Biometrika (2021), **108**, 1, *pp.* 149–166
Printed in Great Britain

doi: 10.1093/biomet/asaa047
Advance Access publication 27 August 2020

An asymptotic and empirical smoothing parameters selection method for smoothing spline ANOVA models in large samples

By XIAOXIAO SUN

Department of Epidemiology and Biostatistics, University of Arizona,
1295 North Martin Avenue, Tucson, Arizona 85724, U.S.A.
xiaosun@email.arizona.edu

WENXUAN ZHONG and PING MA

Department of Statistics, University of Georgia, 310 Herty Drive, Athens,
Georgia 30602, U.S.A.
wenxuan@uga.edu pingma@uga.edu

Summary

Large samples are generated routinely from various sources. Classic statistical models, such as smoothing spline ANOVA models, are not well equipped to analyse such large samples because of high computational costs. In particular, the daunting computational cost of selecting smoothing parameters renders smoothing spline ANOVA models impractical. In this article, we develop an asympirical, i.e., asymptotic and empirical, smoothing parameters selection method for smoothing spline ANOVA models in large samples. The idea of our approach is to use asymptotic analysis to show that the optimal smoothing parameter is a polynomial function of the sample size and an unknown constant. The unknown constant is then estimated through empirical subsample extrapolation. The proposed method significantly reduces the computational burden of selecting smoothing parameters in high-dimensional and large samples. We show that smoothing parameters chosen by the proposed method tend to the optimal smoothing parameters that minimize a specific risk function. In addition, the estimator based on the proposed smoothing parameters achieves the optimal convergence rate. Extensive simulation studies demonstrate the numerical advantage of the proposed method over competing methods in terms of relative efficacy and running time. In an application to molecular dynamics data containing nearly one million observations, the proposed method has the best prediction performance.

Some key words: Asymptotic analysis; Generalized cross-validation; Smoothing parameters selection; Smoothing spline ANOVA model; Subsample.

1. Introduction

In this article, we consider a nonparametric model of the form

$$y_i = \eta(x_i) + \epsilon_i \quad (i = 1, \ldots, n), \tag{1}$$

where $y_i \in \mathbb{R}$ is the response variable for the ith observation, η is a nonparametric function varying in an infinite-dimensional functional space, $x_i = (x_{i\langle 1\rangle}, \ldots, x_{i\langle d\rangle})^{\mathrm{T}}$ is a d-dimensional vector of predictors for the ith observation, and the ϵ_i are independent and identically distributed random errors with mean zero and unknown variance σ^2. We focus on the smoothing spline ANOVA model (Wahba et al., 1995) for a multi-dimensional problem, i.e., with $d > 1$. In the smoothing spline ANOVA model we decompose the function η as

$$\eta(x) = \eta_{\o} + \sum_{j=1}^{d} \eta_j(x_{\langle j \rangle}) + \sum_{j<k} \eta_{j,k}(x_{\langle j \rangle}, x_{\langle k \rangle}) + \cdots + \eta_{1,2,\ldots,d}(x_{\langle 1 \rangle}, x_{\langle 2 \rangle}, \ldots, x_{\langle d \rangle}), \qquad (2)$$

where $\eta_{\o}$ is a constant, the η_j are main-effect functions, the $\eta_{j,k}$ are two-way interaction functions, and $\eta_{1,2,\ldots,d}(x_{\langle 1 \rangle}, x_{\langle 2 \rangle}, \ldots, x_{\langle d \rangle})$ is a d-way interaction function. Side conditions on the components are imposed to guarantee a unique decomposition. The nonparametric function η can be estimated by minimizing the penalized least squares

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - \eta(x_i)\}^2 + \lambda P(\eta), \qquad (3)$$

where $P(\eta) = P(\eta, \eta)$ is a quadratic roughness penalty and the smoothing parameter λ controls the trade-off between the lack of fit of η and the roughness of η. Extra smoothing parameters are included in $P(\eta, \eta)$ to adjust the strength of the components in (2), but for simplicity we omit them from the notation. The explicit formula for $P(\eta, \eta)$ can be found in § 2.1. Since the minimizer of (3), denoted by $\eta_{n,\lambda}$, is sensitive to the selection of λ, it is crucial to choose an effective and efficient method for selecting the smoothing parameter.

Numerous computational methods for smoothing parameter selection have been proposed. One of the earliest is the C_L method (Mallows, 1973). To circumvent the impracticality of the C_L method due to its dependence on an unknown σ^2, Craven & Wahba (1978) proposed generalized cross-validation. They showed that the smoothing parameter estimated by generalized cross-validation minimizes a specific risk function asymptotically. Although their method gives a good estimate of λ without prior knowledge of the variance σ^2, it occasionally has an undersmoothing problem. To overcome this problem, Kim & Gu (2004) developed a modified version of generalized cross-validation by adding a fudge factor. Under the Bayes framework, Wahba (1985) proposed a maximum likelihood estimate for the smoothing parameter. Extensive simulations were performed to demonstrate that the maximum likelihood estimate provides satisfactory estimates. Nonetheless, the minimizer $\eta_{n,\lambda}$ based on the smoothing parameter chosen by maximum likelihood cannot be guaranteed to attain the optimal convergence rate. In contrast to the above methods, the improved Akaike information criterion proposed by Hurvich et al. (1998) aims to avoid the undersmoothing problem of generalized cross-validation. However, the empirical performance of Hurvich et al.'s criterion is not as good as that of other criteria, such as generalized cross-validation, in some situations (Aydin et al., 2013). Moreover, its soundness is hard to justify owing to the lack of theoretical analysis under the smoothing spline ANOVA framework. A more recent line of work for large datasets is the divide-and-recombine method (Shang & Cheng, 2017; Xu & Wang, 2018). In this approach, a large dataset is divided into small subsets to which smoothing spline models are fitted, and the outputs of these models are then recombined. Since the smoothing spline is applied to small subsets, selecting the smoothing parameter is computationally feasible.

For multivariate η, multiple smoothing parameters are used to adjust the strength of the corresponding components in (2). Gu & Wahba (1991) proposed to select multiple smoothing parameters by minimizing the generalized cross-validation function through a modified Newton method. With all smoothing parameters being tunable, the iterative algorithm takes $O(Sn^3)$ flops per iteration, where S is the number of smoothing parameters, and needs tens of iterations to converge. The algorithm is quite efficient when S is small. As the number of multi-way interaction components in (2) increases, the number of smoothing parameters grows dramatically. For example, S is 5 for the full two-way model and 19 for the full three-way model. Thus, the algorithm is computationally expensive for multi-dimensional models with interaction terms. Several methods have been proposed to alleviate the heavy computational burden of these models. An obvious option is to provide good prespecified values for multiple smoothing parameters. Gu & Wahba (1991) proposed an algorithm for calculating these values, and showed that the minimizer of (3) based on them usually yields good estimates. Although the algorithm performs well in additive models, it is unreliable when interaction components are present. The unreliable performance may be exacerbated when the model is misspecified. Helwig & Ma (2015) proposed a reparameterization of smoothing parameters in the smoothing spline ANOVA model. For the reparameterization, there is one smoothing parameter for each predictor and the smoothing parameter for an interaction term is the product of the smoothing parameters of the corresponding predictors. This new algorithm has a computational cost comparable to that of generalized additive models (Hastie & Tibshirani, 1986). Nevertheless, the algorithm may produce a biased estimate when the smoothing spline ANOVA model is misspecified. In addition, its theoretical foundation requires further justification.

Parallel to the work under the smoothing spline ANOVA framework, several authors have proposed efficient smoothing parameter selection methods for generalized additive models. For univariate functions, many attempts have been made to estimate the smoothness of functions (Buja et al., 1989; Marx & Eilers, 1998). These algorithms are fast even for large datasets. For multivariate functions, low-rank tensor product methods were developed (Wood, 2006; Lee & Durbán, 2011). To control the smoothness on different predictors within an interaction term, multiple smoothing parameters are associated with the smoothing penalties corresponding to the interaction. For example, for any bivariate interaction $\eta_{j,\,k}(x_{\langle j\rangle}, x_{\langle k\rangle})$ there are two smoothing parameters for controlling the smoothness on predictors $x_{\langle j\rangle}$ and $x_{\langle k\rangle}$, whereas three smoothing parameters are used under the smoothing spline ANOVA framework to adjust the smoothness on $x_{\langle j\rangle}$, $x_{\langle k\rangle}$, and the interaction of these two predictors separately. The low-rank tensor product methods reduce the number of smoothing parameters and improve the computational efficiency. However, when the bivariate function $\eta_{j,\,k}(x_{\langle j\rangle}, x_{\langle k\rangle})$ is not an additive function with respect to the $x_{\langle j\rangle}$ and $x_{\langle k\rangle}$ directions, the smoothing spline ANOVA models may have numerical advantages since they can model the interaction of these two predictors. Recently, some extensions of the multivariate smoothing approach in generalized additive models have been proposed to estimate the smooth functions (Ruppert et al., 2003; Wand, 2003; Lee et al., 2013; Wood et al., 2013; Rodríguez-Álvarez et al., 2015; Wood & Fasiolo, 2017). Wood et al. (2017) developed an efficient fitting method to estimate generalized additive models in large samples. In particular, a reparameterization is implemented in the fitting iteration, where the smoothing matrix can be computed blockwise. Moreover, instead of fully optimizing the restricted marginal likelihood at each iteration, a single-step Newton update is utilized. To reduce the memory usage for large matrices, a novel covariate discretization scheme is also implemented. While this discretization scheme significantly reduces the computational time of estimation, a rigorous theoretical investigation is still lacking.

The asymptotic behaviour of $\eta_{n,\lambda}$ and the optimal λ has been studied extensively; see Silverman (1982), Rice & Rosenblatt (1983), Cox (1984), Speckman (1985), Cox & O'Sullivan (1990) and Gu & Qiu (1993). The estimator can achieve an optimal convergence rate when the smoothing parameter is of order $O\{n^{-r/(pr+1)}\}$ for $r > 1$ and $p \in [1,2]$. Lin (2000) further studied the optimal convergence rate of the estimator in tensor product space ANOVA models, and showed that the optimal rate of smoothing parameters depends on the highest order of interactions in (2). One can directly use $Cn^{-r/(pr+1)}$ with some predefined C, r and p as the smoothing parameter when fitting the model to a sample of size n (Hall, 1990). This method is referred to as the order-based method. However, the numerical performance of the order-based method is unreliable, which is also observed in our simulation studies.

To make the selection of smoothing parameters practical in large samples, we develop an asympirical, i.e., asymptotic and empirical, smoothing parameters selection approach by combining the theoretical properties of smoothing parameters and the aforementioned computational methods in a synergistic manner. In the proposed method, we choose a subsample of size much smaller than the full sample size n, and we select smoothing parameters for the subsample using the generalized cross-validation method. The smoothing parameters for the full sample are extrapolated based on the selected smoothing parameters and the optimal rate $O\{n^{-r/(pr+1)}\}$. The proposed smoothing parameters selection method reduces the computational complexity from tens of $O(Sn^3)$ flops, as required by the generalized cross-validation method, to $O(B^3)$, where B is the size of the subsamples. The numerical advantage of the proposed algorithm over the other approaches is significant when there are multiple interaction components in the model, as demonstrated through our extensive simulation studies and real data examples. Besides the numerical advantages, the smoothing parameters obtained using our approach share optimal properties with parameters that minimize a specific risk function for full samples. Furthermore, the estimator based on the proposed smoothing parameters attains the optimal convergence rate.

2. Smoothing spline ANOVA models

2.1. *Estimation*

We review the Kimeldorf–Wahba representer theorem (Kimeldorf & Wahba, 1971; Wahba, 1990; Wang, 2011), which states that the solution of penalized least squares defined in an infinite-dimensional functional space actually resides in a finite-dimensional space. Recall that the minimization of (3) is performed in the tensor product reproducing kernel Hilbert space $\mathcal{H} = \{\eta : P(\eta,\eta) < \infty\}$ with the quadratic roughness penalty $P(\eta,\eta) = \sum_{\delta=1}^{S} \theta_\delta^{-1}(\eta,\eta)_\delta$, where the θ_δ are smoothing parameters that adjust the strengths of the corresponding components, $(\cdot\,,\cdot)_\delta$ is the inner product in $\mathcal{H}_\delta$ with reproducing kernel $R_\delta(\cdot\,,\cdot)$, and S is the number of subspaces based on the tensor product decomposition. The space $\mathcal{H}$ has the tensor sum decomposition $\mathcal{H} = \mathcal{N}_P \oplus \mathcal{H}_P$ where $\mathcal{N}_P$, the null space of $\mathcal{H}$, is spanned by $\{\phi_\nu\}_{\nu=1}^{M}$ and $\mathcal{H}_P = \bigoplus_{\delta=1}^{S} \mathcal{H}_\delta$ has the reproducing kernel $R(\cdot\,,\cdot) = \sum_{\delta=1}^{S} \theta_\delta R_\delta(\cdot\,,\cdot)$.

Theorem 1 (Kimeldorf–Wahba representer theorem). *The minimizer of* (3) *is*

$$\eta(x) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} c_i R(x_i, x),$$

where $d = (d_1, \ldots, d_M)^{\mathrm{T}}$ and $c = (c_1, \ldots, c_n)^{\mathrm{T}}$ are unknown coefficients.

Theorem 1 facilitates the estimation by reducing an infinite-dimensional optimization problem to a finite-dimensional one. Based on the representer theorem, the minimization in (3) becomes

$$(Y - Td - Kc)^{\mathrm{T}}(Y - Td - Kc) + n\lambda c^{\mathrm{T}}Kc, \tag{4}$$

where $Y = (y_1, \ldots, y_n)^{\mathrm{T}}$, $T_{n\times M}$ is a matrix with (i,ν)th entry $\phi_\nu(x_i)$, and $K_{n\times n}$ is a matrix with (i,j)th entry $R(x_i, x_j)$. By differentiating (4) with respect to d and c, and setting the derivatives to zero, one obtains the linear system of equations

$$\begin{pmatrix} T^{\mathrm{T}}T & T^{\mathrm{T}}K \\ K^{\mathrm{T}}T & K^{\mathrm{T}}K + n\lambda K \end{pmatrix} \begin{pmatrix} d \\ c \end{pmatrix} = \begin{pmatrix} T^{\mathrm{T}}Y \\ K^{\mathrm{T}}Y \end{pmatrix}. \tag{5}$$

To estimate d and c, one needs to solve the linear system (5). If the smoothing parameters λ and θ_δ are known, the computational cost is typically $O(n^3)$.

2.2. *Roughness penalties*

One can choose different forms of roughness penalties for the estimation. The most popular choice for univariate η on a compact interval $\mathcal{X}$ is

$$P(\eta, \eta) = \int_{\mathcal{X}} (\eta^{(m)})^2 \, \mathrm{d}x,$$

where $\eta^{(m)} = \mathrm{d}^m\eta/\mathrm{d}x^m$. Setting $m = 2$, a cubic spline estimator is obtained by minimizing (3) (Wahba, 1990). One convenient way to define the penalty for multivariate functions that have the form (2) is to construct the tensor product reproducing kernel Hilbert space. The reproducing kernel Hilbert space $\mathcal{H}$ can be decomposed into the space of constants, the spaces of main effects, and the corresponding spaces of interaction terms lying in the tensor product space of the interacting main-effect spaces.

Example 1. For the tensor product cubic spline on $[0, 1]^2$, one has the following space decomposition in each variable:

$$\begin{aligned} \{f : f^{(2)} \in \mathcal{L}_2[0,1]\} &= \{f : f \propto 1\} \oplus \{f : f \propto k_1\} \\ &\quad \oplus \left\{ f : \int_0^1 f \, \mathrm{d}x = \int_0^1 f^{(1)} \, \mathrm{d}x = 0,\ f^{(2)} \in \mathcal{L}_2[0,1] \right\} \\ &= \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1, \end{aligned}$$

where $k_1(x) = x - 0.5$. The space of constant terms is $\mathcal{H}_{00\langle 1\rangle} \otimes \mathcal{H}_{00\langle 2\rangle}$; $\mathcal{H}_{00\langle 1\rangle} \otimes (\mathcal{H}_{01\langle 2\rangle} \oplus \mathcal{H}_{1\langle 2\rangle})$ and $\mathcal{H}_{00\langle 2\rangle} \otimes (\mathcal{H}_{01\langle 1\rangle} \oplus \mathcal{H}_{1\langle 1\rangle})$ span the space of main effects; and the subspace $(\mathcal{H}_{01\langle 1\rangle} \oplus \mathcal{H}_{1\langle 1\rangle}) \otimes (\mathcal{H}_{01\langle 2\rangle} \oplus \mathcal{H}_{1\langle 2\rangle})$ spans the space of interactions. Let $\mathcal{H}_{\nu,\,\mu} = \mathcal{H}_{\nu\langle 1\rangle} \otimes \mathcal{H}_{\mu\langle 2\rangle}$ for $\nu, \mu = 00, 01, 1$, with inner products $(\eta, \eta)_{\nu,\mu}$ and reproducing kernels $R_{\nu,\,\mu} = R_{\nu\langle 1\rangle}R_{\mu\langle 2\rangle}$; see Gu (2013, Theorem 2.6). One may set

$$\begin{aligned} P(\eta, \eta) = \theta_{1,00}^{-1}(\eta, \eta)_{1,00} &+ \theta_{00,1}^{-1}(\eta, \eta)_{00,1} \\ &+ \theta_{1,01}^{-1}(\eta, \eta)_{1,01} + \theta_{01,1}^{-1}(\eta, \eta)_{01,1} + \theta_{1,1}^{-1}(\eta, \eta)_{1,1}. \end{aligned}$$

The null space of $P(\eta, \eta)$ is

$$\mathcal{N}_P = \mathcal{H}_{00,00} \oplus \mathcal{H}_{01,00} \oplus \mathcal{H}_{00,01} \oplus \mathcal{H}_{01,01}.$$

As in Example 1, a two-dimensional η can be decomposed into four main terms: one constant term, two main effect terms, and one two-way interaction term. There are five effective smoothing parameters, namely $\lambda/\theta_{1,00}, \lambda/\theta_{00,1}, \lambda/\theta_{1,01}, \lambda/\theta_{01,1}$ and $\lambda/\theta_{1,1}$. Two of them, $\lambda/\theta_{1,00}$ and $\lambda/\theta_{00,1}$, are for the main effects, and the rest are for the interaction effects.

Example 2. For the tensor product cubic spline on $\{1,\ldots,K\} \times [0,1]$, one can use the kernels $R_{0\langle 1\rangle}(x_{\langle 1\rangle}, \grave{x}_{\langle 1\rangle}) = 1/K$ and $R_{1\langle 1\rangle}(x_{\langle 1\rangle}, \grave{x}_{\langle 1\rangle}) = I_{(x_{\langle 1\rangle}=\grave{x}_{\langle 1\rangle})} - 1/K$ on $\{1,\ldots,K\}$ and the kernels $R_{00\langle 2\rangle}(x_{\langle 2\rangle}, \grave{x}_{\langle 2\rangle}) = 1$, $R_{01\langle 2\rangle}(x_{\langle 2\rangle}, \grave{x}_{\langle 2\rangle}) = k_1(x_{\langle 2\rangle})k_1(\grave{x}_{\langle 2\rangle})$ and $R_{1\langle 2\rangle}(x_{\langle 2\rangle}, \grave{x}_{\langle 2\rangle}) = k_2(x_{\langle 2\rangle})k_2(\grave{x}_{\langle 2\rangle}) - k_4(x_{\langle 2\rangle} - \grave{x}_{\langle 2\rangle})$ on $[0,1]$, where the k_u $(u = 1, 2, 4)$ are scaled Bernoulli polynomials. The tensor product space can be constructed in an analogous way to Example 1.

2.3. *Generalized cross-validation*

When estimating multivariate functions in a tensor product space, multiple smoothing parameters are involved; see Example 1. The multiple smoothing parameters λ/θ control the trade-off between the lack of fit of η and the roughness of η, where $\theta = (\theta_1, \ldots, \theta_S)^{\mathrm{T}}$. Gu & Wahba (1991) proposed a modified Newton method for minimizing the generalized cross-validation score,

$$G(\lambda/\theta) = \frac{n^{-1}Y^{\mathrm{T}}\{I - A(\lambda/\theta)\}^2 Y}{[n^{-1}\,\mathrm{tr}\{I - A(\lambda/\theta)\}]^2},$$

iteratively for multiple smoothing parameters, where the smoothing matrix $A(\lambda/\theta)$ is given in the Supplementary Material. In particular, the method consists of the following steps: (i) for fixed θ, minimize the generalized cross-validation score with respect to $n\lambda$; (ii) update θ based on current information on $n\lambda$.

With all smoothing parameters being tunable, the above iterative algorithm takes $O(Sn^3)$ flops per iteration and needs tens of iterations to converge. The number of smoothing parameters, S, increases dramatically as the number of multi-way interactions grows. In particular, $S = d + 3d(d-1)/2$ for the two-way interaction model which truncates the decomposition in (2) at two-way interactions, so it is impractical to apply smoothing spline ANOVA models to large samples. Even for the additive model with d smoothing parameters being tunable, tens of iterations of $O(n^3)$ flops become infeasible in large samples. Since the iterative algorithm depends heavily on the starting values, Gu & Wahba (1991) proposed an algorithm for calculating good starting values of θ. The software developed by Gu (2014) uses these starting values as the final estimate of θ, and the algorithm is called the skip algorithm. With the aid of the skip algorithm, the multiple smoothing parameters selection problem reduces to the single smoothing parameter selection problem, which takes $O(n^3)$ flops. The skip algorithm comprises two steps: (i) for $\theta_\delta = \{\mathrm{tr}(R_\delta)\}^{-1}$, minimize the generalized cross-validation score with respect to $n\lambda$, and calculate c; (ii) estimate the starting values $\theta_{\delta 0} = \theta_\delta^2 c^{\mathrm{T}} R_\delta c$.

3. Asympirical smoothing parameters selection

3.1. *The optimal smoothing parameter*

We review the optimal smoothing parameters selection method, which motivates the proposed algorithm. The optimality of smoothing parameter selection can be characterized by minimization of the expectation of the loss function, $E\{L(\lambda)\}$, i.e., the risk function, where the loss function is

$$L(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{\eta_{n,\lambda}(x_i) - \eta(x_i)\}^2. \tag{6}$$

Wahba (1975) derived the optimal smoothing parameter by minimizing the risk function for smoothing periodic splines in $\mathcal{H}^{(m)}$, defined by

$$\mathcal{H}^{(m)} = \Big\{ f : f^{(\nu)} \text{ absolutely continuous for } \nu = 0, 1, \ldots, m-1,\ f^{(m)} \in \mathcal{L}_2[0, 1],$$
$$f^{(\nu)}(0) - f^{(\nu)}(1) = 0 \text{ for } \nu = 0, 1, \ldots, m-1 \Big\}.$$

Suppose that $\eta \in \mathcal{H}^{(2m)}$, i.e., η is very smooth, and $\|\eta^{(2m)}\| \dotplus 0$, where $\| \cdot \|$ is the $\mathcal{L}_2$-norm. The optimal choice of smoothing parameter, ignoring $o(1)$ terms, is

$$\left\{ \frac{\tilde{k}_m}{4m} \frac{\sigma^2}{\|\eta^{(2m)}\|^2} \right\}^{2m/(4m+1)} n^{-2m/(4m+1)}, \tag{7}$$

where $\tilde{k}_m = (1/\pi) \int_0^\infty 1/(1 + t^{2m})^2 \,\mathrm{d}t$ is a constant depending on m. We rewrite the smoothing parameter in (7) as $Cn^{-2m/(4m+1)}$, since the first term is a constant unrelated to the full sample size n. Likewise, in the subsample of size $b \to \infty$, the asymptotically optimal smoothing parameter $\lambda_{\text{RISK}}(b)$ is $Cb^{-2m/(4m+1)}$ for the same C. If we can estimate C in a subsample of size b, then the smoothing parameter $\lambda_{\text{RISK}}(b)(n/b)^{-2m/(4m+1)}$ for the full sample of size n is thereby estimated. Under different smoothness conditions, to be defined later, the optimal smoothing parameter that minimizes the risk function has the form $Cb^{-r/(pr+1)}$ for $r > 1$ and $p \in [1, 2]$ in a subsample of size b (Wahba, 1977, 1985). For instance, we have $r = 2m$ and $p = 2$ for the above smoothing periodic spline case. Based on the same rationale, the smoothing parameter for the full sample is

$$\lambda_{\text{RISK}}(b)(n/b)^{-r/(pr+1)}. \tag{8}$$

3.2. *The asympirical algorithm*

It is infeasible to choose the optimal smoothing parameter if the true η and σ^2 are unknown. Therefore, we replace the optimal smoothing parameter $\lambda_{\text{RISK}}(b)$ in (8) with $\lambda_{\text{GCV}}(b)$ chosen by the generalized cross-validation method in a subsample of size b. The detailed procedure is outlined in Algorithm 1.

Algorithm 1. The asympirical smoothing parameters selection algorithm.

Step 1. Take a random subsample of size b from the original data, and apply the generalized cross-validation method to the subsample to estimate the smoothing parameters $\lambda_{\text{GCV}}(b)$ and $\theta_{\text{GCV}}(b)$.

Step 2. Set smoothing parameters $\lambda_{\text{ASP}}(n; b) = \lambda_{\text{GCV}}(b)(n/b)^{-r/(pr+1)}$ and $\theta_{\text{ASP}}(n; b) = \theta_{\text{GCV}}(b)$ to find the minimizer of (3) for the full sample of size n.

In the first step, the random subsample is selected using uniform sampling. More delicate sampling approaches can be found in Ma et al. (2015) and Meng et al. (2020). To make the estimated smoothing parameters more stable, we usually take multiple subsamples and choose the median of a group of smoothing parameters. In the algorithm, we assume that optimal smoothing parameters share the same rate of decrease as n increases (Gu & Wahba, 1991). Since smoothing parameters θ are used to adjust the roughness penalties imposed on different components, see Example 1, we calculate the optimal $\theta_{\text{GCV}}(b)$ for the subsample and perform the minimization based on the estimated $\theta_{\text{GCV}}(b)$ for the full sample. Further details on how to choose b, r and p in practice are given in § 4.

4. Theoretical analysis

This section presents the theoretical analysis of the smoothing parameters selected by Algorithm 1. The selected smoothing parameters tend to the parameter values that minimize the risk function. Our theoretical analysis also provides a guide to choosing b, r and p. We then present results on convergence rates of the estimator based on the proposed smoothing parameters. For simplicity, we suppress λ's dependence on θ and only make λ explicit. All proofs are given in the Supplementary Material.

Let the subsample size be b. The matrix $I - A(\lambda)$ for the smoothing spline ANOVA model has the representation

$$I - A(\lambda) = b\lambda Z(D + b\lambda I)^{-1}Z^{\mathrm{T}},$$

where the matrix Z satisfies $Z^{\mathrm{T}}Z = I_{(b-M)\times(b-M)}$ and D_{b-M} is a $(b-M)\times(b-M)$ diagonal matrix with real-valued entries $\zeta_{\nu b} > 0$; more details are given in the Supplementary Material. We derive theoretical results under the following smoothness assumption.

Assumption 1. The function $\eta \in \mathcal{H}_p$, where the space $\mathcal{H}_p$ is defined as

$$\mathcal{H}_p = \left\{\eta : P(\eta,\eta) > 0,\ \sum_{\nu=1}^{b-M} \frac{h_{\nu b}^2/b}{(\zeta_{\nu b}/b)^p} \leqslant J_p + J_p\, o(1)\right\}.$$

Here the real-valued vector $(h_{1,b},\ldots,h_{b-M,\,b})^{\mathrm{T}} = Z^{\mathrm{T}}H$ with $H = \{\eta(x_1),\ldots,\eta(x_b)\}^{\mathrm{T}}$, J_p for $p \in [1,2]$ is a real-valued constant independent of the subsample size b, and $o(1) \to 0$ as $b \to \infty$.

Under Assumption 1, we only consider the case where $P(\eta,\eta) > 0$. When $P(\eta,\eta) = 0$, both the risk function and the generalized cross-validation function are minimized for $\lambda = \infty$ (Craven & Wahba, 1978).

Theorem 2. *Suppose that Assumption 1 holds for some $p \in [1,2]$. Let $r > 1$, let $\lambda_{\text{GCV}}(b)$ be the smoothing parameter chosen by the generalized cross-validation method for the subsample of size b, let $\lambda_{\text{RISK}}(n)$ be the optimal smoothing parameter minimizing the risk function for the full sample of size n, and let $\lambda_{\text{ASP}}(n;b)$ be the proposed smoothing parameter for the full sample of size n. Suppose that $\lambda_{\text{GCV}}(b) \to 0$ and $b\lambda_{\text{GCV}}^{1/r}(b) \to \infty$. Then $\lambda_{\text{ASP}}(n;b) = \lambda_{\text{RISK}}(n)\{1+o(1)\}$.*

Theorem 2 shows that the proposed smoothing parameter $\lambda_{\text{ASP}}(n;b)$ is an estimate of the minimizer of $E\{L(\cdot)\}$ asymptotically. We have the following immediate corollary under regularity conditions stated in the Supplementary Material.

Corollary 1. *Under regularity conditions in the Supplementary Material, as $\lambda_{\text{GCV}}(b) \to 0$, $b\lambda_{\text{GCV}}^{1/r}(b) \to \infty$ and $n \to \infty$, we have*

$$\frac{E[L\{\lambda_{\text{ASP}}(n;b)\}]}{E[L\{\lambda_{\text{RISK}}(n)\}]} = 1 + o(1).$$

This corollary gives the expectation inefficiency of $\lambda_{\text{ASP}}(n;b)$ relative to $\lambda_{\text{RISK}}(n)$ as the number of observations n tends to infinity.

In Theorem 2 one needs $b\lambda_{\text{GCV}}^{1/r}(b) \to \infty$. We further assume that $\lambda_{\text{GCV}}(b)$ achieves the optimal rate $n^{-r/(pr+1)}$, and it suffices to have $b \asymp n^{1/(pr+1)+\varepsilon}$ for any $\varepsilon > 0$. For $P(\eta,\eta) = \int_0^1 (\eta^{(2)})^2\,\mathrm{d}x$ on $[0,1]$, $r = 4$, and we have $p = 1$ when $\eta^{(2)}$ is square-integrable and $p = 2$ when $\eta^{(4)}$ is square-integrable. For the tensor product cubic spline, r is typically less than 4 (Wahba, 1990; Lin, 2000), so we set $r = 3$ empirically. Taking these facts into consideration, we set $r = 3$, $p = 1$ and $\varepsilon = 0$, and use $b \propto n^{1/4}$ empirically. In real applications, the subsample size b is set to $50n^{1/4}$. The smoothness of η is indexed by p, which is estimated empirically. We first take a random subsample of size B, and minimize the generalized cross-validation score with respect to $p \in \{1,2\}$ by replacing λ in the score with $\lambda_{\text{GCV}}(b)(B/b)^{-r/(pr+1)}$. We take $B = 2b$ in our simulation studies and real data examples. Thus the computational complexity of the proposed algorithm is of order $O(B^3)$. To reduce the computational burden of fitting smoothing spline ANOVA models for large samples, one can implement the fast algorithm proposed by Kim & Gu (2004). In the algorithm, one first randomly selects $\breve{q}$ basis functions from n and then estimates the minimizer of (3). The algorithm requires $O(n\breve{q}^2)$ flops to estimate the minimizer for each choice of smoothing parameters. Thus, the corresponding computational complexities of the generalized cross-validation method and the proposed method are also reduced. The complexity of the proposed method is of order $O(B\breve{q}^2)$ when the fast algorithm is applied.

We now show the convergence rate of the estimator that relies on the proposed smoothing parameters. To study theoretical properties of smoothing spline ANOVA models, one needs the quadratic functional V defined by

$$V(\eta_{n,\lambda} - \eta,\ \eta_{n,\lambda} - \eta) = \int_{\mathcal{X}} \{\eta_{n,\lambda}(x) - \eta(x)\}^2 f(x)\,\mathrm{d}x,$$

where $f(\cdot)$ is the marginal density of x. The functional represents the mean squared error of the estimator $\eta_{n,\lambda}$ in estimating the function η on a compact domain $\mathcal{X} \subset \mathbb{R}^d$. To avoid interpolation, the regularization λP needs to restrict the estimate to an effective model space. To control the bias, the effective model space needs to be increased by letting $\lambda \to 0$ as the sample size $n \to \infty$. It was shown in Gu (2013, Ch. 9) that $(V + \lambda P)(\eta_{n,\lambda} - \eta, \eta_{n,\lambda} - \eta) = O(n^{-1}\lambda^{-1/r} + \lambda^p)$. We establish the following theorem under regularity conditions described in the Supplementary Material.

THEOREM 3. *Under the regularity conditions in the Supplementary Material and for some* $p \in [1,2]$ *and* $r > 1$, *as* $\lambda_{\text{RISK}}(n) \to 0$ *and* $n\lambda_{\text{RISK}}^{2/r}(n) \to \infty$, *we have*

$$\{V + \lambda_{\text{RISK}}(n)P\}(\eta_{n,\lambda_{\text{ASP}}(n;b)} - \eta_{n,\lambda_{\text{RISK}}(n)},\ \eta_{n,\lambda_{\text{ASP}}(n;b)} - \eta_{n,\lambda_{\text{RISK}}(n)}) = O(n^{-pr/(pr+1)}).$$

Remark 1. Our result is for the general smoothing spline estimator. If some structures of the underlying function, e.g., shape-restricted, are known a priori, the convergence rate may be faster, and the estimator may converge in $o(\cdot)$ rather than $O(\cdot)$.

5. SIMULATION STUDIES

5.1. *Simulation settings*

Simulation studies, including univariate and multivariate cases, were carried out to assess the performance of the proposed method in terms of mean squared error. For univariate cases, we compared the proposed method with the generalized cross-validation method and the order-based method of Hall (1990). For multivariate cases, the proposed method was compared with the generalized cross-validation method, the skip method, and another three approaches, namely generalized cross-validation, restricted maximum likelihood, and fast restricted maximum likelihood (Wood et al., 2017) in generalized additive models. For the proposed method we used two sampling schemes to select subsamples: uniform sampling and asymptotic sampling. The former is described in Algorithm 1. The asymptotic sampling strategy is implemented in two steps. First, take random subsamples of size $b_1, \ldots, b_N$ from the original data and apply the generalized cross-validation method to the subsamples to estimate smoothing parameters $\lambda_{\text{GCV}}(b_1), \ldots, \lambda_{\text{GCV}}(b_N)$. Second, apply the constrained optimization method to estimate the constant C and rate parameters r and p by minimizing the objective function $(1/N)\sum_{k=1}^{N}\{\lambda_{\text{GCV}}(b_k) - Cb_k^{-r/(pr+1)}\}^2$ with constraints $p \in [1, 2]$ and $r > 1$. Compared with the uniform sampling scheme, asymptotic sampling provides empirical estimates of parameters needed for the asympirical smoothing parameters selection without using any prior knowledge on rate parameters. In multivariate cases we set $N = 10$, and b_1 and b_{10} were set to $50n^{1/4}$ and $120n^{1/4}$, respectively. In the order-based method, we directly used $n^{-r/(pr+1)}$ as the smoothing parameter λ for sample size n. The skip method is described in § 2.3. The generalized cross-validation, restricted maximum likelihood, and fast restricted maximum likelihood methods under the generalized additive models framework were implemented in the `mgcv` package (Wood, 2004, 2011; Wood et al., 2017) of R (R Development Core Team, 2021). We used the fast algorithm proposed by Kim & Gu (2004) to reduce the computational burden of fitting smoothing spline ANOVA models. To make a fair comparison, the same number of basis functions was used for all methods. We chose the generalized cross-validation method to be the benchmark and report the log-transformed relative efficacy. This relative efficacy is defined as $\sum_{i=1}^{n}\{\hat\eta(x_i) - \eta(x_i)\}^2 / \sum_{i=1}^{n}\{\tilde\eta(x_i) - \eta(x_i)\}^2$, where $\hat\eta$ is the estimator for the method being compared and $\tilde\eta$ is the estimator based on the generalized cross-validation method. A smaller log-transformed relative efficacy indicates better performance. If the log-transformed relative efficacy is zero, the method being compared has the same numerical performance as the generalized cross-validation method. Three univariate and four multivariate functions were evaluated. The full sample size n was set to 20 000, 30 000 and 40 000. Four values, 1, 2, 5 and 7, of the signal-to-noise ratio, defined as $\text{SNR} = \text{SD}\{\eta(x)\}/\sigma$, were used to generate the data. One hundred replicates were generated for each setting.

5.2. *Univariate scenarios*

We simulated the data according to (1) using three univariate functions with different orders of smoothness.

Univariate scenario 1:

$$\eta_{u1}(x) = \frac{1}{3}B_{20,5}(x) + \frac{1}{3}B_{12,12}(x) + \frac{1}{3}B_{7,30}(x),$$

where

$$B_{\alpha,\beta}(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \quad (0 \leqslant x \leqslant 1).$$

Univariate scenario 2:

$$\eta_{u2}(x) = 10\sin^2(2\pi x)\mathbb{1}_{(x\leqslant\frac{1}{2})},$$

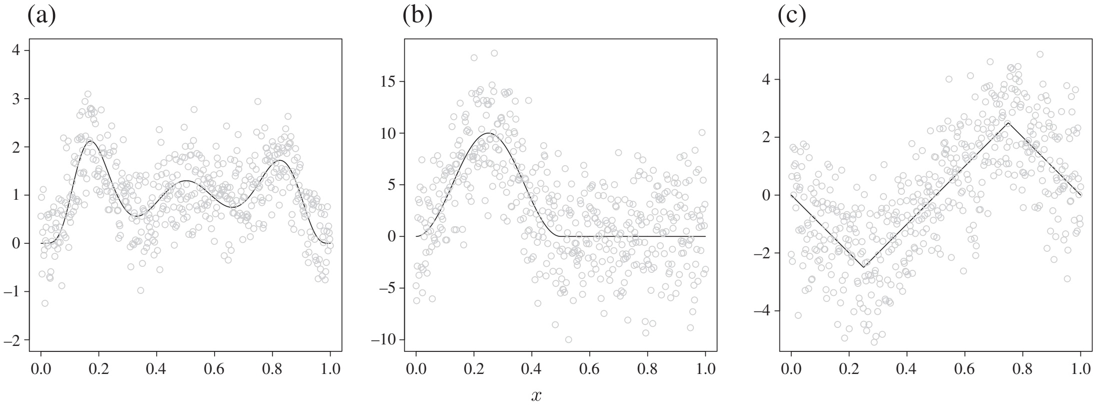

Fig. 1. The univariate true functions (solid lines) of (a) η_{u1}, (b) η_{u2} and (c) η_{u3}, with the data used in the simulation represented by circles.

where $\mathbb{1}_{(x\leqslant\frac{1}{2})}$ is an indicator function that equals 1 for $x \leqslant \frac{1}{2}$ and equals 0 otherwise.

Univariate scenario 3:

$$\eta_{u3}(x) = 10 \times \left\{-x + 2\left(x - \frac{1}{4}\right)\right\} \mathbb{1}_{(x\geqslant\frac{1}{4})} + 2\left(-x + \frac{3}{4}\right) \mathbb{1}_{(x\geqslant\frac{3}{4})},$$

where $\mathbb{1}_{(x\geqslant 1/4)}$ and $\mathbb{1}_{(x\geqslant 3/4)}$ are two indicator functions that equal 1 when the conditions in parentheses are satisfied and equal 0 otherwise.

We generated x from a uniform distribution on $[0, 1]$. The generated data for three univariate functions with SNR $= 1$ and three true function values are shown in Fig. 1. The log-transformed relative efficacies of the proposed method and the order-based method for the three scenarios are shown in Fig. 2. The skip method reduces to the generalized cross-validation method in the single smoothing parameter selection. The performance of the proposed method is comparable to that of the generalized cross-validation method when the signal-to-noise ratio is low, with log-transformed relative efficacies close to zero. The performance of our method is better than that of the generalized cross-validation method as the signal-to-noise ratio increases. Such behaviour may result from unstably estimated smoothing parameters based on subsamples when the signal-to-noise ratio is low. Even though the order-based method performs well in some scenarios, such as univariate scenario 3, it is not reliable because of the large variability in most scenarios.

5.3. *Multivariate scenarios*

We simulated the data according to (1) using four multivariate functions. In these four scenarios, the x values were drawn from the uniform distribution on $[0, 1]$.

Multivariate scenario 1:

$$\eta_{m1}(x) = \frac{0.75}{\pi\sigma_{x_{\langle 1\rangle}}\sigma_{x_{\langle 2\rangle}}} \exp\left\{-\frac{(x_{\langle 1\rangle} - 0.2)^2}{\sigma^2_{x_{\langle 1\rangle}}} - \frac{(x_{\langle 2\rangle} - 0.3)^2}{\sigma^2_{x_{\langle 2\rangle}}}\right\} + \frac{0.45}{\pi\sigma_{x_{\langle 1\rangle}}\sigma_{x_{\langle 2\rangle}}} \exp\left\{-\frac{(x_{\langle 1\rangle} - 0.7)^2}{\sigma^2_{x_{\langle 1\rangle}}} - \frac{(x_{\langle 2\rangle} - 0.8)^2}{\sigma^2_{x_{\langle 2\rangle}}}\right\},$$

where $\sigma_{x_{\langle 1\rangle}} = 0.3$ and $\sigma_{x_{\langle 2\rangle}} = 0.4$.

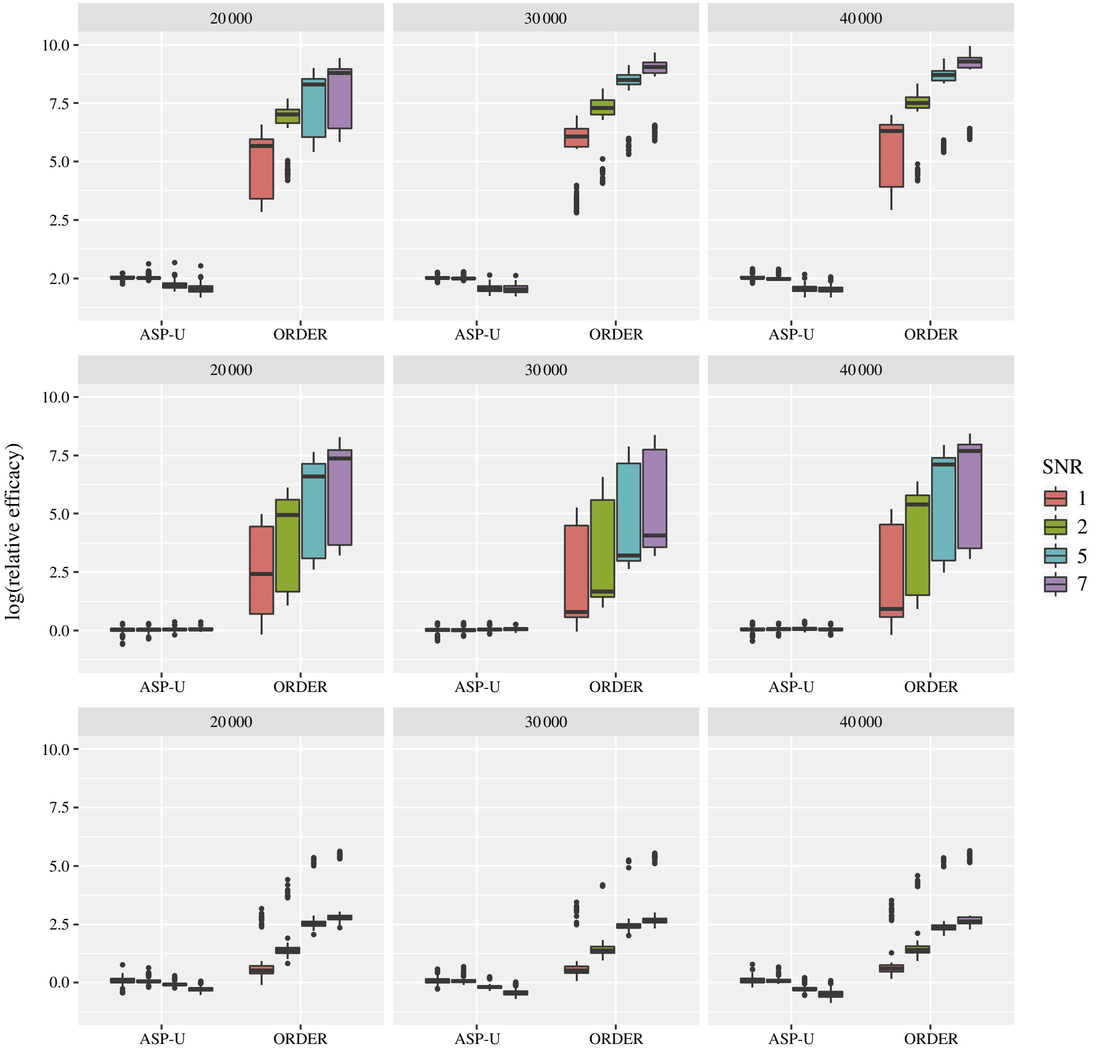

Fig. 2. Log-transformed relative efficacies of the proposed method and the order-based method with respect to the generalized cross-validation method for the three univariate scenarios. The vertical axis represents the log-transformed relative efficacies, and the horizontal axis shows the different methods. Different signal-to-noise ratios are indicated by different colours. The results of univariate scenarios 1, 2 and 3 are displayed in the upper, middle and lower panels, respectively, and the results for full sample sizes 20 000, 30 000 and 40 000 are shown in the left, middle and right columns, respectively. ASP-U, asympirical method using uniform sampling; ORDER, order-based method; SNR, signal-to-noise ratio.

Multivariate scenario 2:

$$\eta_{m2}(x) = 10\sin(\pi x_{\langle 1\rangle}) + \exp(3x_{\langle 2\rangle}) + 10^6 x_{\langle 3\rangle}^{11}(1 - x_{\langle 3\rangle})^6 + 10^4 x_{\langle 3\rangle}^3(1 - x_{\langle 3\rangle})^{10}.$$

Multivariate scenario 3:

$$\eta_{m3}(x) = 10x_{\langle 2\rangle} + 10\sin\{\pi(x_{\langle 3\rangle} - x_{\langle 2\rangle})\} + 5\cos\{2\pi(x_{\langle 1\rangle} - x_{\langle 2\rangle})\}.$$

Multivariate scenario 4:

$$\eta_{m4}(x) = \sum_{j=1}^{18} g_1(x_{\langle j \rangle}) + \sum_{j=1}^{9} g_2(x_{\langle 2j-1 \rangle}, x_{\langle 2j \rangle}) + \sum_{j=1}^{6} g_3(x_{\langle 3j-2 \rangle}, x_{\langle 3j-1 \rangle}, x_{\langle 3j \rangle}),$$

where $g_1(x) = 10^6 x^{11}(1-x)^6$, $g_2(x_{\langle 1 \rangle}, x_{\langle 2 \rangle}) = \exp(3x_{\langle 1 \rangle} x_{\langle 2 \rangle})$ and $g_3(x_{\langle 1 \rangle}, x_{\langle 2 \rangle}, x_{\langle 3 \rangle}) = 15\sin(2\pi x_{\langle 1 \rangle})/\{2 - \sin(2\pi x_{\langle 2 \rangle} x_{\langle 3 \rangle})\}$.

The full model $\eta = \eta_{\o} + \eta_1 + \eta_2 + \eta_{12}$ was considered for multivariate scenario 1, and the additive model $\eta = \eta_{\o} + \eta_1 + \eta_2 + \eta_3$ was fitted in multivariate scenario 2. In multivariate scenario 3, we considered the partial model $\eta = \eta_{\o} + \eta_2 + \eta_{23} + \eta_{12}$. We further considered the high-dimensional case in multivariate scenario 4. Log-transformed relative efficacies of all methods over the generalized cross-validation method are displayed in Fig. 3.

All the methods have similar numerical performance in multivariate scenarios 1 and 2. However, the restricted maximum likelihood method has slightly larger relative efficacies in these two scenarios. In multivariate scenario 3, the proposed method based on uniform sampling has slightly larger relative efficacies when the signal-to-noise ratio is small, but its relative efficacies become smaller as the signal-to-noise ratio increases. The proposed method based on asymptotic sampling has smaller relative efficacies than the one based on uniform sampling in this scenario. The median of relative efficacies of the methods under the generalized additive models framework is more than 35, which implies that the mean squared error of these methods is at least 35 times as large as for the generalized cross-validation method. In addition, the relative efficacies of the skip method are around 15. In the high-dimensional setting, to make the generalized cross-validation method feasible, we used the estimated smoothing parameters after the first iteration as the final smoothing parameters. It is expected that the proposed method will perform better than the one-iteration generalized cross-validation method.

Comparing the performances in multivariate scenario 3, we observe a similar phenomenon for the methods under the generalized additive models framework and the skip method in this high-dimensional setting. The median of the relative efficacies of the methods under the generalized additive models framework is about 4, while that for the proposed methods is around 0.7. The methods under the generalized additive models framework construct the bivariate interaction using two smoothing parameters, which control the smoothness on the directions of two predictors. In the smoothing spline ANOVA framework, there are three smoothing parameters associated with the bivariate interaction. The additional smoothing parameter could improve the numerical performance when the interaction is not an additive function, and this may be the reason that the proposed method performs well in the scenarios where multiple interaction components are present. The number of smoothing parameters is different for the methods under the smoothing spline ANOVA models and under the generalized additive models framework. For methods under the smoothing spline ANOVA framework, there are 5, 3, 7 and 87 effective smoothing parameters in multivariate scenarios 1, 2, 3 and 4, respectively, whereas the corresponding numbers of tunable smoothing parameters for methods under the generalized additive models framework are 4, 3, 5 and 54. Although the number of basis functions is the same for all methods, the generalized cross-validation method under the generalized additive models framework is typically faster than the one under the smoothing spline ANOVA models framework, since the method for generalized additive models has fewer tunable smoothing parameters. This is observed in the running time analysis reported in the Supplementary Material.

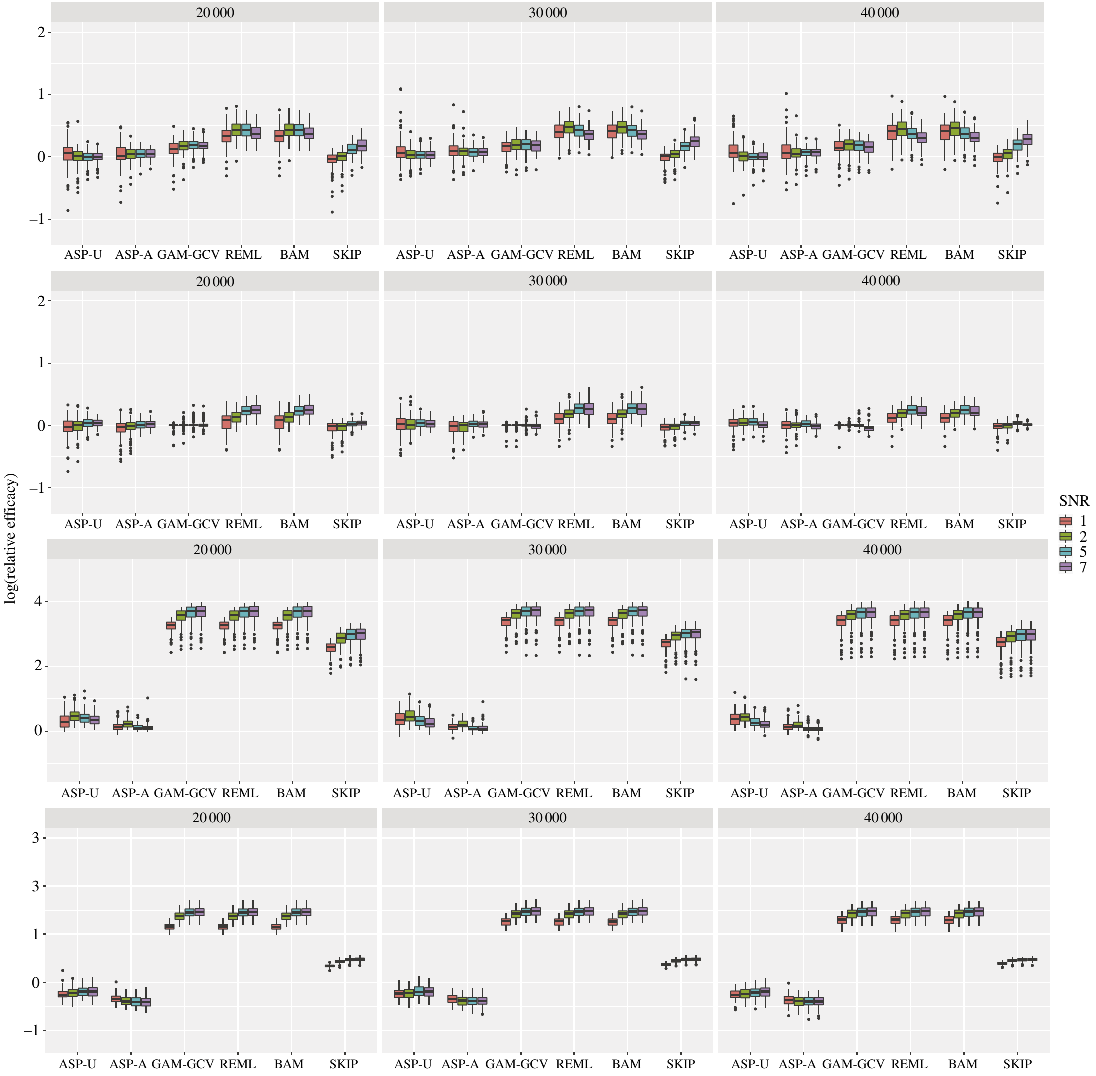

Fig. 3. Log-transformed relative efficacies of the methods under comparison in four multivariate scenarios. The vertical axis represents the log-transformed relative efficacies, and the horizontal axis shows the different methods. Different signal-to-noise ratios are indicated by different colours. From top to bottom the panels display the results for multivariate scenarios 1 to 4, and the results for full sample sizes 20 000, 30 000 and 40 000 are shown in the left, middle and right columns, respectively. ASP-U, asympirical method using uniform sampling; ASP-A, asympirical method using asymptotic sampling; GAM-GCV, generalized cross-validation for generalized additive models; REML, restricted maximum likelihood for generalized additive models; BAM, fast restricted maximum likelihood for generalized additive models; SKIP, the skip method; SNR, signal-to-noise ratio.

6. Real data examples

6.1. *Superconductivity data*

Superconductivity refers to the phenomenon wherein materials can conduct current with zero resistance. Many applications, such as magnetic resonance imaging, are based on superconductivity. Since this phenomenon is only observed at or below a characteristic critical temperature, prediction of the critical temperature of a superconductor is an important problem. In this real data example, we aim to predict the critical temperature by using elemental properties extracted from superconductors. The response is the critical temperature in kelvins. The predictors represent the

Table 1. *Fitting and prediction statistics of the methods under comparison applied to the superconductivity dataset*

Method	R^2	Root fitting MSE	Root prediction MSE (mean)	Root prediction MSE (sd)	CPU time (s)
ASP-U	0.786	15.675	15.871	0.239	0.030
ASP-A	0.785	15.719	15.870	0.281	0.790
GAM-GCV	0.765	16.556	16.627	0.249	0.270
REML	0.764	16.625	16.630	0.252	15.200
BAM	0.763	16.625	16.645	0.248	0.062
GCV	0.789	15.363	15.514	0.289	40.560

MSE, mean squared error; sd, standard deviation; ASP-U, asympirical method using uniform sampling; ASP-A, asympirical method using asymptotic sampling; GAM-GCV, generalized cross-validation for generalized additive models; REML, restricted maximum likelihood for generalized additive models; BAM, fast restricted maximum likelihood for generalized additive models; GCV, generalized cross-validation method.

elemental properties of a superconductor. For instance, one can derive a feature of the superconductor by calculating the average thermal conductivities of the elements in its chemical formula. More details about all the predictors in this example are available in Hamidieh (2018). The dataset contains 21 263 observations. We fit the cubic tensor product smoothing spline ANOVA model to the dataset. Based on the preliminary model diagnostics (Gu, 2004), we consider the following functional ANOVA decomposition:

$$\eta(x) = \eta_{\o} + \sum_{j=1}^{42} \eta_j(x_{\langle j \rangle}),$$

where $\eta_{\o}$ is a constant function and $\eta_1(x_{\langle 1 \rangle}), \ldots, \eta_{42}(x_{\langle 42 \rangle})$ denote the main-effect functions for 42 selected features. Details of the selected features can be downloaded at `https://github.com/shawnstat/Asympirical-Smoothing-Parameters-Selection`. There are 42 effective smoothing parameters in the decomposition. For a fair comparison, the number of basis functions for all methods is taken to be $10n^{2/9}$ (Kim & Gu, 2004).

Table 1 shows the fitting and prediction statistics for the methods under comparison. To evaluate the prediction performance, we compare the five-fold cross-validated root mean squared errors obtained by dividing the full data into five equal parts. The mean and standard deviation of the five root mean squared errors in predicting the testing data are reported. Compared with the proposed methods and the methods under the generalized additive models framework, the generalized cross-validation method has better performance in terms of fitting and prediction mean squared errors. On the other hand, the proposed methods are much faster in terms of CPU time.

6.2. *Molecular dynamics data*

With the aid of modern quantum chemistry methods, researchers can conduct systematic simulations of quantum chemical systems, obtaining accurate results on molecular dynamics at the quantum level. Analysis of such molecular dynamics trajectories is crucial for the discovery of new chemicals (Chmiela et al., 2017; Schütt et al., 2017). The molecular dynamics data on malondialdehyde used in this example contain 893 238 observations. The response is the energy in kcal/mol. The predictors encode the molecular structure, which is measured in terms of the reciprocal of the pairwise Euclidean distance between atoms (Montavon et al., 2013). Since there are nine atoms in malondialdehyde, we have a distance vector of length 36 for each trajectory.

Table 2. *Fitting and prediction statistics of the methods under comparison applied to the molecular dynamics dataset*

Method	R^2	Root fitting MSE	Root prediction MSE (mean)	Root prediction MSE (sd)	CPU time (s)
ASP-U	0.925	1.130	1.134	0.006	1.596
ASP-A	0.926	1.124	1.134	0.003	1.969
GAM-GCV	0.911	1.229	1.226	0.003	4.891
BAM	0.913	1.219	1.224	0.006	0.490
SKIP	0.918	1.173	1.162	0.010	193.788

MSE, mean squared error; sd, standard deviation; ASP-U, asympirical method using uniform sampling; ASP-A, asympirical method using asymptotic sampling; GAM-GCV, generalized cross-validation for generalized additive models; BAM, fast restricted maximum likelihood for generalized additive models; SKIP, skip method.

Therefore, there are 36 predictors for this dataset. We fit the cubic tensor product smoothing spline ANOVA model to the dataset. Based on the preliminary model diagnostics (Gu, 2004), we consider the following functional ANOVA decomposition:

$$\eta(x) = \eta_{\o} + \sum_{j=1}^{36} \eta_j(x_{\langle j\rangle}) \\ + \eta_{9,10}(x_{\langle 9\rangle}, x_{\langle 10\rangle}) + \eta_{9,13}(x_{\langle 9\rangle}, x_{\langle 13\rangle}) + \eta_{24,35}(x_{\langle 24\rangle}, x_{\langle 35\rangle}) + \eta_{25,36}(x_{\langle 25\rangle}, x_{\langle 36\rangle}),$$

where $x_{\langle j\rangle}$ $(j = 1, \ldots, 36)$ is the jth predictor. The numbers of smoothing parameters for the proposed methods and the methods under the generalized additive models framework are 48 and 44, respectively. Bearing in mind the limits on computational resources, we set the number of basis functions to $4.3n^{2/9}$ for all methods (Kim & Gu, 2004).

We compare the fitting and prediction errors of these smoothing parameter selection methods in Table 2. The mean and standard deviation of the five root mean squared error results for the testing datasets are reported as the prediction error. Since the generalized cross-validation method was infeasible even for one iteration, we compared only the proposed methods and the methods under the generalized additive models framework with the skip method. We also compared the proposed methods with the fast restricted maximum likelihood method for generalized additive models (Wood et al., 2017). The proposed method based on asymptotic sampling has the best performance in terms of fitting and prediction errors. The fast restricted maximum likelihood method for generalized additive models is the fastest in terms of CPU time.

Acknowledgement

This research was supported by the U.S. National Science Foundation and National Institutes of Health. We wish to thank Dr Chong Gu and the reviewers for generously providing valuable comments and suggestions.

Supplementary material

Supplementary material available at *Biometrika* online includes further simulation results and detailed proofs of the theoretical results.

References

Aydin, D., Memmedli, M. & Omay, R. E. (2013). Smoothing parameter selection for nonparametric regression using smoothing spline. *Eur. J. Pure Appl. Math.* **6**, 222–38.

Buja, A., Hastie, T. & Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17**, 453–510.

Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T. & Müller, K.-R. (2017). Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015.

Cox, D. D. (1984). Multivariate smoothing spline functions. *SIAM J. Numer. Anal.* **21**, 789–813.

Cox, D. D. & O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18**, 1676–95.

Craven, P. & Wahba, G. (1978). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377–403.

Gu, C. (2004). Model diagnostics for smoothing spline ANOVA models. *Can. J. Statist.* **32**, 347–58.

Gu, C. (2013). *Smoothing Spline ANOVA Models*, vol. 297 of *Springer Series in Statistics*. New York: Springer.

Gu, C. (2014). Smoothing spline ANOVA models: R package gss. *J. Statist. Software* **58**, 1–25.

Gu, C. & Qiu, C. (1993). Smoothing spline density estimation: Theory. *Ann. Statist.* **21**, 217–34.

Gu, C. & Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comp.* **12**, 383–98.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Mult. Anal.* **32**, 177–203.

Hamidieh, K. (2018). A data-driven statistical model for predicting the critical temperature of a superconductor. *Comp. Mater. Sci.* **154**, 346–54.

Hastie, T. & Tibshirani, R. (1986). Generalized additive models. *Statist. Sci.* **1**, 297–310.

Helwig, N. E. & Ma, P. (2015). Fast and stable multiple smoothing parameter selection in smoothing spline analysis of variance models with large samples. *J. Comp. Graph. Statist.* **24**, 715–32.

Hurvich, C. M., Simonoff, J. S. & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* B **60**, 271–93.

Kim, Y.-J. & Gu, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. R. Statist. Soc.* B **66**, 337–56.

Kimeldorf, G. & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33**, 82–95.

Lee, D.-J. & Durbán, M. (2011). P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Statist. Mod.* **11**, 49–69.

Lee, D.-J., Durbán, M. & Eilers, P. (2013). Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. *Comp. Statist. Data Anal.* **61**, 22–37.

Lin, Y. (2000). Tensor product space ANOVA models. *Ann. Statist.* **28**, 734–55.

Ma, P., Huang, J. Z. & Zhang, N. (2015). Efficient computation of smoothing splines via adaptive basis sampling. *Biometrika* **102**, 631–45.

Mallows, C. L. (1973). Some comments on C_P. *Technometrics* **15**, 661–75.

Marx, B. D. & Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Comp. Statist. Data Anal.* **28**, 193–209.

Meng, C., Zhang, X., Zhang, J., Zhong, W. & Ma, P. (2020). More efficient approximation of smoothing splines via space-filling basis selection. *Biometrika* **107**, DOI: 10.1093/biomet/asaa019.

Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K.-R. & Von Lilienfeld, O. A. (2013). Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003.

R Development Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Rice, J. & Rosenblatt, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. *Ann. Statist.* **11**, 141–56.

Rodríguez-Álvarez, M. X., Lee, D.-J., Kneib, T., Durbán, M. & Eilers, P. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: The SAP algorithm. *Statist. Comp.* **25**, 941–57.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. *Nature Commun.* **8**, 13890.

Shang, Z. & Cheng, G. (2017). Computational limits of a distributed algorithm for smoothing spline. *J. Mach. Learn. Res.* **18**, 3809–45.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795–810.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.* **13**, 970–83.

Wahba, G. (1975). Smoothing noisy data with spline functions. *Numer. Math.* **24**, 383–93.

Wahba, G. (1977). Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.* **14**, 651–67.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Statist.* **13**, 1378–402.

Wahba, G. (1990). *Spline Models for Observational Dta*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia: SIAM.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23**, 1865–95.

Wand, M. P. (2003). Smoothing and mixed models. *Comp. Statist.* **18**, 223–49.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. Boca Raton, Florida: CRC Press.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Statist. Assoc.* **99**, 673–86.

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62**, 1025–36.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Statist. Soc.* B **73**, 3–36.

Wood, S. N. & Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics* **73**, 1071–81.

Wood, S. N., Li, Z., Shaddick, G. & Augustin, N. H. (2017). Generalized additive models for gigadata: Modeling the UK black smoke network daily data. *J. Am. Statist. Assoc.* **112**, 1199–210.

Wood, S. N., Scheipl, F. & Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statist. Comp.* **23**, 341–60.

Xu, D. & Wang, Y. (2018). Divide and recombine approaches for fitting smoothing spline models with large datasets. *J. Comp. Graph. Statist.* **27**, 677–83.

[*Received on* 17 *February* 2019. *Editorial decision on* 12 *March* 2020]