

# 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 of predictors  $d \geq 2$ . Such a sizable 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 identical rate when  $q$  is of the order  $O\{n^{2/(pr+1)}\}$ , where  $p \in [1, 2]$  depends on the true function  $\eta$ , and  $r > 1$  depends on the type of spline. Such  $q$  is called the essential number of basis functions. In this article, we develop a more efficient basis selection method. By selecting the ones corresponding to roughly equal-spaced observations, the proposed method chooses a set of basis functions with a large diversity. The asymptotic analysis shows our proposed smoothing spline estimator can decrease  $q$  to roughly  $O\{n^{1/(pr+1)}\}$ , when  $d \leq pr + 1$ . Applications on synthetic and real-world datasets show the proposed method leads to a smaller prediction error compared with other basis selection methods.

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

## 1. INTRODUCTION

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

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

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

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  $d \geq 2$ . Such a 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 of thinking, numerous studies 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 developed to approximate smoothing splines.

Although these methods have already yielded impressive algorithmic benefits, they are usually *ad hoc* in 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 in order to ensure the restricted estimator  $\hat{\eta}_E$  converge to the true function  $\eta$  at the identical rate as the full sample estimator  $\hat{\eta}$ . To answer this question, Gu & Kim (2002), Ma et al. (2015) developed random sampling methods for selecting the basis functions and established the coherent theory for the convergence of the restricted estimator  $\hat{\eta}_E$ . To ensure that  $\hat{\eta}_E$  has the same convergence rate as  $\hat{\eta}$ , both methods in Gu & Kim (2002) and Ma et al. (2015) require  $q$  be of the order  $O\{n^{2/(pr+1)+\delta}\}$ , where  $\delta$  is an arbitrary small positive number,  $p \in [1, 2]$  depends on the true function  $\eta$ , and  $r$  depends on the fitted spline. In Gao (1999), it is shown that fewer basis functions are needed to warrant the aforementioned convergence rate if we select the basis functions  $\{R(z_j, \cdot)\}_{j=1}^q$ , where  $\{z_j\}_{j=1}^q$  are roughly equal-spaced. However, they only provide the theory in the univariate predictor case, and their method cannot be directly applied to multivariate cases.

In this paper, we develop a more efficient computational method to approximate smoothing splines. The distinguishing feature of the proposed method is that it considers the notion of diversity of the selected basis functions. We propose the space-filling basis selection method, which chooses the basis functions with a large diversity, by selecting the ones that correspond to roughly uniformly-distributed observations. When  $d \leq pr + 1$ , we show that the smoothing spline estimator proposed here has the same convergence rate as the full sample estimator, and the order of the essential number of basis function  $q$  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 semi-norm. 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  $\{\xi_i\}_{i=1}^m$  in which  $m$  is the dimension of  $\mathcal{N}_J$ . Let  $\mathcal{H}_J$  be the orthogonal complement of  $\mathcal{N}_J$  in  $\mathcal{H}$  such 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  $d = (d_1, \dots, d_m)^T \in \mathbb{R}^m$  and  $c = (c_1, \dots, c_n)^T \in \mathbb{R}^n$ , such that the minimizer of (1) in  $\mathcal{H}$  is given by  $\eta(x) = \sum_{k=1}^m d_k \xi_k(x) + \sum_{i=1}^n c_i R_J(x_i, x)$ . Let  $Y = (y_1, \dots, y_n)^T$  be the vector of response observations,  $S$  be the  $n \times m$  matrix with the  $(i, j)$ th entry  $\xi_j(x_i)$ , and  $R$  be the  $n \times n$  matrix with the  $(i, j)$ th entry  $R_J(x_i, x_j)$ . Solving the minimization problem in (1) thus is equivalent to solving

$$(\hat{d}, \hat{c}) = \operatorname{argmin}_{d, c} \frac{1}{n} (Y - Sd - Rc)^T (Y - Sd - Rc) + \lambda c^T Rc, \quad (2)$$

where the smoothing parameter  $\lambda$  can be selected based on the generalized cross-validation criterion (Wahba & Craven 1978). In a general case where  $n \gg m$  and  $d \geq 2$ , the computation cost for calculating  $(\hat{d}, \hat{c})$  in equation (2) is of the order  $O(n^3)$ , which is prohibitive when the sample size  $n$  is considerable. To reduce this computational burden, one can restrict the full sample estimator  $\hat{\eta}$  to a subspace  $\mathcal{H}_E \subset \mathcal{H}$ , where  $\mathcal{H}_E = \mathcal{N}_J \oplus \text{span}\{R_J(x_i^*, \cdot), i = 1, \dots, q\}$ . Here,  $\mathcal{H}_E$ , termed as 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 thus called the basis selection method.

Denote a matrix  $R_* \in \mathbb{R}^{n \times q}$  with the  $(i, j)$ th entry  $R_J(x_i, x_j^*)$  and  $R_{**} \in \mathbb{R}^{q \times q}$  with the  $(i, j)$ th entry  $R_J(x_i^*, x_j^*)$ . The minimizer of (1) in the effective model space  $\mathcal{H}_E$  thus can 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,  $d_E = (d_1, \dots, d_m)^T$  and  $c_E = (c_1, \dots, c_q)^T$  can be obtained through solving

$$(\hat{d}_E, \hat{c}_E) = \underset{d_E, c_E}{\operatorname{argmin}} \frac{1}{n} (Y - Sd_E - R_*c_E)^T (Y - Sd_E - R_*c_E) + \lambda c_E^T R_{**} c_E. \quad (3)$$

The evaluation of the restricted estimator  $\hat{\eta}_E$  at sample points thus satisfies  $\hat{\eta}_E = S\hat{d}_E + R_*\hat{c}_E$ , where  $\hat{\eta}_E = \{\hat{\eta}_E(x_1), \dots, \hat{\eta}_E(x_n)\}^T$ . In a general case where  $m \ll q \ll n$ , the overall computational cost for the basis selection method is  $O(nq^2)$ , which is a significant reduction compared with  $O(n^3)$ . 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  $\eta$  at the same rate as  $\hat{\eta}$ , researchers showed that the essential number of basis functions  $q$  needs to be of the order  $O\{n^{2/(pr+1)+\delta}\}$ , where  $\delta$  is an arbitrary small positive number (Kim & Gu 2004, Ma et al. 2015). In the next section, we present the proposed space-filling basis selection method, which reduces such an order to  $O\{n^{(1+\delta)/(pr+1)}\}$ .

### 3. SPACE-FILLING BASIS SELECTION

#### 3.1. Motivation and Notations

To motivate the development of the proposed method, we first re-examine the ensemble learning methods, which are well-known in statistics and machine learning (Dietterich 2002, Rokach 2010). To achieve better predictive performance than a single learner<sup>1</sup>, ensemble learning methods first build a committee which consists of a number of different learners, then aggregate the predictions of these learners in the committee. The aggregation is usually achieved by employing the majority vote or by calculating a weighted average. The diversity among the learners in the committee holds the key to the success of the ensemble learning methods. A large diversity in the committee yields a better performance of ensemble learning methods (Kuncheva & Whitaker 2003).

The restricted smoothing spline estimator  $\hat{\eta}_E$  can be considered as an ensemble learning method. In particular, the prediction of  $\hat{\eta}_E$  is conducted by taking a weighted average of the predictions of the selected basis functions  $R_J(x_i^*, \cdot)$ ,  $i \in \{1, \dots, q\}$ , 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 among a set of basis functions. Notice that adjacent data points, i.e.,  $x_i^* \approx x_j^*$  ( $i, j \in \{1, \dots, q\}$ ) yields 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 inspire 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

<sup>1</sup> A learner is either a model or a learning algorithm.

possible. The uniformly-distributed property, usually termed 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 the data with 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  $\mathcal{D} \subset \mathbb{R}^d$ . Suppose  $\tau$  is a transformation, such that  $\{\tau(x_i)\}_{i=1}^n$  has the uniform distribution on  $[0, 1]^d$ . In a simple case where  $d = 1$  and  $F$  is known, we can find the transformation  $\tau$  by setting  $\tau = F$ . In a more general case where  $d > 1$  and  $F$  is unknown, the transformation  $\tau$  can be calculated via the optimal transport theory (Villani 2008). However, finding the exact solution via the optimal transport theory can be time-consuming. Instead, one may approximate the transformation  $\tau$  using the iterative transformation approach (Pukelsheim 2006) or the sliced optimal transport map approach (Rabin et al. 2011). The computational cost of these two approaches is of the order  $O\{Kn \log(n)\}$ , where  $K$  denotes the number of iterations (Kolouri et al. 2018, Cuturi & Doucet 2014, Bonneel et al. 2015). Such a computational cost is negligible compared with the computational cost of the proposed method. In practice, the data can always be preprocessed using the transformation of  $\tau$ . Without loss of generality,  $\mathcal{X}_n$  may be assumed 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  $a = (a_1, \dots, a_d)^T \in [0, 1]^d$ ,  $[0, a) = \prod_{j=1}^d [0, a_j)$  be a hyper-rectangle and  $1\{\cdot\}$  be the indicator function. The local discrepancy and the star discrepancy are defined as follows (Fang et al. 2005, Pukelsheim 2006).

**DEFINITION 1.** Given  $\mathcal{X}_q = \{x_1, \dots, x_q\}$  in  $[0, 1]^d$  and a hyper-rectangle  $[0, a)$ , the corresponding local discrepancy is defined as  $D(\mathcal{X}_q, a) = |\frac{1}{q} \sum_{i=1}^q 1\{x_i \in [0, a)\} - \prod_{j=1}^d a_j|$ . 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 the data points located in  $[0, a)$ . The local discrepancy is illustrated in the left panel of Fig. 1. The star discrepancy  $D^*(\mathcal{X}_q)$  calculates the supreme of all the local discrepancy over  $a \in [0, 1]^d$ . In other words, the smaller the  $D^*(\mathcal{X}_q)$  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 in  $[0, 1]^d$ ,  $D^*(\mathcal{X}_q)$  converges to 0 with the order of convergence  $O[\{\log \log(q)/q\}^{1/2}]$ . Faster convergence rate can be achieved using space-filling design methods (Pukelsheim 2006) and the low-discrepancy sequence method (Halton 1960, Sobol 1967, Owen 2003). The space-filling design methods, developed in the experimental design literature, aim to generate a set of design points that can cover the space as uniformly as possible. For further details, please refer to Wu & Hamada (2011), Fang et al. (2005), Pukelsheim (2006). The low-discrepancy sequence method Such a method is frequently applied in the field of 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 the order  $O\{\log(q)^d/q\}$ , which is roughly the square order  $D^*(\mathcal{X}_q)$  for fixed  $d$ . For more in-depth discussions on the quasi-Monte Carlo methods, see e.g., Lemieux (2009), Leobacher & Pillichshammer (2014), Dick et al. (2013) or Chapter 5 in Glasserman (2013) and references therein.

Existing studies suggested that 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 select the space-filling basis functions in smooth-

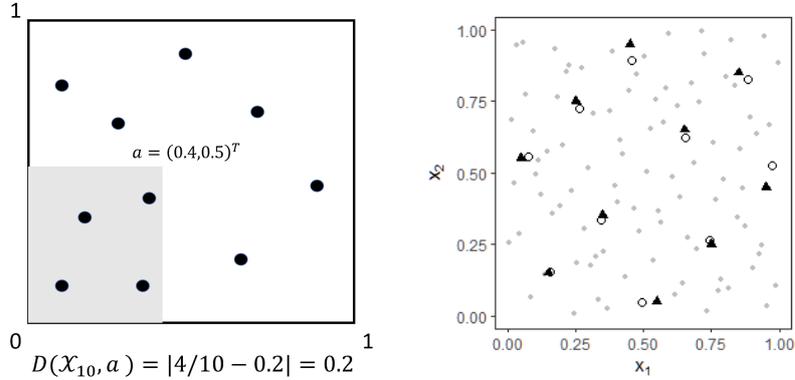


Fig. 1: Left panel: A toy example for local discrepancy. A set of ten data points are generated in  $[0, 1]^2$ , and four of them locate in the rectangular  $[0, a)$ , where  $a = (0.4, 0.5)^T$ . The local discrepancy is  $|4/10 - 0.4 \times 0.5| = 0.2$ . Right panel: An illustration for the proposed basis selection method. The data points are labelled as gray dots. The nearest neighbor data point for each design point (black triangle) is labeled as a circle.

ing splines. Unfortunately, the existing techniques of space-filling design cannot be applied to our basis selection problem directly due to the following fact. The design space in space-filling design methods is usually continuous, whereas our sample space  $\{x_i\}_{i=1}^n$  is finite and discrete. We propose an algorithm to overcome the barrier.

### 3.3. Main algorithm

We shall develop a space-filling basis selection method, in which we select the space-filling data points in a computationally-attractive manner. First, a set of design points  $\mathcal{S}_q = \{s_i\}_{i=1}^q \in [0, 1]^d$  are generated, either using low-discrepancy sequence or space-filling design methods. Subsequently, the nearest neighbor  $x_i^*$  is selected for each  $s_i$ , from the sample points  $\{x_i\}_{i=1}^n$ . Thus,  $\{x_i^*\}_{i=1}^q$  can approximate the design points  $\mathcal{S}_q$  well, if  $x_i^*$  is sufficiently close to  $s_i$ , for  $i = 1, \dots, q$ . The proposed 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 neighbor for 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 following effective model space  $\mathcal{H}_E = \mathcal{N}_J \oplus \text{span}\{R_J(x_i^*, \cdot), i = 1, \dots, q\}$

The proposed algorithm is illustrated through a toy example in the right panel of Fig. 1. One hundred data points (gray dots) are generated from the uniform distribution in  $[0, 1]^2$ , and a set of design points (black triangles) are generated through the max projection design (Joseph et al. 2015), a recently developed space-filling design method. The nearest neighbor to each design point is selected (circle). It is observed 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 (Gu & Kim 2002) and the adaptive basis selection method (Ma et al. 2015) using a two-dimensional toy example. We generate 5000 data points from the uniform distribution in  $[0, 1]^2$ . The leftmost panel in Fig. 2 presents the heat map for the true response

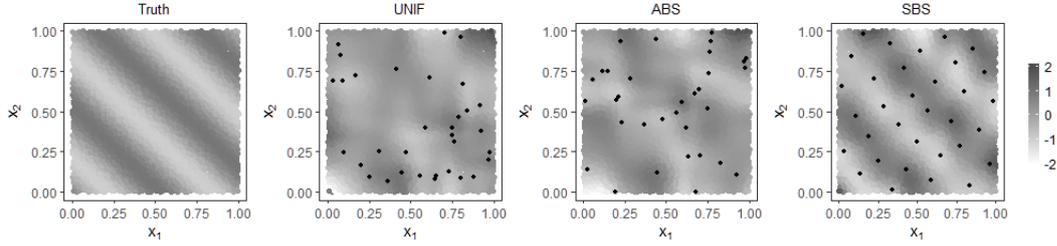


Fig. 2: Comparison of different basis selection methods. The leftmost panel shows the heat map for the true function. The heat maps for the spline estimates based on the uniform basis selection method (UNIF), the adaptive basis selection method (ABS), and the proposed space-filling basis selection method (SBS) are presented in the other three panels, respectively. Black dots are the sampled basis functions. We observe that the proposed method outperforms the other methods in approximating the true function.

surface  $y = \sin\{5(x_1 + x_2)\}$ . The dimension of the effective model space  $q$  is set to be  $5 \times (5000)^{2/9} \approx 33$ , for all basis selection methods. The selected basis functions are labeled as solid dots in each panel. The right three panels of Fig. 2 plot the heat maps of the spline estimates of all three basis selection methods. In the uniform basis selection method, the default random number generator in R is employed to select the basis functions. It is observed that the selected points are not uniformly distributed. This is a very common phenomenon for uniform basis selection since the randomly selected points do not necessarily look uniformly-distributed, especially when the number of selected points is small. In contrast, it is observed that the basis functions selected by the proposed method are space-filling. Using the space-filling design techniques, the proposed method overcomes the pitfall of uniform basis selection method and uniformly-distribute the selected points. The true response can be better estimated using the proposed method than using other methods.

Now we calculate the computational cost of the proposed method. In Step 1, the design points can be generated beforehand; thus, the computational cost in Step 1 can be ignored. For the nearest neighbor 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 searching results, e.g., using those approximate nearest neighbor searching algorithms (Altman 1992, Arya et al. 1994). For Step 3, computing the smoothing spline estimates in the restricted space  $\mathcal{H}_E$  is of the order  $O(nq^2)$ , as discussed in Section 2.2. In summary, the overall computational cost for the space-filling basis selection method is of the order  $O(nq^2)$ .

#### 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  $f_X(\cdot) = 1$ . We define that  $V(g) = \int_{[0,1]^d} g^2 dx$ . The following four regularity conditions are required for the asymptotic analysis, and the first three are the identical conditions employed by Ma et al. (2015), in which 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  $\nu$ ;

Condition 3. for all  $\mu$  and  $\nu$ ,  $\text{var}\{\phi_\nu(x)\phi_\mu(x)\} \leq C_1$ , where  $\phi_\nu, \phi_\mu$  are the eigenfunctions associated with  $V$  and  $J$  in  $\mathcal{H}$ ,  $C_1$  denotes a positive constant; 220

Condition 4. for all  $\mu$  and  $\nu$ ,  $\mathcal{V}(g_{\nu,\mu}) \leq C_2$ , 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. The total variation here is defined in the sense of Hardy and Krause (Owen 2003). As a specific case when  $d = 1$ , the total variation  $\mathcal{V}(g) = \int |g'(x)|dx$ , revealing that a smooth function displays a small total variation. Intuitively, 225  
the total variation measures how wiggly the function  $g$  is.

Condition 1 indicates that there exist a sequence of eigenfunctions  $\phi_\nu \in \mathcal{H}$  and the associated sequence of 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  $\rho_\nu$  dictates how fast  $\lambda$  should approach to 0, and further what the convergence rate of smoothing spline estimates is (Gu 2013). Notice 230  
that the eigenfunctions  $\phi_\nu$ s have a close relationship with the Demmler-Reinsch basis, which are orthogonal vectors representing  $l_2$  norm (Ruppert 2002). The eigenfunctions  $\phi_\nu$ s can be calculated explicitly in several specific scenarios. For instance,  $\phi_\nu$ s are the sine and cosine functions when  $J(\eta) = \int_0^1 (\eta'')^2 dx$ , where  $\eta$  denotes a periodic function on  $[0, 1]$ . For more details on the construction of  $\phi_\nu$ s can be found in Section 9.1 of Gu (2013). 235

We now present our main theoretical results, and all the proofs are relegated to the Supplementary Material. For a set of design points  $\mathcal{S}_q$  of size  $q$ , we now assume the star discrepancy  $D^*(\mathcal{S}_q)$  converges to zero at the rate of  $O\{\log(q)^d/q\}$ , or  $O\{q^{-(1-\delta)}\}$  for an arbitrary small positive number  $\delta$ . Such a convergence rate is warranted if  $\mathcal{S}_q$  is generated from a low-discrepancy sequence or space-filling design methods, as discussed in Section 3.1. Recall that the proposed method 240  
aims to select a subsample that is space-filling, and the key to success depends on 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 star discrepancy.

LEMMA 1. Suppose  $d$  is fixed and  $D^*(\mathcal{S}_q) = O\{q^{-(1-\delta)}\}$ , for any arbitrary small  $\delta > 0$ . If  $q = O(n^{1/d})$ , as  $n \rightarrow \infty$ , we have  $D^*(\mathcal{X}_q^*) = O_p\{q^{-(1-\delta)}\}$ . 245

Lemma 1 states 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  $\mathcal{T}_q = \{t_1, \dots, t_q\}$  be a set of data 250  
points in  $[0, 1]^d$ , and  $h$  be a function defined on  $[0, 1]^d$  with bounded total variation  $\mathcal{V}(h)$ . We have  $|\int_{[0,1]^d} h(x)dx - \sum_{i=1}^q h(t_i)/q| \leq D^*(\mathcal{T}_q)\mathcal{V}(h)$ .

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

LEMMA 2. If  $q = O(n^{1/d})$ , under Condition 4, for all  $\mu$  and  $\nu$ , we have

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

Lemma 2 shows the advantage of  $\{x_i^*\}_{i=1}^q$ , the subsample selected by the proposed method, 255  
over a randomly selected subsample  $\{x_i^+\}_{i=1}^q$ . To be specific, as a direct consequence of Condition 3, we have  $E[\int_{[0,1]^d} \phi_\nu \phi_\mu dx - \sum_{j=1}^q \phi_\nu(x_j^+) \phi_\mu(x_j^+)/q]^2 = O(q^{-1})$ , for all  $\mu$  and  $\nu$ . Lemma 2 therefore implies the subsample  $\mathcal{X}_q^*$  can more efficiently approximate the integration  $\int_{[0,1]^d} \phi_\nu \phi_\mu dx$ , for all  $\mu$  and  $\nu$ . We now present our main theoretical result.

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

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

Given  $q = O(n^{1/d})$ , when  $d \leq pr + 1$ , it follows naturally when  $q^{1-\delta} \lambda^{1/r} \rightarrow \infty$  is satisfied. Otherwise, when  $d > pr + 1$ ,  $q = O(n^{1/d})$  becomes sufficient but not necessary for  $q^{1-\delta} \lambda^{1/r} \rightarrow \infty$ . We thus stress that the essential number of basis functions for the proposed method,  $q = O\{n^{(1+\delta)/(pr+1)}\}$ , can only be achieved when  $d \leq pr + 1$ . The parameter  $p$  in Theorem 2 is closely associated with the true function  $\eta_0$  and will affect the convergence rate of the proposed estimator. Intuitively, the larger the  $p$  is, the smoother the function  $\eta_0$  will be. For  $p \in [1, 2]$ , the optimal convergence rate of  $(V + \lambda J)(\hat{\eta}_E - \eta_0)$  falls in the interval  $[O_p(n^{-r/(r+1)}), O_p(n^{-2r/(2r+1)})]$ . To the best of our knowledge, the problem of selecting the optimal  $p$  has rarely been studied, and one exception is Serra & Krivobokova (2017), where the author studied such a problem in one-dimensional cases. Serra & Krivobokova (2017) provided a Bayesian approach for selecting the optimal parameter, named  $\beta$ , which is known to be proportional to  $p$ . Nevertheless, since the constant  $\beta/p$  is usually unknown, such an approach still cannot be used to select the optimal  $p$  in practice. Furthermore, whether such an approach can be extended to the high-dimensional cases remains unclear.

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

## 5. SIMULATION RESULTS

295 To assess the performance of the proposed space-filling basis selection method, we carry out extensive analysis on simulated datasets. We report some of them in this section. We compare the proposed method with uniform basis selection and adaptive basis selection, and report both prediction error and running time.

The following four functions on  $[0, 1]$  (Lin & Zhang 2006) are used as the 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 addi-

tion, we also use the following functions on  $[0, 1]^2$  (Wood 2003) as the building blocks,

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

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

where  $\sigma_1 = 0.3$  and  $\sigma_2 = 0.4$ . The signal-to-noise ratio (SNR), defined as  $\text{var}\{\eta(X)\}/\sigma^2$ , is set to be at two levels: 5 and 2. We generate replicated samples with sample sizes  $n = \{2^{10}, 2^{11}, \dots, 2^{14}\}$  and dimensions  $d = \{2, 4, 6\}$  uniformly on  $[0, 1]^p$  from the following four regression settings, 305

- (1) A 2-d function  $g_1(x_1x_2) + g_2(x_2) + g_3(x_1) + g_4(x_2) + g_3\{(x_1 + x_2)/2\}$ ;
  - (2) A 2-d function  $h_1(x_1, x_2) + h_2(x_1, x_2)$ ;
  - (3) A 4-d function  $g_1(x_1) + g_2(x_2) + g_3(x_3) + 2g_1\{(x_1 + x_4)/2\} + 2g_2\{(x_2 + x_3)/2\} + 2g_3\{(x_1 + x_3)/2\}$ ;
  - (4) A 6-d function  $h(x_1, x_2) + h(x_1, x_5)$ .
- 310

In the simulation,  $q$  is set to be  $5n^{2/9}$  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 (MSE), defined as  $[\sum_{i=1}^{n_0} \{\hat{\eta}_E(t_i) - \eta_0(t_i)\}^2]/n_0$ , where  $\{t_i\}_{i=1}^{n_0}$  denotes 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, e.g., the Latin hypercube design (Pukelsheim 2006) and the uniform design (Fang et al. 2000), also yield similar performance. 315

Figure 3 shows the MSE against the sample size on the log-log scale. Each column presents the results of a function setting as described above. We set the signal-to-noise ratio to be five and two in the upper row and the lower row, respectively. We use solid lines for the proposed method, dotted lines for adaptive basis selection method, and dashed lines for uniform basis selection method. The number of the basis functions  $q$  is  $5n^{2/9}$  and  $10n^{1/9}$  for the lines with triangles and the lines with circles, respectively. The vertical bars represent standard error bars obtained from 20 replicates. The full sample estimator is omitted here due to the high computation cost. It is observed that the space-filling basis selection method provides more accurate smoothing spline predictions than the other two methods in almost all settings. Notably, the lines with circles for the space-filling basis selection method displays a linear trend similar to the lines with triangles for the other two methods. This observation reveals the proposed estimator yields a faster convergence rate than the other two methods. 320

More simulation results can be found in the Supplementary Material, in which we consider the 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, whereas neither the space-filling basis selection method nor the adaptive basis selection method dominates each other. Moreover, the proposed space-filling basis selection method outperforms the adaptive basis selection method as the sample size  $n$  gets larger. 325

Table 1 summarizes the computing times of model-fitting using all 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. In Table 1, UNIF, ABS, and SBS represent the uniform basis selection method, the adaptive basis selection method, and the proposed space-filling basis selection method, respectively. The time for calculating the smoothing parameter is not included. The result for the full basis smoothing spline estimator is omitted here due to the huge computational cost. The computational time for generating a set of design points, i.e., Step 1 in the proposed 330

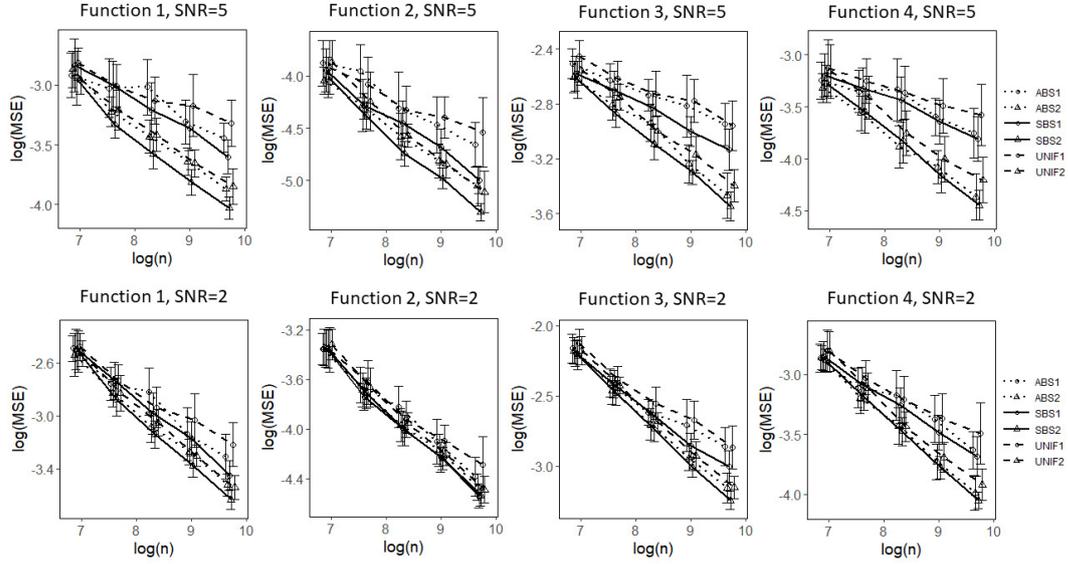


Fig. 3: Simulation under different settings (from left to right) with SNR being five (the upper row) and two (the lower row). The prediction errors are plotted versus sample size  $n$ . The vertical bars are standard error bars obtained from 20 replicates. The solid lines, dotted lines, and dashed lines refer to the space-filling basis selection (SBS), the adaptive basis selection (ABS), and the uniform basis selection (UNIF), respectively. The lines with triangles and those with circles represent  $q = 5n^{2/9}$  and  $q = 10n^{1/9}$ , respectively.

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
Function 1	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)
Function 2	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)
Function 3	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)
Function 4	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)

algorithm, is not included since the design points can be generated beforehand. It is observed that the computing time of the proposed method is comparable with that of the other two basis selection methods under all settings. Combining such an observation with the result in Fig. 3, it is concluded that the proposed method can achieve a more accurate prediction, without requiring much more computational time.

## 6. REAL DATA EXAMPLE

The problem of measuring total column ozone has attracted significant attention for decades. Ozone depletion facilitates the transmission of ultraviolet radiation (290400 nm wavelength) through the atmosphere and causes severe damage to DNA and cellular proteins that are involved in biochemical processes, affecting growth and reproduction. Statistical analysis of total column ozone data has three steps. In the first step, the raw satellite data (level 1) are retrieved by NASA. Subsequently, NASA calibrates and preprocesses the data to generate spatially and temporally irregular total column ozone measurements (level 2). Finally, the level 2 data are processed to yield the level 3 data, which are the daily and spatially regular data product released extensively 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 column ozone measurement at the  $i$ th longitude, i.e.,  $x_{\langle 1 \rangle i}$ , and the  $j$ th latitude, i.e.,  $x_{\langle 2 \rangle j}$ , and  $\epsilon_{ij}$  represent the independent and identically distributed random errors. The heat map of the raw data is presented in the Supplementary Material. The thin-plate smoothing spline is used for the model-fitting, and the proposed method is employed to facilitate 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 yielded from a Sobol sequence (Dutang & Savicky 2013). The heat map of the predicted image on a  $1^\circ \times 1^\circ$  regular grid is presented in Fig. 4. It is seen that the total column ozone value decreases dramatically to form the ozone hole over the South Pole, around  $-55^\circ$  latitudinal zone.

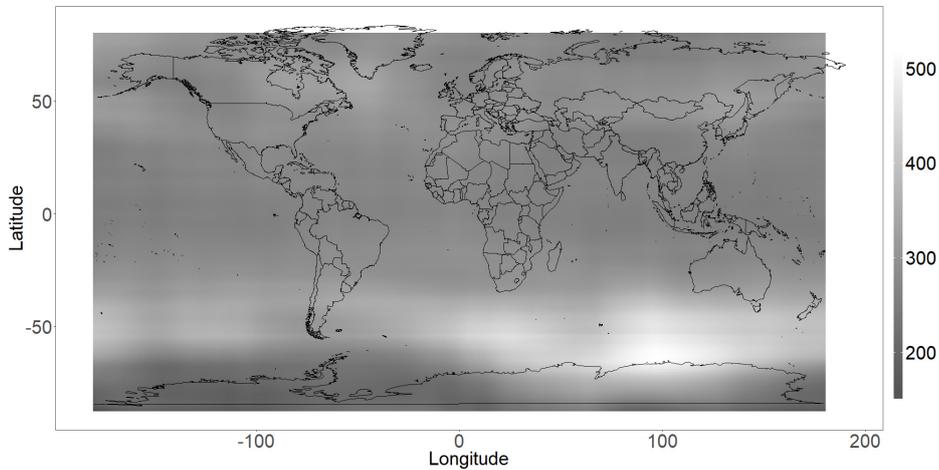


Fig. 4: Smoothing spline prediction of total column ozone value for 10/01/1988, in Dobson units

We now report computing times of the model-fitting that are performed on the identical laptop computer for the simulation studies. The computational times, in CPU seconds, are presented in parentheses, including basis selection (0.1s), model fitting (129s), and prediction (21s). Further comparison between the proposed method and other basis selection methods on this dataset can be found in the Supplementary Material.

## 7. ACKNOWLEDGEMENT

380 This study was partially supported by the National Science Foundation and the National Institutes of Health.

## REFERENCES

- Altman, N. S. (1992), ‘An introduction to kernel and nearest-neighbor nonparametric regression’, *The American Statistician* **46**(3), 175–185.
- 385 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-SIAM Sympos. Discrete Algorithms’, pp. 573–582.
- Bentley, J. L. (1975), ‘Multidimensional binary search trees used for associative searching’, *Communications of the ACM* **18**(9), 509–517.
- 390 Bonneel, N., Rabin, J., Peyré, G. & Pfister, H. (2015), ‘Sliced and radon wasserstein barycenters of measures’, *Journal of Mathematical Imaging and Vision* **51**(1), 22–45.
- Chung, K.-L. (1949), ‘An estimate concerning the kolmogoroff limit distribution’, *Transactions of the American Mathematical Society* **67**(1), 36–50.
- Cressie, N. & Johannesson, G. (2008), ‘Fixed rank kriging for very large spatial data sets’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 209–226.
- 395 Cuturi, M. & Doucet, A. (2014), Fast computation of wasserstein barycenters, in ‘International Conference on Machine Learning’, pp. 685–693.
- Dick, J., Kuo, F. Y. & Sloan, I. H. (2013), ‘High-dimensional integration: the quasi-monte carlo way’, *Acta Numerica* **22**, 133–288.
- 400 Dietterich, T. G. (2002), ‘Ensemble learning’, *The handbook of brain theory and neural networks* **2**, 110–125.
- Dutang, C. & Savicky, P. (2013), ‘randtoolbox: Generating and testing random numbers’, *R package* p. 67.
- Fang, K.-T., Li, R. & Sudjianto, A. (2005), *Design and Modeling for Computer Experiments*, CRC Press.
- Fang, K.-T., Lin, D. K., Winker, P. & Zhang, Y. (2000), ‘Uniform design: Theory and application’, *Technometrics* **42**(3), 237–248.
- 405 Gao, F. (1999), ‘Penalized multivariate logistic regression with a large data set’.
- Glasserman, P. (2013), *Monte Carlo methods in financial engineering*, Vol. 53, Springer Science & Business Media.
- Gu, C. (2013), *Smoothing spline ANOVA models*, Springer Science & Business Media.
- Gu, C. & Kim, Y.-J. (2002), ‘Penalized likelihood regression: general formulation and efficient approximation’, *Canadian Journal of Statistics* **30**(4), 619–628.
- 410 Halton, J. H. (1960), ‘On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals’, *Numerische Mathematik* **2**(1), 84–90.
- Hastie, T. (1996), ‘Pseudosplines’, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 379–396.
- Joseph, V. R., Gul, E. & Ba, S. (2015), ‘Maximum projection designs for computer experiments’, *Biometrika* **102**(2), 371–380.
- 415 Kim, Y.-J. & Gu, C. (2004), ‘Smoothing spline gaussian regression: more scalable computation via efficient approximation’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**(2), 337–356.
- Kolouri, S., Martin, C. E. & Rohde, G. K. (2018), ‘Sliced-wasserstein autoencoder: An embarrassingly simple generative model’, *arXiv preprint arXiv:1804.01947*.
- Kuipers, L. & Niederreiter, H. (2012), *Uniform distribution of sequences*, Courier Corporation.
- 420 Kuncheva, L. I. & Whitaker, C. J. (2003), ‘Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy’, *Machine learning* **51**(2), 181–207.
- Lemieux, C. (2009), *Monte Carlo and quasi-Monte Carlo sampling*, Springer, New York.
- Leobacher, G. & Pillichshammer, F. (2014), *Introduction to quasi-Monte Carlo integration and applications*, Springer.
- Lin, Y. & Zhang, H. H. (2006), ‘Component selection and smoothing in multivariate nonparametric regression’, *The Annals of Statistics* **34**(5), 2272–2297.
- 425 Luo, Z. & Wahba, G. (1997), ‘Hybrid adaptive splines’, *Journal of the American Statistical Association* **92**(437), 107–116.
- Ma, P., Huang, J. Z. & Zhang, N. (2015), ‘Efficient computation of smoothing splines via adaptive basis sampling’, *Biometrika* **102**(3), 631–645.
- 430 Owen, A. B. (2003), ‘Quasi-monte carlo sampling’, *Monte Carlo Ray Tracing: Siggraph* **1**, 69–88.
- Pukelsheim, F. (2006), *Optimal design of experiments*, SIAM.
- Rabin, J., Peyré, G., Delon, J. & Bernot, M. (2011), Wasserstein barycenter and its application to texture mixing, in ‘International Conference on Scale Space and Variational Methods in Computer Vision’, Springer, pp. 435–446.
- Reinsch, C. H. (1967), ‘Smoothing by spline functions’, *Numerische mathematik* **10**(3), 177–183.
- 435 Rokach, L. (2010), ‘Ensemble-based classifiers’, *Artificial Intelligence Review* **33**(1-2), 1–39.

- Ruppert, D. (2002), ‘Selecting the number of knots for penalized splines’, *Journal of computational and graphical statistics* **11**(4), 735–757.
- Ruppert, D., Wand, M. P. & Carroll, R. J. (2009), ‘Semiparametric regression during 2003–2007’, *Electronic journal of statistics* **3**, 1193.
- Serra, P. & Krivobokova, T. (2017), ‘Adaptive empirical bayesian smoothing splines’, *Bayesian Analysis* **12**(1), 219–238. 440
- Sobol, I. (1967), ‘The distribution of points in a cube and the approximate evaluation of integrals’, *USSR Computational Mathematics and Mathematical Physics* **7**(4), 86–112.
- Villani, C. (2008), *Optimal transport: old and new*, Springer Science & Business Media.
- Wahba, G. (1990), *Spline models for observational data*, SIAM. 445
- Wahba, G. & Craven, P. (1978), ‘Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation.’, *Numerische Mathematik* **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 ‘Interactive Ray Tracing 2006, IEEE Symposium on’, IEEE, pp. 61–69.
- Wang, X., Shen, J. & Ruppert, D. (2011), ‘On the asymptotics of penalized spline smoothing’, *Electronic Journal of Statistics* **5**, 1–17. 450
- Wood, S. N. (2003), ‘Thin plate regression splines’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(1), 95–114.
- Wu, C. J. & Hamada, M. S. (2011), *Experiments: planning, analysis, and optimization*, 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’, *Journal of the American Statistical Association* **99**(467), 659–672. 455