., 2018; Park and Jeong, 2018; Francom and Sansó, 2020; Jeong et al., 2021), but a thorough investigation into how the priors differ from the unit information prior is lacking. In this section, we will illuminate how the behavior of mixtures of g-priors differs from that of the unit information prior with $g = n$ and investigate why the unit information prior may not be a good choice for estimation in GAMs.

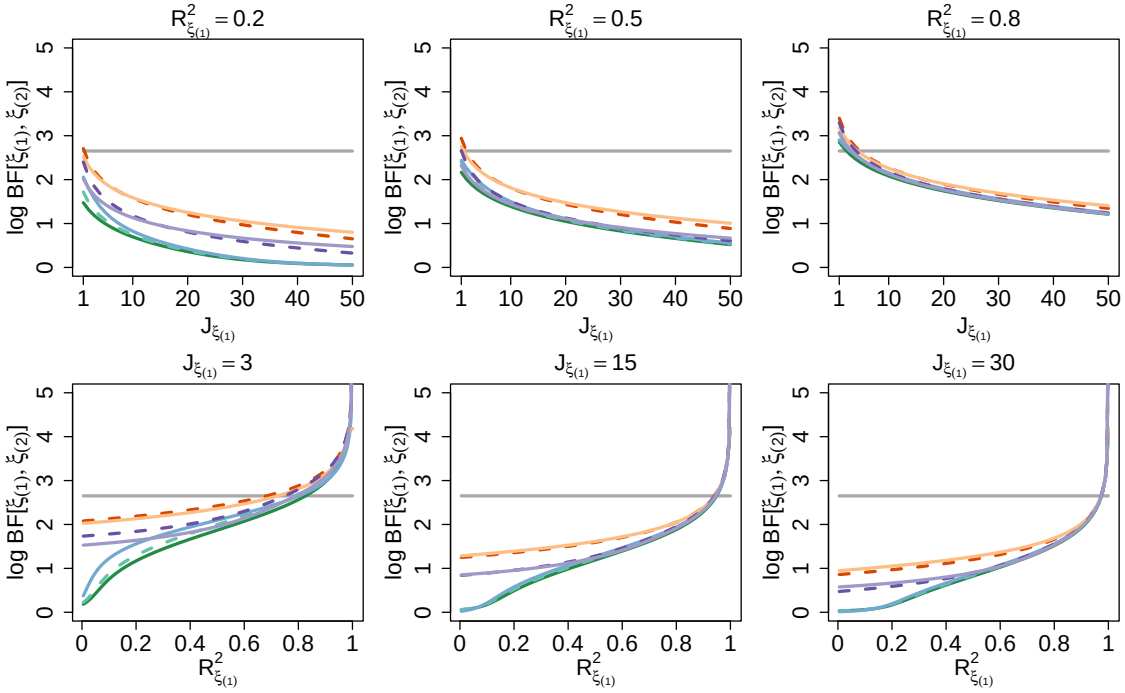
We will use the Bayes factor for comparisons. For two knots $\xi_{(1)}$ and $\xi_{(2)}$, the Bayes factor of $\xi_{(1)}$ to $\xi_{(2)}$ is defined as $BF[\xi_{(1)}; \xi_{(2)}] = p(Y \mid \xi_{(1)})/p(Y \mid \xi_{(2)})$. For the exponential family models with known ϕ, the marginal likelihood $p(Y \mid \xi)$ is expressed by (12) with $g = n$ for the unit information prior and by (14) for mixtures of g-priors induced by the tCCH prior on $(g+1)^{-1}$. Similarly, the Gaussian regression model defines the Bayes factor using the expressions for the marginal likelihood provided in Section 4.2.

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 $\hat{\eta}_{\xi_{(1)}} = \hat{\eta}_{\xi_{(2)}}$, i.e., the two knots contribute equally to model fit, but $\xi_{(1)}$ has one more redundant knot-point than $\xi_{(2)}$. The Bayes factor $BF[\xi_{(1)}; \xi_{(2)}]$ can be interpreted as a penalty against the more complex model with one more redundant knot-point; the larger value of the log Bayes factor, the stronger penalty to the larger model. We will see how the Bayes factor behaves as $J_{\xi_{(1)}}$ and the goodness-of-fit change. For Gaussian regression, the goodness-of-fit can be measured by 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 instead be used (Cox and Snell, 1989; Magee, 1990), albeit with the caveat that the maximum value may be less than one depending on the specific model (Nagelkerke, 1991). Since the Bayes factor is expressed with the Wald statistic Q_ξ, we approximate the pseudo-R^2 as a function of Q_ξ based on 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 $R^2_{\xi,\text{pseudo}} = 1 - \exp(-Q_\xi/n)$ to measure the goodness-of-fit in the exponential family models.

Figure 3 visualizes a toy example with $n = 1000$, which shows how the log Bayes factor $\log BF[\xi_{(1)}; \xi_{(2)}]$ behaves as $J_{\xi_{(1)}}$ (which is $J_{\xi_{(2)}} + 1$) and the goodness-of-fit change. It is clear that the unit information prior always yields a constant penalty regardless of $J_{\xi_{(1)}}$ and the goodness-of-fit. On the other hand, the first rows of Figure 3a and Figure 3b illustrate that the penalty functions produced by the mixture priors get weaker as the model size $J_{\xi_{(1)}}$ increases. This indicates that, when small models are compared, mixtures of g-priors favor the sparser model $(\xi_{(2)})$ unless there is a significant improvement in the marginal likelihood. In contrast, when large models are compared, the mixture priors advocate the more complex model $(\xi_{(1)})$ even if the gain is not clear enough. This property is desirable in improving the performance in GAMs, because the basis expansion produces a large model in general and we wish to detect local and global signals of the target functions that may be easily missed. The second rows of Figure 3a and Figure 3b show that the penalty functions induced by the mixture priors get stronger as the goodness-of-fit



(a) The exponential family model



(b) The Gaussian regression model

Figure 3: The log Bayes factor $\log BF[\xi_{(1)}; \xi_{(2)}]$ as a function of $J_{\xi_{(1)}}$ $(= J_{\xi_{(2)}} + 1)$ and $R^2_{\xi_{(1)},\text{pseudo}}$ $(= R^2_{\xi_{(2)},\text{pseudo}})$ for $n = 200$.

measurement increases. This is in accordance with intuition because if the goodness-of-fit is good enough, we may not want to go for the more complex model and instead choose for the sparser one, unless the complex model significantly improves the marginal likelihood. We underline once again that the unit information prior yields a constant penalty.

We saw that the unit information prior may be not be a suitable choice for estimation in GAMs, and mixtures of g-priors can be an alternative. However, it still remains unclear which mixture prior would perform the best for GAMs. Figure 3 also demonstrates differences among the mixtures of g-priors. The beta-prime and the ZS-adapted priors behave similarly and exhibit the strongest penalties among the candidates. The robust and intrinsic priors show similar decays to each other and are weaker than the beta-prime and the ZS-adapted priors. The two $O(1)$-type priors (uniform and hyper-g) yield the weakest penalties and are weaker than other $O(n)$-type priors. The hyper-g/n prior is somewhat special and appears to be close to the $O(1)$-type priors, though it belongs to the $O(n)$ family.

Although the figure shows the mixtures of g-priors behave differently, it hardly reveals the best mixture prior for GAMs. The answer to this question should be based on suitable simulation studies as provided in the following section. However, the following proposition provides an interpretation of where such discrepancies arise.

Proposition 3. *For the model in (1) and (2) with the prior in (11), 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. Then, for any positive integer k,*

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

The result also holds for the model in (16) with the prior in (17) if either $\kappa = 1$ or $s = 0$.

A proof of Proposition 3 is provided in the supplementary material. Proposition 3 implies that the Bayes factor $BF[\xi_{(1)}; \xi_{(2)}]$ is the conditional posterior mean of $(1+g)^{-k/2}$ induced by the unit information prior or the tCCH prior. Hence the proposition explicitly shows why the penalty function induced by the unit information prior is constant. The differences in the penalties by the mixture priors can be attributed to different posterior means of the shrinkage factor $(1+g)^{-k/2}$.

5 Numerical study

The main objective of this study is to understand how the mixtures of g-priors behave in the BMS-based approaches to estimation of GAMs. Section 4.3 provides a basic understanding of how the mixtures of g-priors penalize models, but it remains unclear which mixture prior is the most appropriate for GAMs. We also want to examine whether the BMS-based methods for GAMs outperform other Bayesian approaches for function estimation, e.g., Bayesian P-splines. To this end, this section provides extensive simulation studies.

5.1 Comparison among the mixtures of g-priors

We first conduct a simulation study that illuminates the difference in performance between the mixtures of g-priors for estimation of GAMs. As for the test functions, we consider the following

three uncentered functions $f_j^* : [-1, 1] \to \mathbb{R}$, $j = 1, 2, 3$:

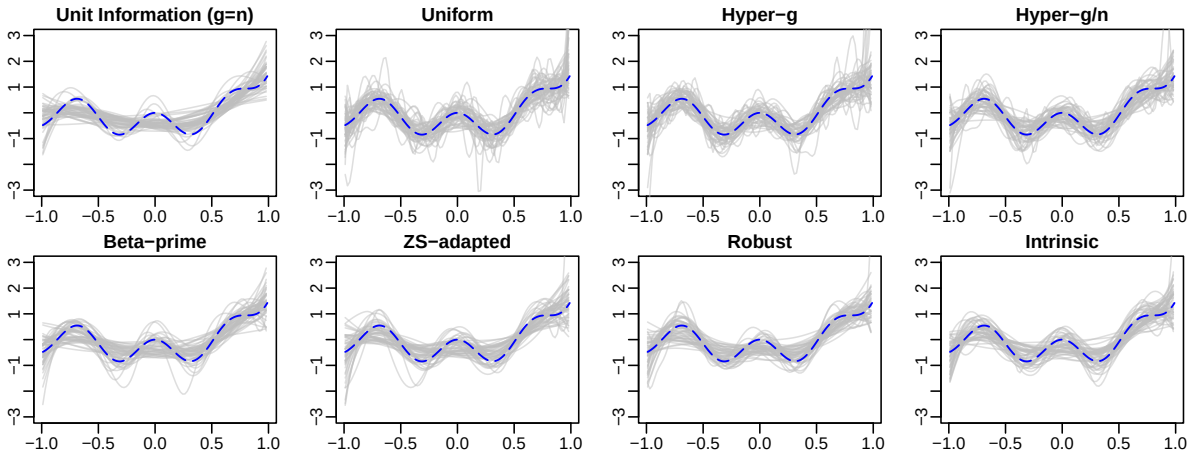
$$
\begin{aligned}
f_1^*(x) &= 0.5(2x^5 + 3x^2 + \cos(3\pi x) - 1), \\
f_2^*(x) &= \frac{21(3x + 1.5)^3}{8000} + \frac{21(3x - 2.5)^2 e^{3x+1.5}}{400} \sin\left(\frac{(3x + 1.5)^2 \pi}{3.2}\right) \mathbb{1}(-0.5 < x < 0.85), \\
f_3^*(x) &= x.
\end{aligned} \tag{20}
$$

Specifically, f_1^* is a nonlinear function which is not polynomial, f_2^* is a nonlinear function with locally varying smoothness, and f_3^* is a linear function. The two nonlinear functions f_1^* and f_2^* are modified from Gressani and Lambert (2021) and Francom and Sansó (2020), respectively. The functions are visualized in Figure 4 with suitable centering.

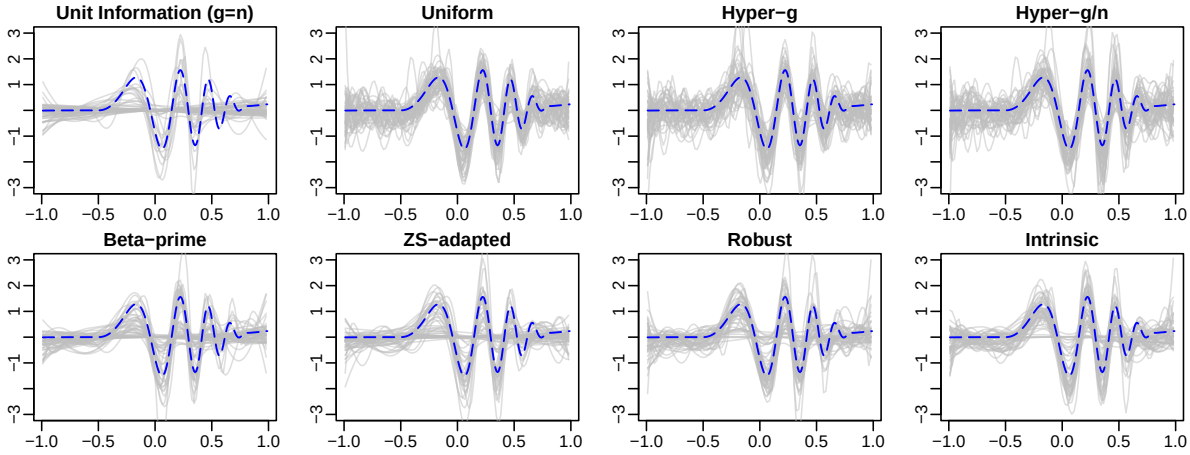
For each of $j = 1, 2, 3$, we generate the predictor $\eta_i = f_j^*(x_i) = \alpha + f_j(x_i)$ with x_i drawn independently from Unif$(-1, 1)$, where f_j is the centered version of f_j^* and α is the induced intercept. The test dataset is then generated by the exponential family model with η_i, $i = 1, \ldots, n$. That is, the regression models considered here are not additive, but rather comprise a single univariate function. It is certainly possible to carry out simulation using an additive structure, but we observe that a single univariate function sharpens the difference between the mixtures of g-priors. In this section, we only provide a simulation result for a nonlinear logistic regression model given by $Y_i \sim \text{Bernoulli}(e^{\eta_i}/(1 + e^{\eta_i}))$. The supplementary material includes simulation studies for Poisson regression $Y_i \sim \text{Poi}(e^{\eta_i})$ and Gaussian regression $Y_i \sim \text{N}(\eta_i, 1)$.

Among the three strategies for choosing Ξ described in Section 3, here we focus on the VS-knot splines approach since we observe that it generally outperforms the other methods in terms of both empirical performance and computational efficiency (the three strategies will be compared in Section 5.2). For each of the logistic regression models with $n = 500, 1000, 2000$, we generate 500 replications of datasets and estimate f_j using the VS-knot splines made up of 30 knot candidates with the unit information prior and the mixture priors summarized in Table 1. We calculate the root mean squared error (RMSE) and the coverage probabilities of the 95% pointwise credible bands at a few given points.

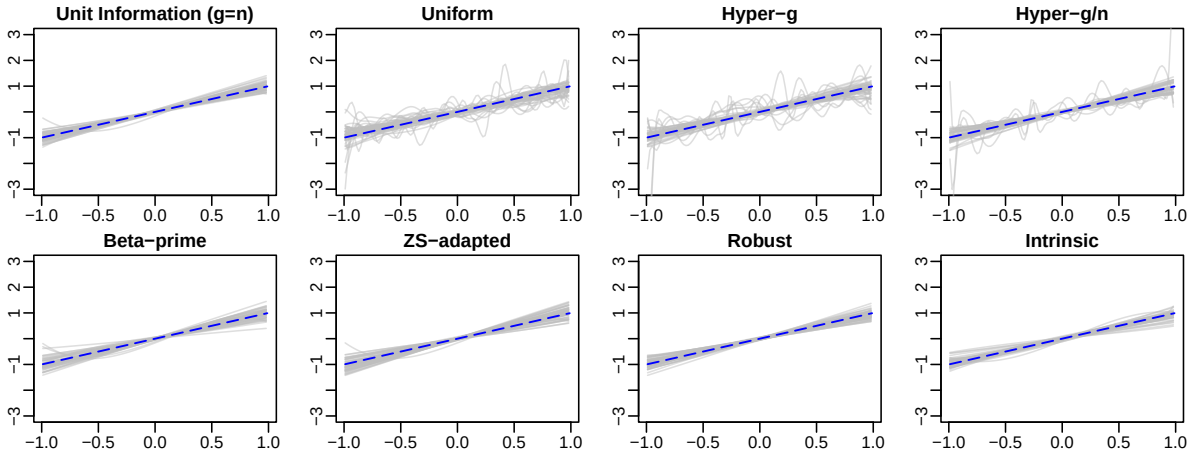
Figures 4 and 5 summarize the simulation results. As discussed in Section 4.3, our results show that the unit information prior behaves very differently from the mixture priors, and it generally underperforms for nonlinear function estimation. Specifically, the unit information prior often prefers simplistic models, which indicates that the current convention of employing the the unit information prior for function estimation may be inappropriate. The major challenge is determining which mixture prior is the most suited for function estimation. Based on our simulation analysis, we conclude that the intrinsic and robust priors are solid choices. Although the difference between the mixtures of g-priors is blurred for a large sample size, it is quite clear that intrinsic and robust priors outperform the other priors almost always; the beta-prime and ZS-adapted priors tend to exhibit undersmoothing, whereas the uniform, hyper-g, and hyper-g/n priors tend to exhibit oversmoothing. The simulation results for Gaussian regression and Poisson regression in the supplementary material also lead to the similar conclusion. The remaining question is how to determine the default prior between the intrinsic prior and the robust prior. We find that sampling from the posterior distribution of g is facilitated with the robust prior (the posterior reduces to a truncated gamma distribution for the exponential family model and to a



(a) Pointwise posterior mean estimates of f_1 in 50 replications

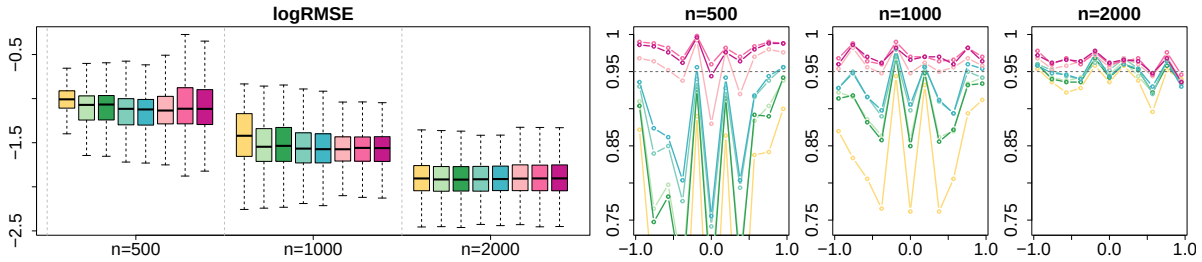


(b) Pointwise posterior mean estimates of f_2 in 50 replications

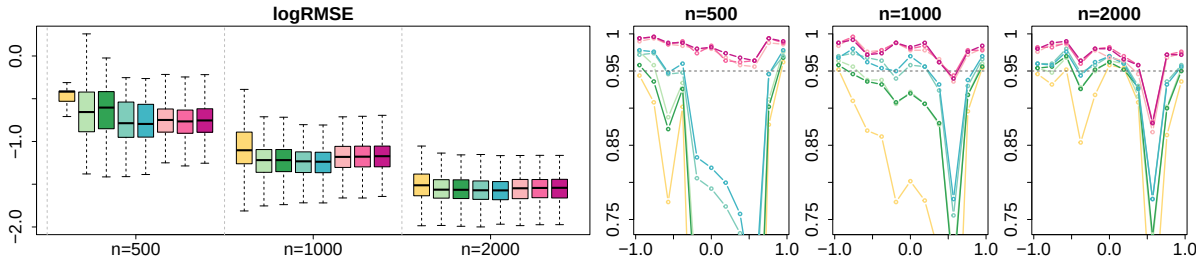


(c) Pointwise posterior mean estimates of f_3 in 50 replications

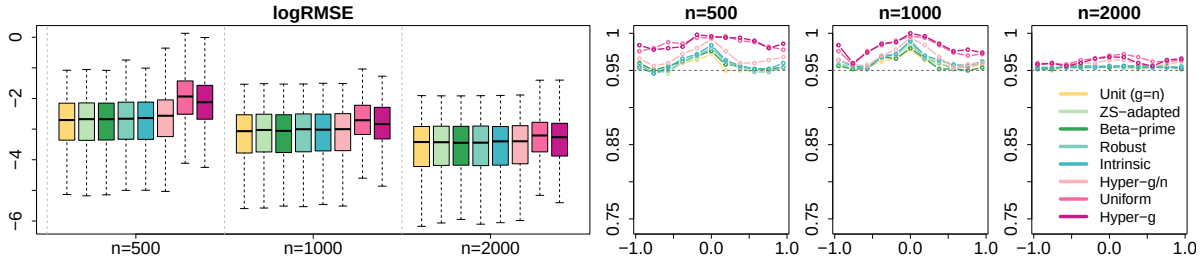
Figure 4: Estimates of (a) f_1, (b) f_2, and (c) f_3 in the nonparametric logistic regression models with $n = 500$. Pointwise posterior means of randomly chosen 50 replications (gray solid) and the true function (blue dashed).



(a) The log RMSE and the coverage probabilities for f_1



(b) The log RMSE and the coverage probabilities for f_2



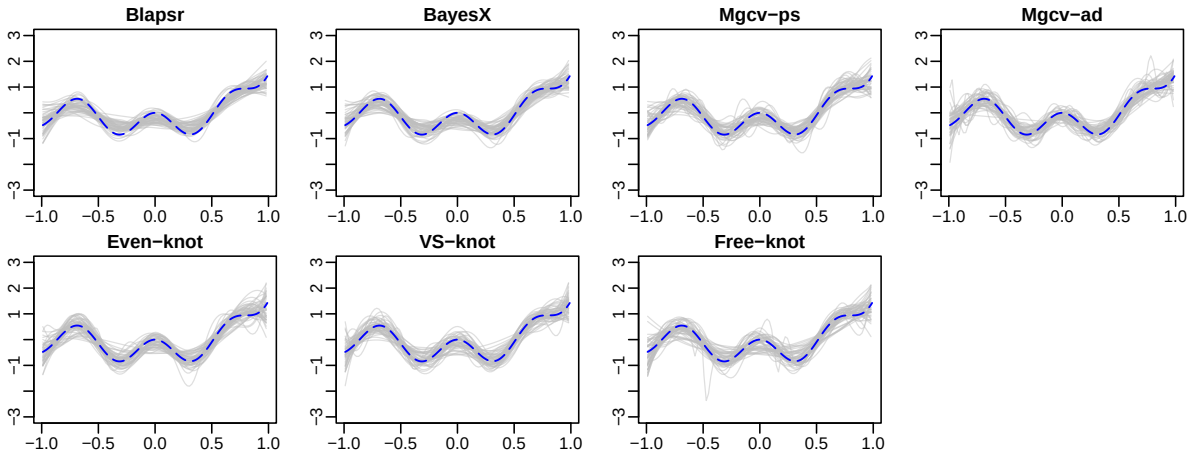
(c) The log RMSE and the coverage probabilities for f_3

Figure 5: The log RMSE and the coverage probabilities for (a) f_1, (b) f_2, and (c) f_3 in the nonparametric logistic regression models with $n = 500, 1000, 2000$.

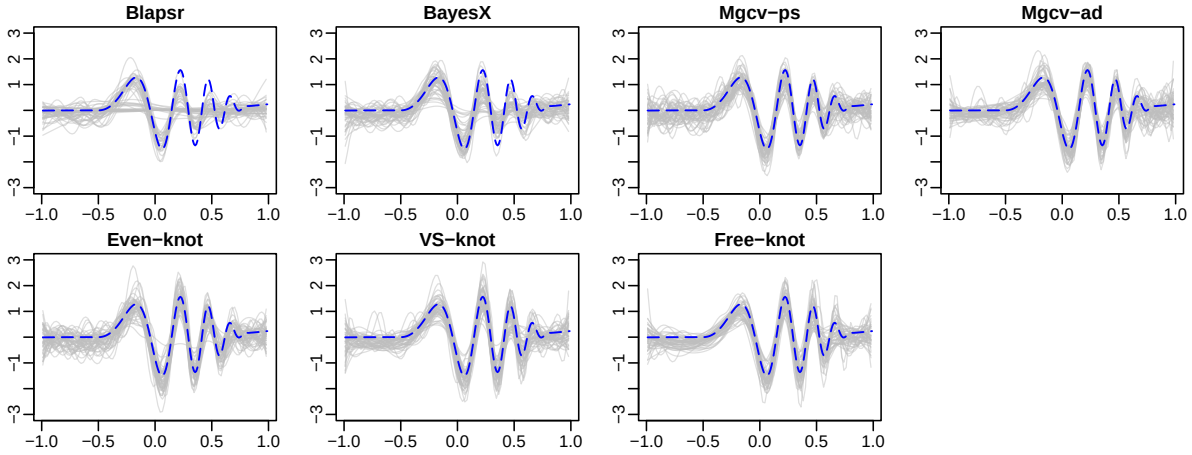
Gaussian hypergeometric distribution (Armero and Bayarri, 1994) for the Gaussian regression model). We thus choose the robust prior as our default prior for a hierarchy of g.

5.2 Comparison with other methods

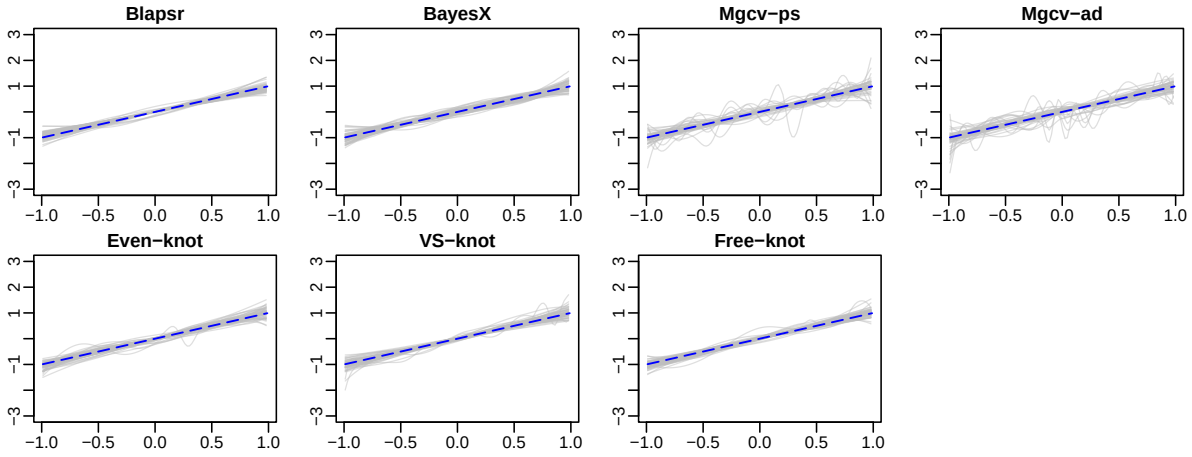
Now we compare the BMS-based methods for GAMs with other Bayesian methods for function estimation. We consider the three strategies described in Section 3: the even-knot splines, the VS-knot splines, and the free-knot splines, together with a few Bayesian competitors available through R packages: `R2BayesX` (Umlauf et al., 2012; Belitz et al., 2022), `Blapsr` (Oswaldo and Philippe, 2020), and `mgcv` (Wood, 2017). The BMS-based approaches are equipped with the robust prior based on the simulation study in Section 5.1. In that all competitors are based on the idea of Bayesian P-splines (Lang and Brezger, 2004), they are intrinsically different from the BMS-based approaches at a philosophical level. The difference between the competitors arises from the way they handle the smoothness parameters. Whereas `R2BayesX` provides standard MCMC estimates, `Blapsr` has an option to utilize the Laplace approximation for a small number



(a) Pointwise posterior mean estimates of f_1 in 50 replications

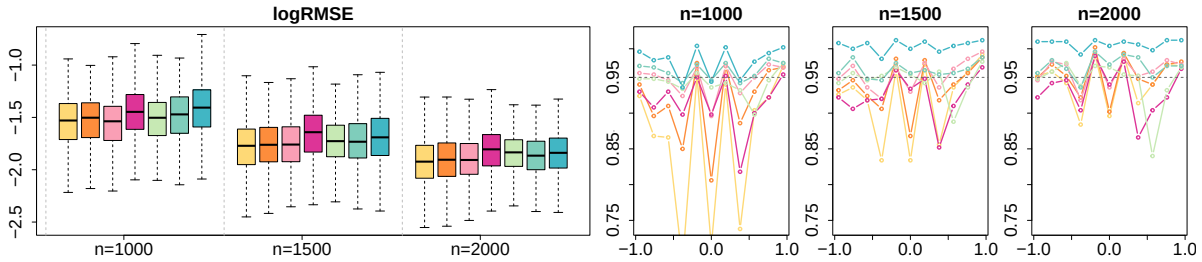


(b) Pointwise posterior mean estimates of f_2 in 50 replications

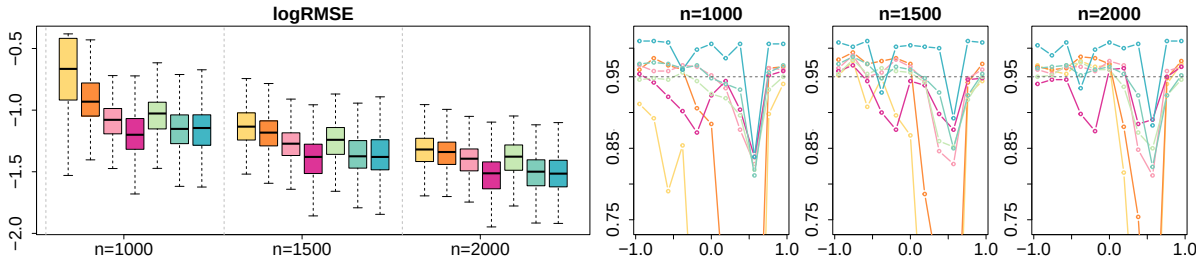


(c) Pointwise posterior mean estimates of f_3 in 50 replications

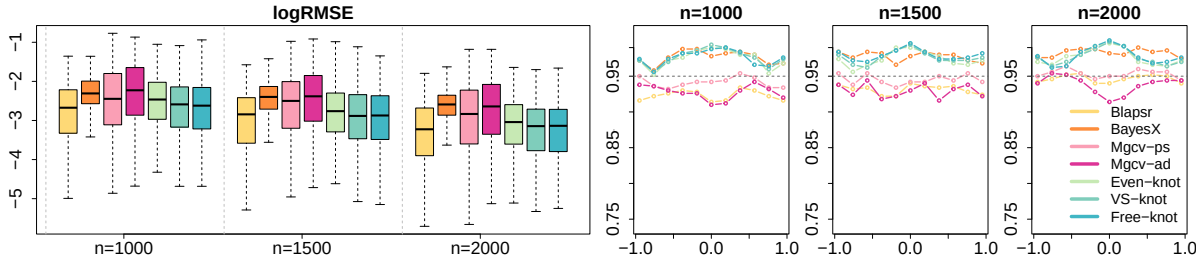
Figure 6: Estimates of (a) f_1, (b) f_2, and (c) f_3 in the nonparametric logistic regression model with $n = 1000$. Pointwise posterior means of randomly chosen 50 replications (gray solid) and the true functions (blue dashed).



(a) The log RMSE and the coverage probabilities for f_1



(b) The log RMSE and the coverage probabilities for f_2



(c) The log RMSE and the coverage probabilities for f_3

Figure 7: The log RMSE and the coverage probabilities for (a) f_1, (b) f_2, and (c) f_3 in the nonparametric logistic regression models with $n = 1000, 1500, 2000$.

of additive components to facilitate the computation. In general, `mgcv` is the fastest and the most computationally efficient, but it is not fully Bayesian as it resorts to the generalized cross-validation in choosing the smoothness parameters. The simulation specification is carefully chosen for a fair comparison among the methods. For the BMS-based approaches (i.e., the even-knot, VS-knot, and free-knot splines), the maximum number of knots M_j is fixed to 30 for every $j = 1, 2, 3$. We also use $M_j = 30$ for the competitors based on the Bayesian P-splines (i.e., `R2BayesX`, `Blapsr`, and `mgcv`), so that both the BMS-based approaches and the Bayesian P-spline approaches have comparable least penalized models. The `mgcv` package provides an option to pursue locally adaptive estimation for smooth functions.

Using the functions in (20), the test datasets are generated by the additive predictor $\eta_i = \sum_{j=1}^{3} f_j^*(x_{ij}) = \alpha + \sum_{j=1}^{3} f_j(x_{ij})$ with x_{ij} drawn independently from Unif$(-1, 1)$, where f_j is the centered version of f_j^* in (20) and α is the induced intercept. Here, 500 replicated datasets are generated for the nonlinear logistic regression model $Y_i \sim \text{Bernoulli}(e^{\eta_i}/(1 + e^{\eta_i}))$ with $n = 1000, 1500, 2000$. We also consider Poisson regression $Y_i \sim \text{Poi}(e^{\eta_i})$ and Gaussian regression

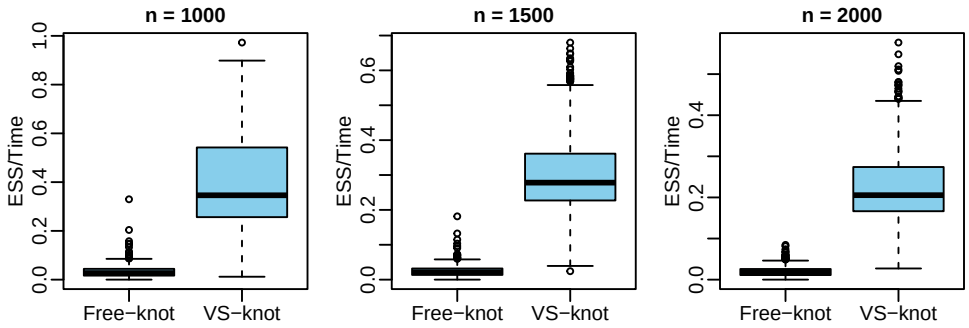


Figure 8: The efficiency ratio, the number of effective sample per one second of CPU runtime, in the nonparametric logistic regression models with $n = 1000, 1500, 2000$ for VS-knot and free-knot.

$Y_i \sim \mathrm{N}(\eta_i, 1)$, whose simulation results are provided in the supplementary material. Similar to Section 5.1, we calculate the RMSE and the 95% pointwise credible bands at a few given points for each method.

Figures 6 and 7 summarize the simulation results for the nonlinear logistic regression models. One can easily see from the figures that `R2BayesX` and `Blapsr` often oversmooth the target functions with excessive penalization. In contrast, `mgcv` provides too wiggly estimates of the linear function, meaning that the method often leads to undersmoothing for simple functions. As seen in Figures 6 and 7, `R2BayesX` and `Blapsr` do not work properly for locally varying smoothness, which is to be expected given that the Bayesian P-splines are not designed for locally adaptive estimation without major modifications (Crainiceanu et al., 2007; Jullion and Lambert, 2007; Scheipl and Kneib, 2009). We observe that `mgcv` with local adaptation performs very well for estimation of locally varying smoothness. However, as can be seen from the performance summaries for f_1 and f_3, the use of `mgcv` for adaptive estimation may lead to higher RMSEs and incorrect coverage probabilities. The major drawback of `mgcv` is that the use of adaptive estimation must be correctly specified in advance for optimal performance, but such a characteristic of the target function is generally unknown to us. The results for Poisson regression and Gaussian regression in the supplementary material also lead to a similar conclusion.

Among the BMS-based approaches, the even-knot splines cannot adapt to locally varying smoothness of f_2 due to its simple construction with equidistant knots. On the other hand, both the VS-knot splines and the free-knot splines correctly detect the local features of f_2. However, we find that the VS-knot splines outperform the free-knot splines in terms of sampling efficiency (measured by the ratio of effective sample size to runtime), as shown in Figure 8, while the empirical performances of the two methods for adaptive estimation are very similar unless the target function is extremely complicated. Compared to `mgcv`, the VS-knot approach does not require the correct determination for the use of adaptive estimation, while outperforming the competitors in most cases. Therefore, we recommend the VS-knot splines as the default option. Even so, it is worth mentioning that the even-knot splines approach is faster than the other BMS-based methods, and even eliminates the need for MCMC when p is reasonably small.

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

Table 2: Description of the variables in the Boston housing dataset.

6 Applications

In this section, we analyze two real-world datasets using the BMS-based methods. The datasets are analyzed by the VS-knot splines approach because it generally outperforms the other methods as shown in Section 5.

6.1 Boston house price data

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

$$\begin{aligned} Y_i = \alpha + \beta_1 chas_i + f_1(crim_i) + f_2(zn_i) + f_3(indus_i) + f_4(nox_i) + f_5(rm_i) + f_6(age_i) \\ + f_7(dis_i) + f_8(rad_i) + f_9(tax_i) + f_{10}(ptratio_i) + f_{11}(black_i) + f_{12}(lstat_i) + \epsilon_i, \\ \epsilon_i \sim N(0, 1/\phi). \end{aligned} \tag{21}$$

The variable $chas$ is binary and assumed to have a linear effect. Each fixed dimensional parameter and nonparametric function is estimated by the VS-knot splines approach. For each nonparametric function, the number of knots M_j is reasonably chosen based on the observed predictor variables. The results of Bayesian inference are summarized in Table 3 and Figure 9.

The results are generally acceptable, but there are also a few variables that are hard to interpret, e.g., nox and rm. We surmise that these are confounded with other unobserved variables or higher order interactions. Among the variables, the effect of dis is particularly interesting; the counties that are very close to the employment centers are highly expensive, but the effect levels off very quickly and is close to zero after some point. Although our main objective is not function

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

Table 3: Summary statistics of the posterior distribution for the model in (21).

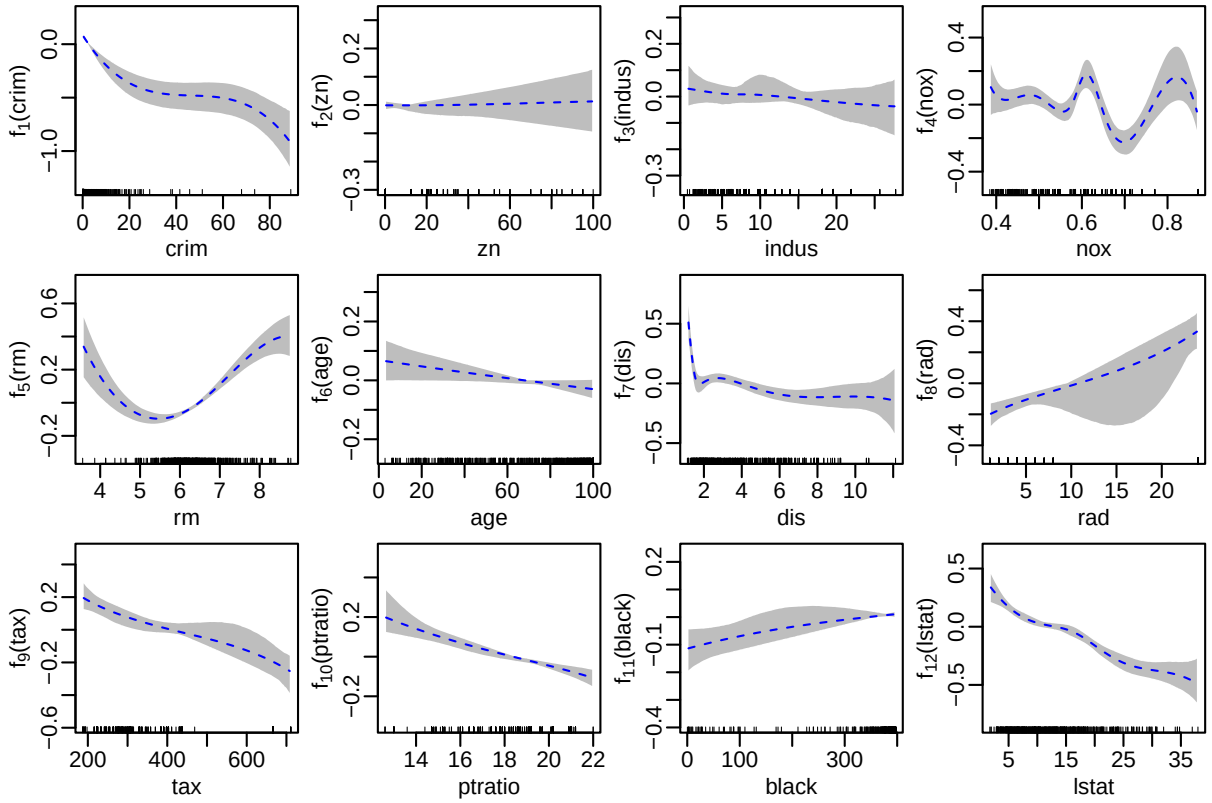


Figure 9: Pointwise posterior mean (blue dashed curve) and pointwise 95% credible band (gray shade) of the functions for the model in (21).

selection, one may be interested in looking at the posterior probability that a variable has a linear effect. This can be fulfilled by examining the marginal posterior distribution of $|\xi_j|$. Table 4 shows the marginal posterior probability of $|\xi_j| = 0$ for each j. The results suggest the possibility of using another model that assumes more variables to have linear effects.

6.2 Pima diabetes data

The Pima diabetes dataset includes signs of diabetes and 7 potential risk factors of $n = 532$ Pima Indian women in Arizona (Smith et al., 1988). The variables are summarized in Table 5. We aim to examine the relationship between the signs of diabetes and other risk factors for Pima Indian women. To model the sign of diabetes (0 or 1) as a response variable Y_i, we consider the following

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

Table 4: Marginal posterior probabilities of linear effects.

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

Table 5: Description of the variables in the Pima diabetes data.

GAM with a logit link,

$$\log \frac{E(Y_i)}{1 - E(Y_i)} = \alpha + f_1(pregnant_i) + f_2(glucose_i) + f_3(pressure_i) + f_4(triceps_i) + f_5(mass_i) + f_6(pedigree_i) + f_7(age_i). \tag{22}$$

The observations with missing values are removed for analysis. Each nonparametric function is estimated by the VS-knot splines. The results are summarized in Table 6 and Figure 10. Many variables have near-linear effects, but a few variables clearly have nonlinear effects, e.g., *mass* and *age*. Similar to Section 6.1, we also provide the posterior probability that each variable has a linear effect in Table 7. The table implies that a few nonparametric functions may be replaced by linear functions for parsimoniousness.

7 Discussion

In this paper, we have reviewed and extended the BMS-based estimation methods for GAMs via Laplace approximation of the integrated marginal likelihood as proposed in Li and Clyde (2018). Furthermore, we recommended a default choice for mixtures of g-priors and provided a brief explanation with a toy example as to how the model selection behaviors of different g-priors, including the conventional unit information prior, could vary. Although the framework of interpreting the Bayes factor of hypothetical two models with equal fit but one with a spurious knot as a penalty function could partially be of help, an overarching theory behind the empirical differences within various g-priors, not only in estimating GAMs but for variable selection for GLMs in general, need further investigation, as it is of primary interest for practitioners to have informed choice for default prior to use.

To the best of our knowledge, Liang et al. (2008) showed that the mixtures of g-priors of $g = \mathcal{O}(n)$ satisfy model selection desiderata of Bayarri et al. (2012) but did not explore into why

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

Table 6: Summary statistics of the posterior distribution for the model in (22).

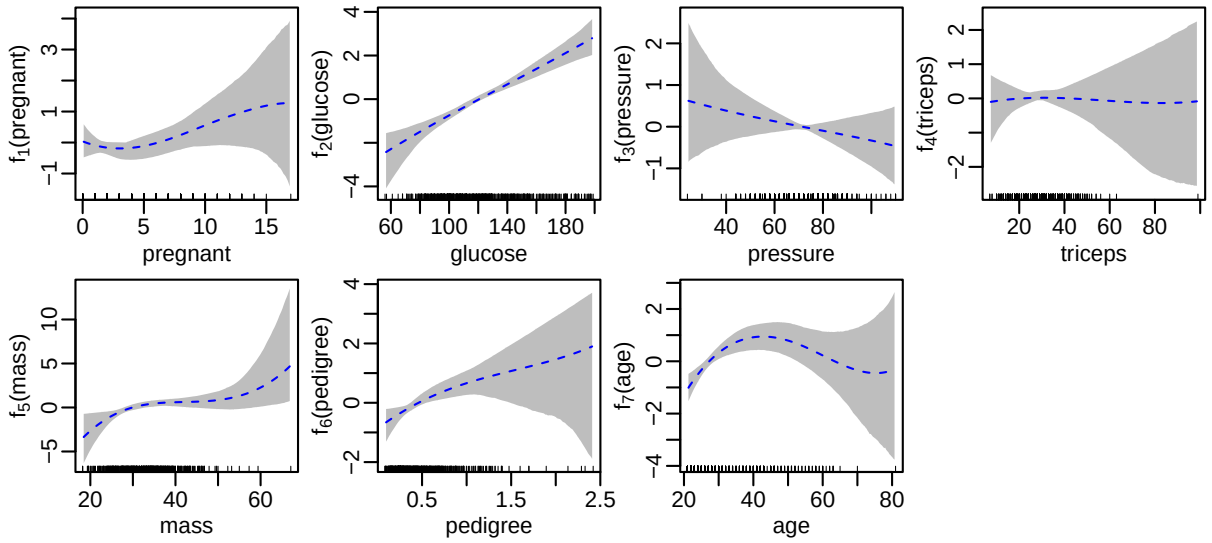


Figure 10: Pointwise posterior mean (blue dashed curve) and pointwise 95% credible band (gray shade) of the functions for the model in (22).

the model selection and estimation performances differ. To this end, one might conduct additional simulations on GLMs with varying conditions including sample size and signal-to-noise ratio, i.e., the magnitude of the true regression coefficient, similar to the absolute scale of the functions in our paper. Our proposed perspective of interpreting the Bayes factor of two hypothetical knots could provide an useful starting point to establish a more general and robust foundation.

A bottleneck in BMS-based algorithms in our paper is evaluating the marginal likelihood of the models as it involves the MLE whose cost grows by the order of $p^3 + np^2$ for $n \times p$ design matrix. A possible remedy could be to use approximate Laplace approximation, as proposed in Rossell et al. (2021), where for a given model Taylor expansion on the regression coefficients is centered at the origin, not at the MLE. For VS-knot, instead of Gibbs sampling for sampling with replacements, one can employ adaptive sampling scheme proposed in Clyde et al. (2011), which is practically to sample without replacements models near the median probability model.

Acknowledgment

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

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

Table 7: Marginal posterior probabilities of linear effects.

Appendix. R Package `gambms`

The R package `gambms` is installed and loaded by running

```
devtools::install_github("hun-learning94/gambms")
library(gambms)
```

The `README.md` in the github repository above illustrates the use of the package on both a simulated data and real data examples. With the R codes provided in `test` directory, one can reproduce the results in this paper.

References

Al-Awadhi, F., Hurn, M., and Jennison, C. (2004). Improving the acceptance rate of reversible jump mcmc proposals. *Statistics & probability letters*, 69(2):189–198.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Antoniadis, A. (1997). Wavelets in statistics: a review. *Journal of the Italian Statistical Society*, 6(2):97–130.

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

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.

Belitz, C., Brezger, A., Kneib, T., Lang, S., and Umlauf, N. (2022). *BayesX: Software for Bayesian Inference in Structured Additive Regression Models*. Version 1.1.

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, 50(4):967–991.

Buhmann, M. D. (2003). *Radial basis functions: theory and implementations*, volume 12. Cambridge university press.

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.

Chen, M.-H. and Ibrahim, J. G. (2003). Conjugate priors for generalized linear models. *Statistica Sinica*, pages 461–476.

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101.

Cox, D. and Snell, E. (1989). *The Analysis of Binary Data*, volume 32. CRC Press.

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.

Cripps, E., Carter, C., and Kohn, R. (2005). Variable selection and covariance selection in multivariate regression models. *Handbook of Statistics*, 25:519–552.

Cui, W. and George, E. I. (2008). Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900.

De Boor, C. (1978). *A practical guide to splines*, volume 27. springer-verlag New York.

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.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36.

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

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

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.

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2018). Power-expected-posterior priors for generalized linear models. *Bayesian Analysis*, 13(3):721–748.

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(LA-UR-20-23587).

Francom, D., Sansó, B., Kupresanin, A., and Johannesson, G. (2018). Sensitivity analysis and emulation for functional data using bayesian adaptive splines. *Statistica Sinica*, pages 791–816.

Gordy, M. B. (1998). A generalization of generalized beta distributions.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

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.

Guisan, A., Edwards Jr, T. C., and Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological modelling*, 157(2-3):89–100.

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.

Gustafson, P. (2000). Bayesian regression modeling with interactions and smooth effects. *Journal of the American Statistical Association*, 95(451):795–806.

Hansen, M. H. and Yu, B. (2003). Minimum description length model selection criteria for generalized linear models. *Lecture Notes-Monograph Series*, pages 145–163.

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

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Sicence*, pages 297–318.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Held, L., Sabanés Bové, D., and Gravestock, I. (2015). Approximate Bayesian model selection with the deviance statistic. *Statistical Science*, pages 242–257.

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

Jeong, S., Park, M., and Park, T. (2017). Analysis of binary longitudinal data with time-varying effects. *Computational Statistics & Data Analysis*, 112:145–153.

Jeong, S. and Park, T. (2016). Bayesian semiparametric inference on functional relationships in linear mixed models. *Bayesian Analysis*, 11(4):1137–1163.

Jeong, S., Park, T., and van Dyk, D. A. (2021). Bayesian model selection in additive partial linear models via locally adaptive splines. *Journal of Computational and Graphical Statistics*, pages 1–13.

Jeong, S. and Rockova, V. (2020). The art of bart: Minimax optimality over nonhomogeneous smoothness in high dimension. *arXiv preprint arXiv:2008.06620*.

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.

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.

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

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.

Katznelson, Y. (2004). *An introduction to harmonic analysis*. Cambridge University Press.

Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11(4):313–322.

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212.

Ley, E. and Steel, M. F. (2012). Mixtures of g-priors for bayesian model averaging with economic applications. *Journal of Econometrics*, 171(2):251–266.

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.

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.

Magee, L. (1990). R^2 measures based on Wald and likelihood ratio joint significance tests. *The American Statistician*, 44(3):250–253.

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

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., and Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, 23(1):249–269.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692.

Nott, D. J. and Kohn, R. (2005). Adaptive sampling for bayesian variable selection. *Biometrika*, 92(4):747–763.

Oswaldo, G. and Philippe, L. (2020). *The blapsr package for fast Bayesian inference in latent Gaussian models by combining Laplace approximations and P-splines.*

Park, T. and Jeong, S. (2018). Analysis of poisson varying-coefficient models with autoregression. *Statistics*, 52(1):34–49.

Rivoirard, V. and Rousseau, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Analysis*, 7(2):311–334.

Ročková, V. and van der Pas, S. (2020). Posterior concentration for bayesian regression trees and forests. *The Annals of Statistics*, 48(4):2108–2131.

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.

Sabanés Bové, D. and Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410.

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.

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.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619.

Shen, W. and Ghosal, S. (2015). Adaptive bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, 42(4):1194–1213.

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*, page 261. American Medical Informatics Association.

Smith, M. and Kohn, R. (1996). Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343.

Sohn, J., Jeong, S., Cho, Y., and Park, T. (2022). Functional clustering methods for binary longitudinal data with temporal heterogeneity. *arXiv preprint arXiv:2210.10273*.

Umlauf, N., Adler, D., Kneib, T., Lang, S., and Zeileis, A. (2012). Structured additive regression models: An r interface to bayesx. Technical report, Working Papers in Economics and Statistics.

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.

Williams, C. and Rasmussen, C. (1995). Gaussian processes for regression. *Advances in neural information processing systems*, 8.

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.

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

Wood, S. N., Goude, Y., and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 64(1):139–155.

Yee, T. W. and Mitchell, N. D. (1991). Generalized additive models in plant ecology. *Journal of vegetation science*, 2(5):587–602.

Yee, T. W. and Wild, C. (1996). Vector generalized additive models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(3):481–493.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de estadística y de investigación operativa*, 31(1):585–603.