Biometrika (2020), **107**, 3, *pp.* 723–735
Printed in Great Britain
doi: 10.1093/biomet/asaa019
Advance Access publication 7 May 2020

More efficient approximation of smoothing splines via space-filling basis selection

BY CHENG MENG, XINLIAN ZHANG, JINGYI ZHANG,
WENXUAN ZHONG AND PING MA

Department of Statistics, University of Georgia, 310 Herty Dr., Athens, Georgia 30602, U.S.A.
cheng.meng25@uga.edu xinlian.zhang25@uga.edu jingyi.zhang25@uga.edu
wenxuan@uga.edu pingma@uga.edu

SUMMARY

We consider the problem of approximating smoothing spline estimators in a nonparametric regression model. When applied to a sample of size n, the smoothing spline estimator can be expressed as a linear combination of n basis functions, requiring $O(n^3)$ computational time when the number d of predictors is two or more. Such a sizeable computational cost hinders the broad applicability of smoothing splines. In practice, the full-sample smoothing spline estimator can be approximated by an estimator based on q randomly selected basis functions, resulting in a computational cost of $O(nq^2)$. It is known that these two estimators converge at the same rate when q is of order $O\{n^{2/(pr+1)}\}$, where $p \in [1, 2]$ depends on the true function and $r > 1$ depends on the type of spline. Such a q is called the essential number of basis functions. In this article, we develop a more efficient basis selection method. By selecting basis functions corresponding to approximately equally spaced observations, the proposed method chooses a set of basis functions with great diversity. The asymptotic analysis shows that the proposed smoothing spline estimator can decrease q to around $O\{n^{1/(pr+1)}\}$ when $d \leqslant pr + 1$. Applications to synthetic and real-world datasets show that the proposed method leads to a smaller prediction error than other basis selection methods.

Some key words: Nonparametric regression; Penalized least squares criterion; Space-filling design; Star discrepancy; Subsampling.

1. INTRODUCTION

Consider the nonparametric regression model $y_i = \eta(x_i)+\epsilon_i$ $(i = 1, \ldots, n)$, where $y_i \in \mathbb{R}$ is the ith observation of the response, η represents an unknown function to be estimated, $x_i \in \mathbb{R}^d$ is the ith observation of the predictor variable, and $\{\epsilon_i\}_{i=1}^n$ are independent and identically distributed random errors with zero mean and unknown variance σ^2. The function η can be estimated by minimizing the penalized least squares criterion,

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

where $J(\eta)$ is a quadratic roughness penalty (Wahba, 1990; Wang et al., 2011; Gu, 2013). The smoothing parameter λ controls the trade-off between the goodness-of-fit of the model and the roughness of the function η. In this paper, expression (1) is minimized in a reproducing kernel Hilbert space $\mathcal{H}$, which leads to a smoothing spline estimate for η.

Although univariate smoothing splines can be computed in $O(n)$ time (Reinsch, 1967), it takes $O(n^3)$ time to find the minimizer of (1) when $d \geqslant 2$. Such a high computational cost hinders the use of smoothing splines for large samples. To reduce the computational cost for smoothing splines, extensive efforts have been made to approximate the minimizer of (1) by restricting the estimator $\hat{\eta}$ to a subspace $\mathcal{H}_E \subset \mathcal{H}$. Let the dimension of the space $\mathcal{H}_E$ be q and the restricted estimator be $\hat{\eta}_E$. Compared with the $O(n^3)$ computational cost of calculating $\hat{\eta}$, the computational cost of $\hat{\eta}_E$ can be reduced to $O(nq^2)$. Along this line, numerous methods have been developed in recent decades. Luo & Wahba (1997) and Zhang et al. (2004) approximated the minimizer of (1) using variable selection techniques. Pseudosplines (Hastie, 1996) and penalized splines (Ruppert et al., 2009) were also used to approximate smoothing splines.

Although these methods have already yielded impressive algorithmic benefits, they are usually ad hoc in terms of choosing the value of q. The value of q regulates the trade-off between the computational time and the prediction accuracy. One fundamental question is how small q can be to ensure that the restricted estimator $\hat{\eta}_E$ converges to the true function η at the same rate as the full-sample estimator $\hat{\eta}$. To answer this question, Gu & Kim (2002) and Ma et al. (2015) developed random sampling methods for selecting the basis functions and established a coherent theory for the convergence of the restricted estimator $\hat{\eta}_E$. To ensure that $\hat{\eta}_E$ has the same convergence rate as $\hat{\eta}$, the methods in both Gu & Kim (2002) and Ma et al. (2015) require q to be of order $O\{n^{2/(pr+1)+\delta}\}$, where δ is an arbitrary small positive number, $p \in [1,2]$ depends on the true function η, and r depends on the fitted spline. In a 1999 PhD thesis by F. Gao from the University of Wisconsin-Madison, it is shown that fewer basis functions are needed to guarantee the aforementioned convergence rate if one selects the basis functions $\{R(z_j,\cdot)\}_{j=1}^q$ with $\{z_j\}_{j=1}^q$ approximately equally spaced. However, Gao provided theory only in the univariate predictor case, and that method cannot be directly applied to the multivariate setting.

In this paper, we develop a more efficient computational method for approximating smoothing splines. The distinguishing feature of our method is that it incorporates the notion of diversity of the selected basis functions. We propose a space-filling basis selection method, which chooses basis functions with large diversity by selecting the ones that correspond to roughly uniformly distributed observations. When $d \leqslant pr+1$, we show that the proposed smoothing spline estimator has the same convergence rate as the full-sample estimator, and the order of the essential number q of basis functions is reduced to $O\{n^{(1+\delta)/(pr+1)}\}$.

2. SMOOTHING SPLINES AND THE BASIS SELECTION METHOD

Let $\mathcal{H} = \{\eta : J(\eta) < \infty\}$ be a reproducing kernel Hilbert space, where $J(\cdot)$ is a squared seminorm. Let $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ be the null space of $J(\eta)$ and assume that $\mathcal{N}_J$ is a finite-dimensional linear subspace of $\mathcal{H}$ with basis $\{\xi_i\}_{i=1}^m$, where m is the dimension of $\mathcal{N}_J$. Let $\mathcal{H}_J$ be the orthogonal complement of $\mathcal{N}_J$ in $\mathcal{H}$, so that $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$. The space $\mathcal{H}_J$ is a reproducing kernel Hilbert space with $J(\cdot)$ as the squared norm. The reproducing kernel of $\mathcal{H}_J$ is denoted by $R_J(\cdot,\cdot)$. The well-known representer theorem (Wahba, 1990) states that there exist vectors $d = (d_1,\ldots,d_m)^{\mathrm{T}} \in \mathbb{R}^m$ and $c = (c_1,\ldots,c_n)^{\mathrm{T}} \in \mathbb{R}^n$ such that the minimizer of (1) in $\mathcal{H}$ is $\eta(x) = \sum_{k=1}^m d_k\xi_k(x) + \sum_{i=1}^n c_iR_J(x_i,x)$. Let $Y = (y_1,\ldots,y_n)^{\mathrm{T}}$ be the vector of response observations, S the $n \times m$ matrix with (i,j)th entry $\xi_j(x_i)$, and R the $n \times n$ matrix with (i,j)th entry $R_J(x_i,x_j)$. Solving the problem of minimizing (1) is therefore equivalent to solving

$$(\hat{d},\hat{c}) = \arg\min_{d,c} \frac{1}{n}(Y - Sd - Rc)^{\mathrm{T}}(Y - Sd - Rc) + \lambda c^{\mathrm{T}}Rc, \tag{2}$$

where the smoothing parameter λ can be selected based on the generalized cross-validation criterion (Wahba & Craven, 1978). In the general case of $n \gg m$ and $d \geqslant 2$, the computational cost of calculating $(\hat{d}, \hat{c})$ in (2) is $O(n^3)$, which is prohibitive when the sample size n is large. To reduce the computational burden, one can restrict the full-sample estimator $\hat{\eta}$ to a subspace $\mathcal{H}_E \subset \mathcal{H}$, where $\mathcal{H}_E = \mathcal{N}_J \oplus \text{span}\{R_J(x_i^*, \cdot) : i = 1, \ldots, q\}$. Here $\mathcal{H}_E$, called the effective model space, can be constructed by selecting a subsample $\{x_i^*\}_{i=1}^q$ from $\{x_i\}_{i=1}^n$. Such an approach is called a basis selection method.

Denote by $R_* \in \mathbb{R}^{n \times q}$ the matrix with (i,j)th entry $R_J(x_i, x_j^*)$ and by $R_{**} \in \mathbb{R}^{q \times q}$ the matrix with (i,j)th entry $R_J(x_i^*, x_j^*)$. The minimizer of (1) in the effective model space $\mathcal{H}_E$ can thus be written as $\eta_E(x) = \sum_{k=1}^m d_k \xi_k(x) + \sum_{i=1}^q c_i R(x_i^*, x)$, and the coefficients $d_E = (d_1, \ldots, d_m)^{\mathrm{T}}$ and $c_E = (c_1, \ldots, c_q)^{\mathrm{T}}$ can be obtained by solving

$$(\hat{d}_E, \hat{c}_E) = \underset{d_E,\, c_E}{\arg\min} \frac{1}{n}(Y - Sd_E - R_* c_E)^{\mathrm{T}}(Y - Sd_E - R_* c_E) + \lambda c_E^{\mathrm{T}} R_{**} c_E.$$

The restricted estimator $\hat{\eta}_E$ evaluated at the sample points therefore satisfies $\hat{\eta}_E = S\hat{d}_E + R_* \hat{c}_E$, where $\hat{\eta}_E = \{\hat{\eta}_E(x_1), \ldots, \hat{\eta}_E(x_n)\}^{\mathrm{T}}$. In the general case where $m \ll q \ll n$, the overall computational cost of the basis selection method is $O(nq^2)$, which is a significant reduction from $O(n^3)$. Recall that the value of q controls the trade-off between the computational time and the prediction accuracy. To ensure that $\hat{\eta}_E$ converges to the true function η at the same rate as $\hat{\eta}$, researchers have shown that the essential number q of basis functions needs to be of order $O\{n^{2/(pr+1)+\delta}\}$, where δ is an arbitrary small positive number (Kim & Gu, 2004; Ma et al., 2015). In the next section, we present a space-filling basis selection method which reduces this order to $O\{n^{(1+\delta)/(pr+1)}\}$.

3. Space-filling basis selection

3.1. *Motivation and notation*

To motivate the development of the proposed method, we first re-examine ensemble learning methods that are well-known in the statistics and machine learning communities (Dietterich, 2002; Rokach, 2010). To achieve better predictive performance than a single learner, which is either a model or a learning algorithm, ensemble learning methods first build a committee consisting of a number of different learners, and then aggregate the predictions of the learners in the committee. The aggregation is usually achieved by employing majority vote or by calculating a weighted average. The diversity among the learners in the committee is key to the success of ensemble learning methods; greater diversity in the committee yields better performance (Kuncheva & Whitaker, 2003).

The restricted smoothing spline estimator $\hat{\eta}_E$ can be considered an ensemble learning method. In particular, the prediction of $\hat{\eta}_E$ is done by taking a weighted average of the predictions of the selected basis functions $R_J(x_i^*, \cdot)$ $(i \in \{1, \ldots, q\})$ in addition to the basis functions in the null space $\mathcal{N}_J$. Inspired by Kuncheva & Whitaker (2003), we propose to select a subsample $\{x_i^*\}_{i=1}^q$ such that the diversity among the basis functions $\{R_J(x_i^*, \cdot)\}_{i=1}^q$ is as large as possible. One crucial question is how to measure the diversity in a set of basis functions. Notice that adjacent data points, $x_i^* \approx x_j^*$ $(i,j \in \{1, \ldots, q\})$, yield similar basis functions, i.e., $R_J(x_i^*, \cdot) \approx R_J(x_j^*, \cdot)$. On the other hand, if x_i^* is far away from x_j^*, the basis function $R_J(x_i^*, \cdot)$ tends to be different from $R_J(x_j^*, \cdot)$. These observations motivate us to select a set of basis functions $\{R_J(x_i^*, \cdot)\}_{i=1}^q$ where $\{x_i^*\}_{i=1}^q$ are as uniformly distributed as possible. The uniformly distributed property, usually referred to

as the space-filling property in the experimental design literature (Pukelsheim, 2006), can be systematically measured by the star discrepancy.

Since the star discrepancy is defined for data in $[0,1]^d$, in practice we need to map data with an arbitrary distribution to this domain. Suppose $\mathcal{X}_n = \{x_i\}_{i=1}^n$ are independent and identically distributed observations generated from the cumulative distribution function F with bounded support $\mathcal{D} \subset \mathbb{R}^d$. Let τ be a transformation such that $\{\tau(x_i)\}_{i=1}^n$ follows the uniform distribution on $[0,1]^d$. In the simple case where $d = 1$ and F is known, we can find the transformation τ by setting $\tau = F$. In the more general case where $d > 1$ and F is unknown, the transformation τ can be calculated using optimal transport theory (Villani, 2008). However, finding the exact solution via optimal transport theory can be time-consuming. Instead, one can approximate the transformation τ using the iterative transformation approach of Pukelsheim (2006) or Meng et al. (2019), or the sliced optimal transport map approach of Rabin et al. (2011). The computational cost of these two approaches is $O\{Kn\log(n)\}$, where K is the number of iterations (Cuturi & Doucet, 2014; Bonneel et al., 2015; Kolouri et al., 2018). This cost is negligible compared with that of the proposed method. In practice, the data can always be pre-processed using the τ transformation. Without loss of generality, one may assume the $\mathcal{X}_n$ to be independent and identically distributed observations generated from the uniform distribution on $[0,1]^d$.

3.2. *Star discrepancy and space-filling design*

Let $a = (a_1, \ldots, a_d)^{\mathrm{T}} \in [0,1]^d$, let $[0,a) = \prod_{j=1}^d [0,a_j)$ be a hyper-rectangle, and let $\mathbb{1}\{\cdot\}$ denote the indicator function. The local discrepancy and the star discrepancy are defined as follows (Fang et al., 2005; Pukelsheim, 2006).

DEFINITION 1. *Given $\mathcal{X}_q = \{x_1, \ldots, x_q\}$ in $[0,1]^d$ and a hyper-rectangle $[0,a)$, the corresponding local discrepancy is defined as $D(\mathcal{X}_q, a) = \left|(1/q)\sum_{i=1}^q \mathbb{1}\{x_i \in [0,a)\} - \prod_{j=1}^d a_j\right|$. The star discrepancy corresponding to $\mathcal{X}_q$ is defined as $D^*(\mathcal{X}_q) = \sup_{a\in[0,1]^d} D(\mathcal{X}_q, a)$.*

The local discrepancy $D(\mathcal{X}_q, a)$ measures the difference between the volume of the hyper-rectangle $[0,a)$ and the fraction of data points located in $[0,a)$; it is illustrated in Fig. 1(a). The star discrepancy $D^*(\mathcal{X}_q)$ calculates the supremum over $a \in [0,1]^d$ of all the local discrepancies; in other words, the smaller $D^*(\mathcal{X}_q)$ is, the more space-filling the data points $\mathcal{X}_q$ are (Fang et al., 2005).

Chung (1949) showed that when $\mathcal{X}_q$ is generated from the uniform distribution on $[0,1]^d$, $D^*(\mathcal{X}_q)$ converges to zero with order of convergence $O[\{\log\log(q)/q\}^{1/2}]$. Faster convergence rates can be achieved using space-filling design methods (Pukelsheim, 2006) or the low-discrepancy sequence method (Halton, 1960; Sobol, 1967; Owen, 2003). Space-filling design methods, developed in the experimental design literature, seek to generate a set of design points that can cover the space as uniformly as possible. For further details see, for example, Fang et al. (2005), Pukelsheim (2006) and Wu & Hamada (2011). The low-discrepancy sequence method is frequently used in quasi-Monte Carlo and is extensively employed for numerical integration. For a Sobol sequence $\mathcal{S}_q$, one type of low-discrepancy sequence, it is known that $D^*(\mathcal{S}_q)$ is of order $O\{\log(q)^d/q\}$, which is roughly the squared order of $D^*(\mathcal{X}_q)$ for fixed d. For more in-depth discussions on quasi-Monte Carlo methods see, for example, Lemieux (2009), Dick et al. (2013), Glasserman (2013, Ch. 5) or Leobacher & Pillichshammer (2014) and references therein.

Existing studies suggest that the space-filling property can be exploited to achieve a fast convergence rate for numerical integration and response surface estimation (Fang et al., 2005; Pukelsheim, 2006). These results inspire us to choose space-filling basis functions for smoothing splines. Unfortunately, existing techniques of space-filling design cannot be directly applied to

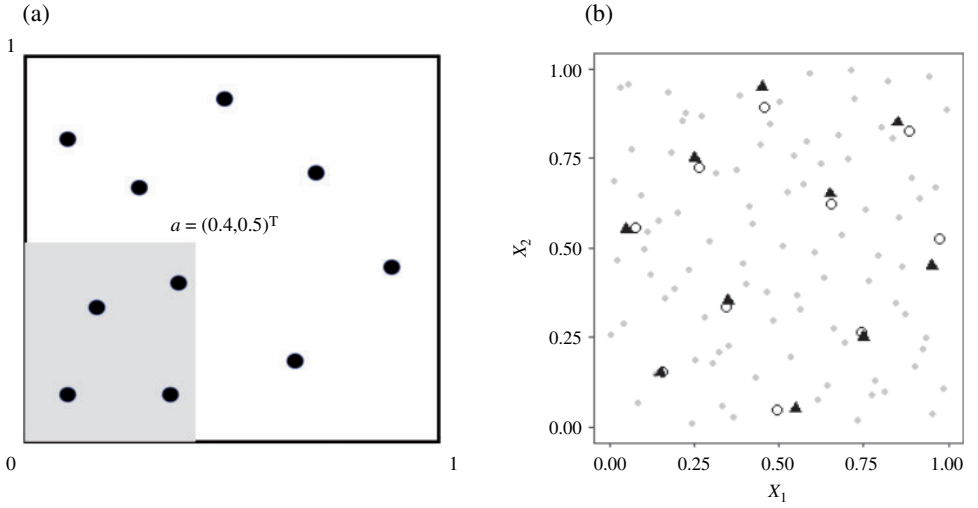

Fig. 1. (a) A toy example to illustrate local discrepancy: 10 data points are generated in $[0,1]^2$, with four of them located in the shaded rectangle $[0,a)$ where $a=(0.4,0.5)^{\mathrm{T}}$; the local discrepancy is $D(\mathcal{X}_{10},a)=|4/10-0.4\times 0.5|=0.2$. (b) Illustration of the proposed basis selection method: the data points are represented by grey dots, and the nearest-neighbour data point of each design point (black triangle) is represented by a circle.

our basis selection problem because, while the design space in space-filling design methods is typically continuous, our sample space $\{x_i\}_{i=1}^n$ is finite and discrete. We propose an algorithm that will enable us to overcome this obstacle.

3.3. *Main algorithm*

We develop a space-filling basis selection method in which we select the space-filling data points in a computationally efficient manner. First, a set of design points $\mathcal{S}_q=\{s_i\}_{i=1}^q\in[0,1]^d$ is generated, using either a low-discrepancy sequence or a space-filling design method. Next, the nearest neighbour x_i^* of each s_i is selected from the sample points $\{x_i\}_{i=1}^n$. Thus, $\{x_i^*\}_{i=1}^q$ can approximate the design points $\mathcal{S}_q$ well, provided each x_i^* $(i=1,\ldots,q)$ is sufficiently close to s_i. The method is summarized as follows.

Step 1. Generate a set of design points $\{s_i\}_{i=1}^q$ from $[0,1]^d$.

Step 2. Select the nearest neighbour of each design point s_i from $\{x_i\}_{i=1}^n$. Let the selected data points be $\{x_i^*\}_{i=1}^q$.

Step 3. Minimize the penalized least squares criterion (1) over the effective model space $\mathcal{H}_E=\mathcal{N}_J\oplus\mathrm{span}\{R_J(x_i^*,\cdot):i=1,\ldots,q\}$.

The proposed algorithm is illustrated through a toy example in Fig. 1(b). One hundred data points, grey dots, are generated from the uniform distribution in $[0,1]^2$, and a set of design points, black triangles, is generated via the max projection design (Joseph et al., 2015), a recently developed space-filling design method. The nearest neighbour of each design point is selected, circles. One can see that the selected subsample is space-filling since it can effectively approximate the design points.

In Fig. 2 the proposed space-filling basis selection method is compared with the uniform basis selection method of Gu & Kim (2002) and the adaptive basis selection method of

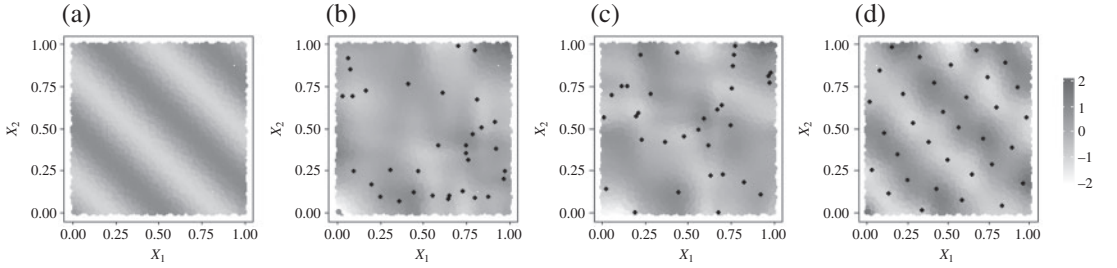

Fig. 2. Comparison of different basis selection methods: heatmaps for (a) the true function and spline estimates based on (b) the uniform basis selection method, (c) the adaptive basis selection method, and (d) the proposed space-filling basis selection method; black dots represent the sampled basis functions. The proposed method is observed to outperform the other methods in approximating the true function.

Ma et al. (2015) on a two-dimensional toy example. We generate 5000 data points from the uniform distribution in $[0, 1]^2$. Figure 2(a) shows the heatmap for the true response surface $y = \sin\{5(x_1 + x_2)\}$. The dimension q of the effective model space is set to $5 \times (5000)^{2/9} \approx 33$ for all basis selection methods. The selected basis functions are represented by black dots in each panel. Panels (b)–(d) display the heatmaps of the spline estimates of the three basis selection methods. In the uniform basis selection method, the default random number generator in R (R Development Core Team, 2020) is used to select the basis functions. It is observed that the selected points are not uniformly distributed. This is a very common phenomenon in uniform basis selection because the randomly selected points do not necessarily look uniformly distributed, especially when the number of selected points is small. In contrast, it can be seen that the basis functions selected by the proposed method are space-filling. Using space-filling design techniques, our method avoids the pitfalls of the uniform basis selection method and yields uniformly distributed selected points. The proposed method appears better than the other methods at estimating the true response.

Next we calculate the computational cost of the proposed method. In Step 1 the design points can be generated beforehand, so the computational cost of Step 1 can be ignored. For the nearest-neighbour search in Step 2 we employ the k-d tree method, which takes $O\{n \log(n)\}$ flops (Bentley, 1975; Wald & Havran, 2006). The computational cost of this step can be further reduced if we are willing to sacrifice the accuracy of the search results, for example by using approximate nearest-neighbour search algorithms (Altman, 1992; Arya et al., 1994). In Step 3, computation of the smoothing spline estimates in the restricted space $\mathcal{H}_E$ is $O(nq^2)$, as discussed in § 2. In summary, the overall computational cost of the space-filling basis selection method is $O(nq^2)$.

4. Convergence rates for function estimation

Recall that the data points are assumed to be generated from the uniform distribution on $[0, 1]^d$. Thus, for each coordinate x, the corresponding marginal density is $f_X(\cdot) = 1$. We define $V(g) = \int_{[0,1]^d} g^2 \, \mathrm{d}x$. The following four regularity conditions are required for the asymptotic analysis; the first three are identical to conditions used in Ma et al. (2015), where one can find more technical discussions.

Condition 1. The function V is completely continuous with respect to J.

Condition 2. For some $\beta > 0$ and $r > 1$, $\rho_\nu > \beta \nu^r$ for sufficiently large ν.

Condition 3. For all μ and ν, $\mathrm{var}\{\phi_\nu(x)\phi_\mu(x)\} \leqslant C_1$, where ϕ_ν and ϕ_μ are the eigenfunctions associated with V and J in $\mathcal{H}$, and C_1 is a positive constant.

Condition 4. For all μ and ν, $\mathcal{V}(g_{\nu,\mu}) \leqslant C_2$, where $\mathcal{V}(\cdot)$ denotes the total variation, $g_{\nu,\mu}(x) = \phi_\nu(x)\phi_\mu(x)$, and C_2 represents a positive constant. Here the total variation is defined in the sense of Hardy and Krause (Owen, 2003). As a specific case, when $d = 1$, the total variation is $\mathcal{V}(g) = \int |g'(x)|\,\mathrm{d}x$, indicating that a smooth function exhibits small total variation. Intuitively, the total variation measures how wiggly the function is.

Condition 1 indicates that there exists a sequence of eigenfunctions $\phi_\nu \in \mathcal{H}$ and associated eigenvalues $\rho_\nu \uparrow \infty$ satisfying $V(\phi_\nu, \phi_\mu) = \delta_{\nu\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu\mu}$, where $\delta_{\nu\mu}$ is the Kronecker delta. The growth rate of the eigenvalues ρ_ν dictates how fast λ should approach zero and, further, what the convergence rate of the smoothing spline estimates is (Gu, 2013). The eigenfunctions ϕ_ν have a close relationship with the Demmler–Reinsch basis, which consists of orthogonal vectors representing the l_2-norm (Ruppert, 2002). The eigenfunctions ϕ_ν can be calculated explicitly in several particular scenarios. For instance, they are the sine and cosine functions when $J(\eta) = \int_0^1 (\eta'')^2\,\mathrm{d}x$, where η is a periodic function on $[0, 1]$. More details on the construction of the ϕ_ν functions can be found in Gu (2013, § 9.1).

We now present our main theoretical results; all proofs are given in the Supplementary Material. For a set of design points $\mathcal{S}_q$ of size q, we assume that the star discrepancy $D^*(\mathcal{S}_q)$ converges to zero at a rate of $O\{\log(q)^d/q\}$, or $O\{q^{-(1-\delta)}\}$ for an arbitrary small positive number δ. Such a convergence rate is warranted if $\mathcal{S}_q$ is generated from a low-discrepancy sequence or space-filling design method, as discussed earlier. Recall that the proposed method aims to select a subsample that is space-filling, and success is determined by whether the chosen subsample $\mathcal{X}_q^*$ can effectively approximate the design points $\mathcal{S}_q$. The following lemma bounds the difference between $\mathcal{X}_q^*$ and $\mathcal{S}_q$ in terms of the star discrepancy.

Lemma 1. *Suppose that d is fixed and $D^*(\mathcal{S}_q) = O\{q^{-(1-\delta)}\}$ for an arbitrary small $\delta > 0$. If $q = O(n^{1/d})$ as $n \to \infty$, then $D^*(\mathcal{X}_q^*) = O_{\mathrm{p}}\{q^{-(1-\delta)}\}$.*

Lemma 1 says that the selected subsample $\mathcal{X}_q^*$ can effectively approximate the design points $\mathcal{S}_q$ in the sense that the convergence rate of $D^*(\mathcal{X}_q^*)$ is similar to that of $D^*(\mathcal{S}_q)$. The following theorem is the Koksma–Hlawka inequality, which will be used in proving our main theorem; see Kuipers & Niederreiter (2012) for a proof.

Theorem 1 (Koksma–Hlawka inequality). *Let $\mathcal{T}_q = \{t_1, \ldots, t_q\}$ be a set of data points in $[0, 1]^d$, and let h be a function defined on $[0, 1]^d$ with bounded total variation $\mathcal{V}(h)$. We have $\left|\int_{[0,1]^d} h(x)\,\mathrm{d}x - \sum_{i=1}^q h(t_i)/q\right| \leqslant D^*(\mathcal{T}_q)\mathcal{V}(h)$.*

Combining Lemma 1 and Theorem 1 and setting $h = g_{\nu,\mu}$ and $\mathcal{T}_q = \mathcal{X}_q^*$ yields the following lemma.

Lemma 2. *If $q = O(n^{1/d})$, then under Condition 4, for all μ and ν we have*

$$\left|\int_{[0,1]^d} \phi_\nu\phi_\mu\,\mathrm{d}x - \frac{1}{q}\sum_{j=1}^{q} \phi_\nu(x_j^*)\phi_\mu(x_j^*)\right| = O_{\mathrm{p}}\{q^{-(1-\delta)}\}.$$

Lemma 2 demonstrates the superiority of $\{x_i^*\}_{i=1}^q$, the subsample selected by the proposed method, over a randomly selected subsample $\{x_i^+\}_{i=1}^q$. To be specific, as a direct consequence of Condition 3, we have $E\{\int_{[0,1]^d} \phi_\nu\phi_\mu\,\mathrm{d}x - \sum_{j=1}^q \phi_\nu(x_j^+)\phi_\mu(x_j^+)/q\}^2 = O(q^{-1})$ for all μ and ν.

Lemma 2 therefore implies that the subsample $\mathcal{X}_q^*$ can more efficiently approximate the integral $\int_{[0,1]^d} \phi_\nu \phi_\mu \, \mathrm{d}x$ for all μ and ν. We now state our main theoretical result.

THEOREM 2. *Suppose that* $\sum_i \rho_i^p V(\eta_0, \phi_i)^2 < \infty$ *for some* $p \in [1,2]$, *and let* δ *be an arbitrary small positive number. Under Conditions 1–4 and assuming that* $q = O(n^{1/d})$ *and* $q^{1-\delta}\lambda^{1/r} \to \infty$ *as* $\lambda \to 0$, *we have* $(V + \lambda J)(\hat{\eta}_E - \eta_0) = O_{\mathrm{p}}(n^{-1}\lambda^{-1/r} + \lambda^p)$. *In particular, if* $\lambda \asymp n^{-r/(pr+1)}$, *the estimator achieves the optimal convergence rate* $(V + \lambda J)(\hat{\eta}_E - \eta_0) = O_{\mathrm{p}}\{n^{-pr/(pr+1)}\}$.

It is shown in Theorem 9.17 of Gu (2013) that the full-sample smoothing spline estimator $\hat{\eta}$ has convergence rate $(V + \lambda J)(\hat{\eta} - \eta_0) = O_{\mathrm{p}}\{n^{-pr/(pr+1)}\}$ under some regularity conditions. Theorem 2 states that the proposed estimator $\hat{\eta}_E$ achieves the same convergence rate as the full-sample estimator $\hat{\eta}$ under two extra conditions on q: (i) $q = O(n^{1/d})$ and (ii) $q^{1-\delta}\lambda^{1/r} \to \infty$ as $\lambda \to 0$. Moreover, Theorem 2 indicates that in order to achieve the same convergence rate as the full-sample estimator $\hat{\eta}$, the proposed approach requires a much smaller number of basis functions, in the case where $\lambda \asymp n^{-r/(pr+1)}$. The condition $q^{1-\delta}\lambda^{1/r} \to \infty$ indicates that an essential choice of q for the proposed estimator should satisfy $q = O\{n^{(1+\delta)/(pr+1)}\}$ when $\lambda \asymp n^{-r/(pr+1)}$. As a comparison, for both the random basis selection method (Gu & Kim, 2002) and the adaptive basis selection method (Ma et al., 2015), the essential number of basis functions is $q = O\{n^{2/(pr+1)+\delta}\}$. Thus, the proposed estimator is more efficient in that it reduces the order of the essential number of basis functions.

Given $q = O(n^{1/d})$, when $d \leqslant pr + 1$ it follows naturally that $q^{1-\delta}\lambda^{1/r} \to \infty$ is satisfied. On the other hand, when $d > pr + 1$, $q = O(n^{1/d})$ becomes sufficient, but not necessary for $q^{1-\delta}\lambda^{1/r} \to \infty$ to hold. We therefore stress that the essential number of basis functions for the proposed method, $q = O\{n^{(1+\delta)/(pr+1)}\}$, can be achieved only when $d \leqslant pr + 1$. The parameter p in Theorem 2 is closely associated with the true function η_0 and will affect the convergence rate of the proposed estimator. Intuitively, the larger the p is, the smoother the function η_0 will be. For $p \in [1,2]$, the optimal convergence rate of $(V + \lambda J)(\hat{\eta}_E - \eta_0)$ falls in the interval $[O_{\mathrm{p}}(n^{-r/(r+1)}), O_{\mathrm{p}}(n^{-2r/(2r+1)})]$. To the best of our knowledge, the problem of selecting the optimal p has rarely been studied. One exception is the work of Serra & Krivobokova (2017), who studied such a problem in the one-dimensional setting; they proposed a Bayesian approach to selecting an optimal parameter, called β, which is known to be proportional to p. However, because the constant β/p is usually unknown, this approach still cannot be used to select the optimal p in practice. Furthermore, whether such a method can be extended to high-dimensional cases remains unclear.

For the dimension q of the effective model space, a suitable choice is $q = n^{(1+\delta)/(4p+1)+\delta}$ in the following two cases: (I) univariate cubic smoothing splines with penalty $J(\eta) = \int_0^1 (\eta'')^2$, $r = 4$ and $\lambda \asymp n^{-4/(4p+1)}$; (II) tensor-product splines with $r = 4 - \delta^*$ where $\delta^* > 0$. For $p \in [1,2]$, the dimension lies approximately between $O(n^{1/9})$ and $O(n^{1/5})$.

5. SIMULATION RESULTS

To assess the performance of the proposed space-filling basis selection method, we carry out extensive analyses on simulated datasets. We compare the proposed method with uniform basis selection and adaptive basis selection, and report both prediction errors and running times.

The following four functions on $[0,1]$ (Lin & Zhang, 2006) are used as building blocks in our simulation study: $g_1(t) = t$, $g_2(t) = (2t-1)^2$, $g_3(t) = \sin(2\pi t)/\{2 - \sin(2\pi t)\}$ and $g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin(2\pi t)^2 + 0.4\cos(2\pi t)^3 + 0.5\sin(2\pi t)^3$. In addition,

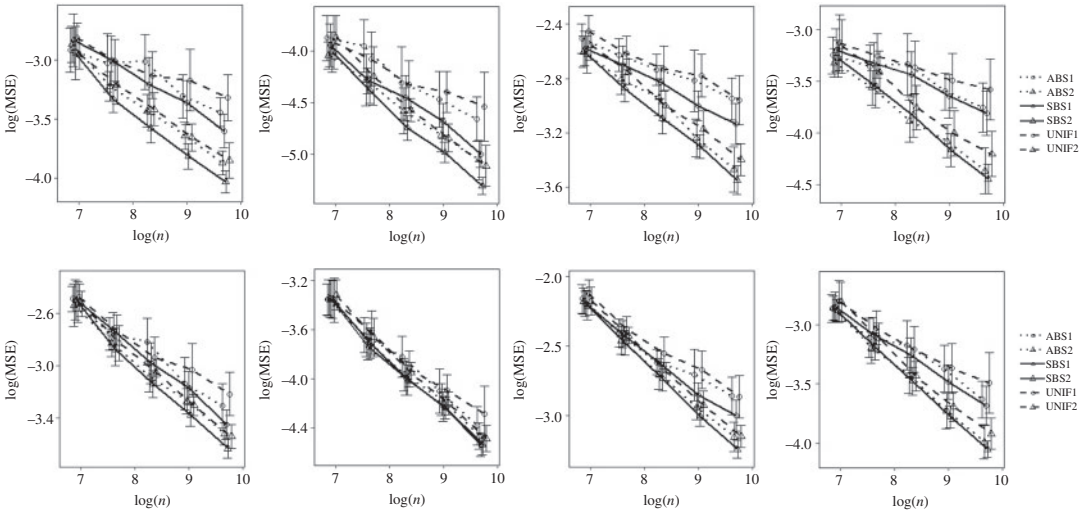

Fig. 3. Simulation under regression settings (i)–(iv) (from left to right), with the signal-to-noise ratio taken to be 5 (upper panels) and 2 (lower panels). The mean squared prediction error (MSE) is plotted against the sample size (n) on the log-log scale, and the vertical bars represent standard errors obtained from 20 replicates. The methods compared are the proposed space-filling basis selection method (solid), the adaptive basis selection method (dotted), and the uniform basis selection method (dashed); lines with triangles and circles correspond to $q = 5n^{2/9}$ and $q = 10n^{1/9}$, respectively.

we use the following two functions on $[0,1]^2$ (Wood, 2003):

$$h_1(t_1,t_2) = \{0.75/(\pi\sigma_1\sigma_2)\} \times \exp\{-(t_1-0.2)^2/\sigma_1^2 - (t_2-0.3)^2/\sigma_2^2\},$$
$$h_2(t_1,t_2) = \{0.45/(\pi\sigma_1\sigma_2)\} \times \exp\{-(t_1-0.7)^2/\sigma_1^2 - (t_2-0.8)^2/\sigma_2^2\},$$

where $\sigma_1 = 0.3$ and $\sigma_2 = 0.4$. The signal-to-noise ratio, defined as $\mathrm{var}\{\eta(X)\}/\sigma^2$, is set at two levels: 5 and 2. We generate replicated samples of sizes $n = \{2^{10}, 2^{11}, \ldots, 2^{14}\}$ and dimensions $d = \{2, 4, 6\}$ uniformly on $[0,1]^p$ from the following four regression settings:

(i) a two-dimensional function $g_1(x_1x_2) + g_2(x_2) + g_3(x_1) + g_4(x_2) + g_3\{(x_1+x_2)/2\}$;
(ii) a two-dimensional function $h_1(x_1,x_2) + h_2(x_1,x_2)$;
(iii) a four-dimensional function $g_1(x_1) + g_2(x_2) + g_3(x_3) + 2g_1\{(x_1+x_4)/2\} + 2g_2\{(x_2+x_3)/2\} + 2g_3\{(x_1+x_3)/2\}$;
(iv) a six-dimensional function $h(x_1,x_2) + h(x_1,x_5)$.

In the simulation, q is set to $5n^{2/9}$ and $10n^{1/9}$ based on the asymptotic results. To combat the curse of dimensionality, we fit smoothing spline analysis of variance models with all main effects and two-way interactions. The prediction error is measured by the mean squared error, defined as $[\sum_{i=1}^{n_0}\{\hat\eta_E(t_i) - \eta_0(t_i)\}^2]/n_0$, where $\{t_i\}_{i=1}^{n_0}$ is an independent testing dataset uniformly generated on $[0,1]^p$ with $n_0 = 5000$. The max projection design (Joseph et al., 2015) is used to generate design points in Step 1 of the proposed method. Our empirical studies suggest that the Sobol sequence and other space-filling techniques, such as the Latin hypercube design (Pukelsheim, 2006) and the uniform design (Fang et al., 2000), also yield similar performance.

Figure 3 plots the mean squared error against the sample size on the log-log scale. The full-sample estimator is omitted because of its high computational cost. The figure shows that the space-filling basis selection method provides more accurate smoothing spline predictions than the other two methods in almost all settings. It can be seen that the lines with circles, $q = 10n^{1/9}$,

Table 1. *Means and standard errors (in parentheses) of the computational time, in CPU seconds, for multivariate cases, based on* 20 *replicates*

True function	SNR	UNIF	ABS	SBS
(i)	5	0.97 (0.15)	0.90 (0.05)	0.90 (0.04)
	2	0.92 (0.10)	0.87 (0.04)	0.87 (0.06)
(ii)	5	0.88 (0.04)	0.87 (0.03)	0.90 (0.06)
	2	0.86 (0.05)	0.85 (0.02)	0.90 (0.06)
(iii)	5	3.92 (0.24)	3.95 (0.24)	4.04 (0.19)
	2	4.08 (0.30)	4.51 (0.66)	4.27 (0.39)
(iv)	5	12.95 (0.61)	15.10 (3.20)	15.45 (3.04)
	2	14.33 (1.44)	13.72 (1.02)	14.25 (1.09)

SNR, signal-to-noise ratio; UNIF, uniform basis selection method; ABS, adaptive basis selection method; SBS, the proposed space-filling basis selection method.

for the space-filling basis selection method display a linear trend, as do the lines with triangles, $q = 5n^{2/9}$, for the other two methods. This indicates that the proposed estimator has a faster convergence rate than the other two methods.

Further simulation results are reported in the Supplementary Material, in which we consider regression functions that exhibit several sharp peaks. In those cases, the results suggest that both the space-filling basis selection method and the adaptive basis selection method outperform the uniform basis selection method, and neither the space-filling basis selection method nor the adaptive basis selection method is superior to the other. Moreover, the proposed space-filling basis selection method outperforms the adaptive basis selection method as the sample size n gets larger.

Table 1 summarizes the computing times for model-fitting with all the methods on a synthetic dataset with $n = 2^{14}$ and $q = 5n^{2/9}$. The simulation is replicated for 20 runs using a computer with an Intel 2.6 GHz processor. The time taken to calculate the smoothing parameter is not included. The result for the full-sample smoothing spline estimator is omitted because of the huge computational cost. The computational time for generating a set of design points, i.e., Step 1 of the proposed algorithm, is not included since the design points can be generated beforehand. One can see that the computing time of the proposed method is comparable to that of the other two basis selection methods under all settings. Combining this observation with the result in Fig. 3 leads to the conclusion that the proposed method can achieve more accurate prediction without requiring much more computational time.

6. REAL-DATA EXAMPLE

The problem of measuring total column ozone has attracted significant attention in the past few decades. Ozone depletion facilitates transmission through the atmosphere of ultraviolet radiation, which can cause severe damage to DNA and cellular proteins involved in biochemical processes, affecting growth and reproduction. Statistical analysis of total column ozone data involves three steps. In the first step, raw satellite data (level 1) are retrieved by NASA, which then calibrates and pre-processes the data to generate spatially and temporally irregular total column ozone measurements (level 2). Finally, the level 2 data are processed to yield level 3 data, which are the daily and spatially regular data released to the public.

We fit the nonparametric model $y_{ij} = \eta(x_{\langle 1\rangle i}, x_{\langle 2\rangle j}) + \epsilon_{ij}$ to a level 2 total column ozone dataset ($n = 173\,405$) compiled by Cressie & Johannesson (2008). Here, y_{ij} is the level 2 total

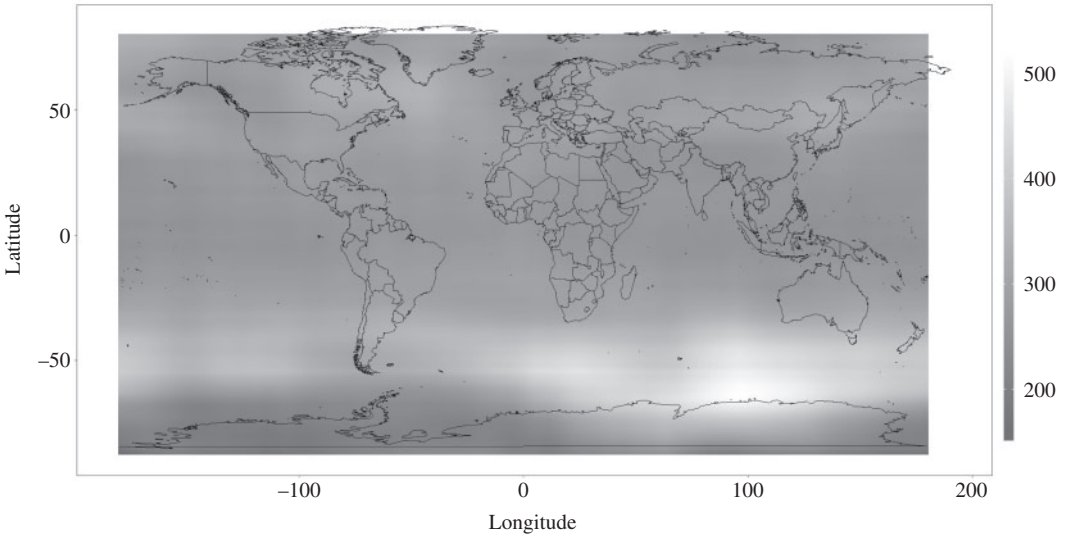

Fig. 4. Smoothing spline prediction of total column ozone value for 1 October 1988, in Dobson units.

column ozone measurement at the ith longitude, $x_{\langle 1\rangle i}$, and jth latitude, $x_{\langle 2\rangle i}$, and ϵ_{ij} represent independent and identically distributed random errors. The heatmap of the raw data is presented in the Supplementary Material. The thin-plate smoothing spline is used for model-fitting, and the proposed method is used for the estimation. The number of basis functions is set to $q = 20n^{2/9} \approx 292$. The design points employed in the proposed basis selection method are obtained from a Sobol sequence (Dutang & Savicky, 2013). The heatmap of the predicted image on a $1^\circ \times 1^\circ$ regular grid is shown in Fig. 4. It is seen that the total column ozone value decreases dramatically to form the ozone hole over the South Pole, around the -55° latitudinal zone.

The computing times, in CPU seconds, on the same computer as for the simulation studies are 0.1 s for basis selection, 129 s for model-fitting and 21 s for prediction. Further results on comparison of the proposed method and other basis selection methods using this dataset can be found in the Supplementary Material.

Acknowledgement

This work was partially supported by the U.S. National Science Foundation and National Institutes of Health.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of the theoretical results, further simulation results, and details of the real-data example.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Statistician* **46**, 175–85.

Arya, S., Mount, D. M., Netanyahu, N., Silverman, R. & Wu, A. Y. (1994). An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. In *Proc. 5th ACM-Society for Industrial and Applied Mathematics Sympos. Discrete Algorithms*. New York: Association for Computing Machinery, pp. 573–82.

Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Commun. ACM* **18**, 509–17.

BONNEEL, N., RABIN, J., PEYRÉ, G. & PFISTER, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *J. Math. Imag. Vis.* **51**, 22–45.
CHUNG, K.-L. (1949). An estimate concerning the Kolmogroff limit distribution. *Trans. Am. Math. Soc.* **67**, 36–50.
CRESSIE, N. & JOHANNESSON, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc.* B **70**, 209–26.
CUTURI, M. & DOUCET, A. (2014). Fast computation of Wasserstein barycenters. *Proc. Mach. Learn. Res.* **32**, 685–93.
DICK, J., KUO, F. Y. & SLOAN, I. H. (2013). High-dimensional integration: The quasi-Monte Carlo way. *Acta Numer.* **22**, 133–288.
DIETTERICH, T. G. (2002). Ensemble learning. In *The Handbook of Brain Theory and Neural Networks*, vol. 2. Cambridge, Massachusetts: MIT Press, pp. 110–25.
DUTANG, C. & SAVICKY, P. (2013). *randtoolbox: Toolbox for Pseudo and Quasi Random Number Generation and Random Generator Tests*. R package version 1.30.1, available at https://rdrr.io/cran/randtoolbox/.
FANG, K.-T., LI, R. & SUDJIANTO, A. (2005). *Design and Modeling for Computer Experiments*. Boca Raton, Florida: CRC Press.
FANG, K.-T., LIN, D. K., WINKER, P. & ZHANG, Y. (2000). Uniform design: Theory and application. *Technometrics* **42**, 237–48.
GLASSERMAN, P. (2013). *Monte Carlo Methods in Financial Engineering*. New York: Springer.
GU, C. (2013). *Smoothing Spline ANOVA Models*. New York: Springer.
GU, C. & KIM, Y.-J. (2002). Penalized likelihood regression: General formulation and efficient approximation. *Can. J. Statist.* **30**, 619–28.
HALTON, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numer. Math.* **2**, 84–90.
HASTIE, T. (1996). Pseudosplines. *J. R. Statist. Soc.* B **58**, 379–96.
JOSEPH, V. R., GUL, E. & BA, S. (2015). Maximum projection designs for computer experiments. *Biometrika* **102**, 371–80.
KIM, Y.-J. & GU, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. R. Statist. Soc.* B **66**, 337–56.
KOLOURI, S., POPE, P. E., MARTIN, C. E. & ROHDE, G. K. (2018). Sliced-Wasserstein autoencoder: An embarrassingly simple generative model. *arXiv:* 1804.01947v3.
KUIPERS, L. & NIEDERREITER, H. (2012). *Uniform Distribution of Sequences*. Mineola, New York: Dover Publications.
KUNCHEVA, L. I. & WHITAKER, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.* **51**, 181–207.
LEMIEUX, C. (2009). *Monte Carlo and Quasi-Monte Carlo Sampling*. New York: Springer.
LEOBACHER, G. & PILLICHSHAMMER, F. (2014). *Introduction to Quasi-Monte Carlo Integration and Applications*. Cham, Switzerland: Springer.
LIN, Y. & ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272–97.
LUO, Z. & WAHBA, G. (1997). Hybrid adaptive splines. *J. Am. Statist. Assoc.* **92**, 107–16.
MA, P., HUANG, J. Z. & ZHANG, N. (2015). Efficient computation of smoothing splines via adaptive basis sampling. *Biometrika* **102**, 631–45.
MENG, C., KE, Y., ZHANG, J., ZHANG, M., ZHONG, W. & MA, P. (2019). Large-scale optimal transport map estimation using projection pursuit. *Advances in Neural Information Processing Systems*, pp. 8116–27.
OWEN, A. B. (2003). Quasi-Monte Carlo sampling. In *SIGGRAPH: Monte Carlo Ray Tracing*, vol. 1. New York: Association for Computing Machinery, pp. 69–88.
PUKELSHEIM, F. (2006). *Optimal Design of Experiments*. Philadelphia: Society for Industrial and Applied Mathematics.
R DEVELOPMENT CORE TEAM (2020). *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.
RABIN, J., PEYRÉ, G., DELON, J. & BERNOT, M. (2011). Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision (3rd Int. Conf., SSVM 2011)*. Heidelberg: Springer, pp. 435–46.
REINSCH, C. H. (1967). Smoothing by spline functions. *Numer. Math.* **10**, 177–83.
ROKACH, L. (2010). Ensemble-based classifiers. *Artif. Intel. Rev.* **33**, 1–39.
RUPPERT, D. (2002). Selecting the number of knots for penalized splines. *J. Comp. Graph. Statist.* **11**, 735–57.
RUPPERT, D., WAND, M. P. & CARROLL, R. J. (2009). Semiparametric regression during 2003–2007. *Electron. J. Statist.* **3**, 1193–256.
SERRA, P. & KRIVOBOKOVA, T. (2017). Adaptive empirical Bayesian smoothing splines. *Bayesian Anal.* **12**, 219–38.
SOBOL, I. M. (1967). The distribution of points in a cube and the approximate evaluation of integrals. *USSR Comp. Math. Math. Phys.* **7**, 86–112.
VILLANI, C. (2008). *Optimal Transport: Old and New*. New York: Springer.
WAHBA, G. (1990). *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

WAHBA, G. & CRAVEN, P. (1978). Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377–404.
WALD, I. & HAVRAN, V. (2006). On building fast kd-trees for ray tracing, and on doing that in $O(N \log N)$. In *2006 IEEE Sympos. Interactive Ray Tracing*. New York: Institute of Electrical and Electronics Engineers, pp. 61–9.
WANG, X., SHEN, J. & RUPPERT, D. (2011). On the asymptotics of penalized spline smoothing. *Electron. J. Statist.* **5**, 1–17.
WOOD, S. N. (2003). Thin plate regression splines. *J. R. Statist. Soc.* B **65**, 95–114.
WU, C. F. J. & HAMADA, M. S. (2011). *Experiments: Planning, Analysis, and Optimization*. Hoboken, New Jersey: John Wiley & Sons.
ZHANG, H. H., WAHBA, G., LIN, Y., VOELKER, M., FERRIS, M., KLEIN, R. & KLEIN, B. (2004). Variable selection and model building via likelihood basis pursuit. *J. Am. Statist. Assoc.* **99**, 659–72.

[*Received on* 22 *October* 2018. *Editorial decision on* 17 *October* 2019]