Smoothing splines approximation using Hilbert curve basis selection

Cheng Meng[1], **Jun Yu**[2], **Yongkai Chen**[3], **Wenxuan Zhong**[3], **Ping Ma**[3,*]

[1]Institute of Statistics and Big Data, Renmin University of China

[2]School of Mathematics and Statistics, Beijing Institute of Technology

[3]Department of Statistics, University of Georgia

Abstract

Smoothing splines have been used pervasively in nonparametric regressions. However, the computational burden of smoothing splines is significant when the sample size n is large. When the number of predictors $d \geq 2$, the computational cost for smoothing splines is at the order of $O(n^3)$ using the standard approach. Many methods have been developed to approximate smoothing spline estimators by using q basis functions instead of n ones, resulting in a computational cost of the order $O(nq^2)$. These methods are called the basis selection methods. Despite algorithmic benefits, most of the basis selection methods require the assumption that the sample is uniformly-distributed on a hyper-cube. These methods may have deteriorating performance when such an assumption is not met. To overcome the obstacle, we develop an efficient algorithm that is adaptive to the unknown probability density function of the predictors. Theoretically, we show the proposed estimator has the same convergence rate as the full-basis estimator when q is roughly at the order of $O[n^{2d/\{(pr+1)(d+2)\}}]$, where $p \in [1, 2]$ and $r \approx 4$ are some constants depend on the type of the spline. Numerical studies on various synthetic datasets demonstrate the superior performance of the proposed estimator in comparison with mainstream competitors.

Keywords

Space-filling curve; Nonparametric regression; Penalized least squares; Subsampling

1 Introduction

Smoothing spline estimators have been used pervasively in nonparametric regression models

$$y_i = \eta(\boldsymbol{x}_i) + \epsilon_i \quad i = 1, \dots n, \tag{1}$$

where $y_i \in \mathbb{R}$ is the response, $\boldsymbol{x}_i \in \mathbb{R}^d$ is the predictor, η is the unknown function to be estimated, and $\{\epsilon_i\}_{i=1}^{n}$ are i.i.d. normal random errors with zero mean and unknown variance σ^2 (Wahba, 1990; Wang, 2011; Gu, 2013; Zhang et al., 2018a). Despite their impressive performance, smoothing splines suffer from a huge computational burden when

the sample size n is large. Although univariate smoothing splines can be computed in $O(n)$ time (Reinsch, 1967), in general cases when the number of predictors $d \geq 2$, the classical method for calculating smoothing splines requires computing the inverse of a $n \times n$ matrix. The standard algorithm for calculating matrix inversion requires $O(n^3)$ computational time. To reduce such a huge computational cost, existing methods approximate smoothing spline estimators by using $q \ll n$ basis functions instead of n ones. These methods are called the basis selection methods, which can reduce the computational cost to $O(nq^2)$. Notice that one can further refine the order $O(n^3)$ and $O(nq^2)$ to $o(n^3)$ and $o(nq^2)$, respectively, using Strassen algorithm, Coppersmith–Winograd algorithm or Optimized CW-like algorithms (Bernstein, 2009; Golub and Van Loan, 2013). These algorithms are beyond the scope of this paper.

Various basis selection methods have been proposed. Luo and Wahba (1997) and Zhang et al. (2004) selected the basis functions through variable selection techniques. Hastie (1996) and Ruppert (2002) considered pseudosplines, also called P-splines, which utilize q fixed basis functions to approximate splines. Such fixed basis functions are also called knots and differ from the construction of the basis functions in smoothing splines. He et al. (2001); Sklar et al. (2013), and Yuan et al. (2013) considered the cases that the regression function has non-homogeneous smoothness across the design space. They developed data-driven methods to select basis functions or knots, such that the selected ones are adaptive to non-homogeneous smoothness of the regression function.

There also exist other strategies that aim to approximate splines or other nonparametric regression estimators in a computationally efficient manner through parallel computing. Zhang et al. (2013) and Zhang et al. (2015) studied the divide-and-conquer kernel ridge regression (dacKRR), and showed that it achieves minimax optimal convergence rates under relatively mild conditions. Wood et al. (2017) accelerated the fitting of penalized regression spline based generalized additive models. They showed that their method could run reliably and efficiently on a desktop workstation for d up to 10^4 and n up to 10^8. Xu and Wang (2018) and Xu et al. (2019) considered the problem of how to estimate the tuning parameter effectively for dacKRR. They proposed a variant of the generalized cross-validation for dacKRR, and showed that their proposed technique is computationally scalable for massive datasets and is asymptotically optimal under mild conditions. Shang and Cheng (2017) analyzed the theoretical properties of one-dimensional smoothing splines under the divide-and-conquer setting. Liu et al. (2018) and Liu et al. (2020) studied the theoretical properties of dacKRR respecting the number of machines. They showed that there exists a specific bound for the number of machines in order to let the dacKRR estimators to achieve statistical minimax. (Shang et al., 2019) developed scalable Bayesian inference procedures for a general class of nonparametric regression models using distributed learning. In practice, it is possible to combine the aforementioned parallel-based strategies with the proposed method for more computational savings.

One fundamental question for basis selection methods is how to determine the size of q, which balances the trade-off between the computation time and the prediction accuracy. In this paper, we focus on the widely-used asymptotic criterion, which aims to determine the smallest order of q such that the q-basis estimator converges to the true function η at

the same rate as the full-basis estimator. Zhou and Shen (2001) proposed an estimator for regression spline using the spatial adaptive basis functions. This method has been applied in univariate cases; however, it is not clear whether it can be extended to multivariate cases. Xiao et al. (2013) proposed an estimator for P-spline under the scenario that the observations are supported on a $n_1 \times n_2$ grids, and showed that the essential number of basis $q = n_1 n_2 / 4$. One limitation of their estimator is that it can only be applied in the cases when the observations are supported on a two-dimensional grid. Gu and Kim (2002) and Ma et al. (2015a) developed the uniform basis selection method and the adaptive basis selection method, respectively. Both methods require q roughly be of the order $O\{n^{2/(pr+1)}\}$, where $p \in [1, 2]$ and $r \approx 4$ are some constants depend on the type of the spline. We provide a discussion on these two constants in Section 4. Recently, Meng et al. (2020b) proposed a more efficient basis selection method that only require q roughly be of the order $O\{n^{1/(pr+1)}\}$, when $d \quad pr + 1$. Their method aims to select approximately uniformly-distributed observations by utilizing space-filling designs or low-discrepancy sequences, resulting in a faster convergence rate compared with the uniform basis selection method.

Despite algorithmic benefits, the key to the success of most of the existing basis selection methods depends on the assumption that the sample is uniformly-distributed in a hyper-cube or a hyper-rectangular. In practice, most basis selection methods may suffer from deteriorating performance when such an assumption is not met. We now demonstrate the case that the sample is not uniformly-distributed using a toy example. In this example, we generate two thousand data points from a banana-shape distribution on $[0,1]^2$, and we show the heat map of the true response surface $y = \sin\{20(x_1 + x_2)\}$ in the leftmost panel of Fig. 1. The marginal distribution of such banana-shape distribution conditional on x_1 is a Gaussian distribution; thus, more data points are located in the middle than on the boundary. We compare the proposed method, denoted by HBS, with the mainstream competitors, which includes the uniform basis selection (Gu and Kim, 2002), the adaptive basis selection (Ma et al., 2015a), and the space-filling basis selection (Meng et al., 2020b). We set $q = 5 \times (2000)^{2/9} \approx 27$ for all basis selection methods, and we mark the selected basis functions as black triangles. The right four panels of Fig. 1 show the heat maps of the spline estimates of all four basis selection methods, respectively. We observe the uniform basis selection method and the adaptive basis selection method perform similarly: both select very few basis functions on the boundary. As a result, these two methods fail to capture the periodic pattern of the response surface on the boundary. In contrast, the space-filling basis selection method selects very few basis functions in the middle, resulting in degenerated performance in such an area. These observations suggest that the performance of a basis selection method may deteriorate significantly when the sample is not uniformly generated in a hypercube. Finally, we observe the proposed method is adaptive to the arbitrary distribution of the sample, resulting in the best estimation of the true function, compared with other methods.

In practice, the distribution of the sample is almost always unknown to practitioners. The basis selection method hence is highly desirable to be robust to arbitrary distribution of the sample. To achieve the goal, it is suggested in Chapter 4 of Gu (2013) to select the basis functions corresponding to roughly equally-spaced observations, even when the sample is not uniformly-distributed. Analogously, Eilers and Marx (2010) found that equally-spaced knots, which can be regarded as the basis functions here, are always preferred in practice.

These discoveries are consistent with the common key idea in importance sampling techniques, which are widely-used for variance-reduction in numerical integration (Liu, 1996, 2008). We now briefly introduce such an idea in the following.

Let f be an integrand and g be a probability density function on $\Omega \subseteq \mathbb{R}^d$. To estimate the integration $\int_{\Omega} f(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}$, one can simply generate an i.i.d. sample $\{\boldsymbol{x}_i\}_{i=1}^{n}$ from g, then calculate the mean of $\{f(\boldsymbol{x}_i)\}_{i=1}^{n}$. Instead, one can also generate an i.i.d. sample from a probability density function h, then calculate the mean of $\{f(\boldsymbol{x}_i) g(\boldsymbol{x}_i)/h(\boldsymbol{x}_i)\}_{i=1}^{n}$. Kahn and Marshall (1953) showed when both f and g are known, the optimal $h(\boldsymbol{x})$ in terms of variance-reduction is proportional to $|f(\boldsymbol{x})| g(\boldsymbol{x})$, outside of trivial cases where $\int |f(\boldsymbol{x})| g(\boldsymbol{x}) \mathrm{d} \boldsymbol{x} = 0$. The intuition is that $h(\boldsymbol{x})$ needs to have sufficiently large value for the $\boldsymbol{x}$ such that $|f(\boldsymbol{x})| g(\boldsymbol{x})$ is close to zero. Otherwise, if $h(\boldsymbol{x})$ is extremely small for such $\boldsymbol{x}$, the variance of $\mathbb{E}_g(|f(\boldsymbol{x})|) = |f(\boldsymbol{x})| g(\boldsymbol{x})/h(\boldsymbol{x})$ can be inflated to be arbitrarily large. Consequently, in the cases when either f or g is unknown, a safe strategy is to let h be the uniform distribution on Ω; thus simply avoids the scenario that $h(\boldsymbol{x})$ is extremely small for any $\boldsymbol{x} \in \Omega$.

Inspired by such a strategy in importance sampling techniques, we propose a novel basis selection method by selecting the basis functions corresponding to roughly equally-spaced observations. To achieve the goal, we develop an efficient algorithm that utilizes the Hilbert space-filling curve. The proposed algorithm can be used to select a uniformly-distributed subsample without knowing the probability density function of the predictors. Theoretically, we show the proposed estimator has the same convergence rate as the full-basis estimator. Furthermore, we show the order of q for the proposed method is reduced from roughly $O(n^{2/(pr+1)})$ in the uniform basis selection method to roughly $O(n^{2d/\{(pr+1)(d+2)\}})$. To the best of our knowledge, in the cases when the sample follows an arbitrary distribution, the proposed estimator is the one that requires the smallest order of q. Numerical studies on various synthetic datasets demonstrate the superior performance of the proposed estimator in comparison with mainstream competitors.

Although we mainly focus on smoothing splines in this paper, it is possible that the proposed method could also accelerate the estimation of other nonparametric regression estimators, includes the thin plate regression splines, kernel ridge regression and etc (Geer and van de Geer, 2000; Wood, 2003; Györfi et al., 2006; Wasserman, 2006; Hastie et al., 2009; Yang et al., 2017). Some simulation results are provided in Supplementary Material to support this claim.

The rest of the paper is organized as follows. We review smoothing splines and basis selection methods in Section 2. In Section 3, we first introduce the Hilbert curve and some of its properties. We then introduce our basis selection method utilizing the Hilbert curve. In Section 4, we present theoretical properties of the proposed method. We evaluate the empirical performance of the proposed method via extensive simulation studies and a real-world data analysis in Section 5 and Section 6, respectively. Section 7 includes some discussion of the paper. Proofs of the theorems are collected in Supplementary Material.

2.   Preliminaries

2.1   Background of Smoothing Splines

To estimate the unknown function η in Model (1), a common strategy is to minimize the penalized least squares criterion (Wahba, 1990; Wang et al., 2011; Gu, 2013; Wang et al., 2013),

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

Where $J(\cdot)$ is a squared semi-norm, and $J(\eta)$ is called the roughness penalty. The λ here is the smoothing parameter, which balances the trade-off between the goodness-of-fit of the model and the roughness of the function η. Such λ can be selected based on the generalized cross-validation (GCV) criterion (Wahba and Craven, 1978). Xu and Wang (2018) and Xu et al. (2019) generalized the GCV criterion to the setting of distributed learning. Recently, Sun et al. (2021) proposed a more efficient approach for accelerating the calculation of λ.

In this paper, we focus on minimizing the objective function (2) in a reproducing kernel Hilbert space, resulting in a smoothing spline estimate for η. Let $\mathscr{H} = \{\eta : J(\eta) < \infty\}$ be a reproducing kernel Hilbert space and $\mathscr{N}_J = \{\eta : J(\eta) = 0\}$ be the null space of $J(\eta)$. Let $\mathscr{N}_J$ be a m-dimensional linear subspace of $\mathscr{H}$ and $\{\xi_i\}_{i=1}^{m}$ be a set of basis for $\mathscr{N}_J$. Moreover, let $\mathscr{H}_J$ to denote the orthogonal complement of $\mathscr{N}_J$ in $\mathscr{H}$ such that $\mathscr{H} = \mathscr{N}_J \oplus \mathscr{H}_J$. It can be shown that $\mathscr{H}_J$ is a still a reproducing kernel Hilbert space, and we use $R_J(\cdot,\cdot)$ to denote the reproducing kernel of $\mathscr{H}_J$. Let $\mathbf{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be the response vector, $S \in \mathbb{R}^{n \times m}$ be a matrix where the (i, j)-th element equals $\xi_j(\boldsymbol{x}_i)$, and $R \in \mathbb{R}^{n \times n}$ be a matrix where the (i, j)-th element equals $R_J(\boldsymbol{x}_i, \boldsymbol{x}_j)$. According to the representer theorem (Wahba, 1990), the minimizer of the objective function (2) in the space $\mathscr{H}$ takes the form $\eta(\boldsymbol{x}) = \sum_{k=1}^{m} \alpha_k \xi_k(\boldsymbol{x}) + \sum_{i=1}^{n} \beta_i R_J(\boldsymbol{x}_i, \boldsymbol{x})$. Let $\boldsymbol{a} = (a_1, \ldots, a_m)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^{\mathrm{T}}$ be the coefficient vectors. With trivial modification, it can be shown that finding the minimizer of the objective function (2) is equivalent to solving

$$(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \operatorname*{argmin}_{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^n} \frac{1}{n}(Y - S\boldsymbol{\alpha} - R\boldsymbol{\beta})^{\mathrm{T}}(Y - S\boldsymbol{\alpha} - R\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}R\boldsymbol{\beta}. \tag{3}$$

Although the solution of the minimization problem (3) has a closed form (Gu and Kim, 2002; Ma et al., 2015a), the computational cost for calculating the solution is of the order $O(n^3)$, in a general case where $n \gg m$ and $d \ \ 2$.

2.2   Basis Selection Methods

To alleviate the computation burden for smoothing splines, various basis selection methods have been developed. These methods are of similar nature to the subsampling methods, which are widely used in large-scale data analysis (Mahoney, 2011; Drineas et al., 2012; Ma

et al., 2015b; Ma and Sun, 2015; Meng et al., 2017; Zhang et al., 2018b; Ai et al., 2021b; Xie et al., 2019; Ma et al., 2020; Yu et al., 2020; Ai et al., 2021a; Meng et al., 2020a; Zhong et al., 2021). We refer to Li and Meng (2020) for a recent review.

The standard basis selection method works as follows. One first use subsampling techniques to select a subsample $\{x_i^*\}_{i=1}^q$ from the observed sample $\{x_i\}_{i=1}^n$. The selected subsample is then used to construct the so-called effective model space $\mathscr{H}_E = \mathscr{N}_J \oplus \mathrm{span}\{R_J(x_i^*, \cdot), i = 1, ..., q\}$. Finally, the objective function (2) is minimized in the effective model space $\mathscr{H}_E$, and the solution thus can be written as $\eta_E(x) = \sum_{k=1}^m \alpha_k \xi_k(x) + \sum_{i=1}^q \beta_i R_J(x_i^*, x)$. Analogous to Equation (3), the coefficients $\boldsymbol{a}_{E} = (a_1, ..., a_m)^\mathrm{T}$ and $\boldsymbol{\beta}_E = (\beta_1, ..., \beta_q)^\mathrm{T}$ can be obtained through solving

$$
\begin{aligned}
(\boldsymbol{\alpha}_E, \boldsymbol{\beta}_E) = \underset{\boldsymbol{\alpha}_E \in \mathbb{R}^m, \boldsymbol{\beta}_E \in \mathbb{R}^q}{\mathrm{argmin}} \frac{1}{n}(Y - S\boldsymbol{\alpha}_E - R_*\boldsymbol{\beta}_E)^\mathrm{T}(Y - S\boldsymbol{\alpha}_E - R_*\boldsymbol{\beta}_E) \\
+ \lambda \boldsymbol{\beta}_E^\mathrm{T} R_{**} \boldsymbol{\beta}_E,
\end{aligned}
\tag{4}
$$

where $R_* \in \mathbb{R}^{n \times q}$ is a matrix where the (i, j)-th element equals $R_J(x_i, x_j^*)$ and $R_{**} \in \mathbb{R}^{q \times q}$ is a matrix where the (i, j)-th element equals $R_J(x_i^*, x_j^*)$. In general cases where $m \ll q \ll n$, solving the optimization problem (4) requiring only $O(nq^2)$ computation time, which is a significant reduction compared with $O(n^3)$.

Despite algorithmic benefits, most of the existing basis selection methods heavily rely on the condition that the sample is uniformly-distributed on a hyper-cube. When such a condition is not met, they may suffer from deteriorating performance, as shown in Fig 1. Recall that a common strategy in importance sampling techniques is to select a roughly uniformly-distributed subsample, which tends to be beneficial for numerical integration. Such a strategy motivates us to select the basis functions corresponding to roughly equally-spaced observations, even when the sample is not uniformly-distributed. Intuitively, such a goal can be easily achieved when $d = 1$, in which cases one can first divide the sample space into equally-spaced bins, and then select an equal number of observations within each bin. The selected subsample is roughly uniformly-distributed when the number of bins is carefully determined. Unfortunately, such a naive strategy is not easily extendable to the cases that $d \geq 2$ due to the curse-of-dimensionality.

To overcome the barrier, a natural strategy is to find a continuous mapping $F : \Omega \to \mathbb{R}$ that preserves local structures, where $\Omega \subset \mathbb{R}^d$ is a bounded design space. In other words, we aim to find a mapping F such that, for any $x_i, x_j \in \Omega$, $i, j \in \{1, ..., n\}$, a small value of $\|x_i = x_j\|$ is associated with a small value of $\|F(x_i) = F(x_j)\|$, where $\|\cdot\|$ represents the Euclidean norm. Loosely speaking, let $\{F(x_i^*)\}_{i=1}^q$ be a roughly uniformly-distributed subset selected from $\{F(x_i)\}_{i=1}^n$, the subsample $\{x_i^*\}_{i=1}^q$ thus tends to be uniformly-distributed in Ω. One family of the mappings that approximately achieve this goal is the family of space-filling curves, which include the Hilbert curve, the Peano curve, and the Z-order curve (Sagan, 2012). We

develop a novel basis selection method utilizing space-filling curves, as detailed in the next section.

3 Basis Selection using Space-Filling Curves

3.1 Hilbert Curves

Space-filling curves have long been studied in mathematics and have become important computational tools since the 1980s (Bader, 2012). Nowadays, space-filling curves have been widely used for computer graphics, approximately nearest neighbor searching, solving partial differential equations, and so on (Zumbusch, 2012). We now briefly introduce the Hilbert curve, a representative of space-filling curves, and some of its properties that we need. The formal definition of the Hilbert curve is relegated to Supplementary Material. Other space-filling curves enjoy similar properties, and we refer to Sagan (2012); Zumbusch (2012) for more details.

We first introduce a sequence of the so-called Hilbert space-filling curves, denoted by $\{H_k\}_{k=1}^{\infty}$. Intuitively, for each k, the k-th Hilbert space-filling curve H_k is a bijection between a partition of $[0,1]$ and a partition of $[0,1]^d$. In particular, the curve H_k partitions both $[0,1]$ and $[0,1]^d$ into $(2^k)^d$ blocks, respectively, and construct a bijection between these blocks. Figure 2 illustrates how a partition of $[0,1]$ is mapped to a partition of $[0,1]^2$ using H_1, H_2, and H_3, respectively. The Hilbert curve is defined as $H(x) = \lim_{k \to \infty} H_k(x)$, which becomes a mapping from $[0,1]$ to $[0,1]^d$. It is well-known that the Hilbert curve H enjoys the locality-preserving property (Zumbusch, 2012). In particular, for any $x, y \in [0,1]$, one has

$$\|H(x) - H(y)\| \le 2\sqrt{d+3}|x - y|^{1/d}. \tag{5}$$

Inequality (5) indicates a small value of $|x - y|$ is associated with a small value of $\|H(x) - H(y)\|$, despite the fact that the converse can't always be true. Inequality (5) inspired us to select approximately equally-spaced observations using the Hilbert curve H. In practice, H_k is used as a surrogate of H due to the computational concern.

3.2 Hilbert Basis Selection Method

We develop a novel basis selection method utilizing Hilbert space-filling curves, called the Hilbert basis selection method. The proposed method works as follows. We first scale the sample $\{x_i\}_{i=1}^{n} \in \mathbb{R}^d$ to $[0,1]^d$ as a pre-processing step. Recall that the Hilbert space-filling curve H_k partitions both $[0,1]$ and $[0,1]^d$ into 2^{kd} blocks, denoted by $\{c'_j\}_{j=1}^{2^{dk}}$, and $\{c_j\}_{j=1}^{2^{dk}}$, respectively, and construct a bijection between these blocks. For any data point $x \in [0,1]^d$, we assign x to its corresponding block c_j in $[0,1]^d$, $j \in \{1,\ldots, 2^{kd}\}$, then map x to the center of the block $c_{j'} = H_k^{-1}(c_j)$. This is to say, all the data points that belong to the same block are mapped to the same point in $[0,1]$. Next, given a positive integer C, we draw the histogram for the mapped data points with C bins. Let $\tilde{C}$ be the number of non-empty bins and q be the subsample size. We then randomly select roughly $\lceil q/\tilde{C}\rceil$ data points from each

non-empty bin. The subsample corresponding to the selected data points is used to construct the effective subspace $\mathcal{H}_E$. Finally, we calculate the smoothing spline estimator η_E in such a subspace. The algorithm is summarized below.

Algorithm *1* Hilbert basis selection method

Step 1. The sample $\{\boldsymbol{x}_i\}_{i=1}^n$ is first scaled to $[0,1]^d$.

Step 2. Calculate the bijection between $\{c'_j\}_{j=1}^{2^{dk}}$ and $\{c_j\}_{j=1}^{2^{dk}}$ using the Hilbert space-filling
curve H_k.

Step 3. For each data point $\boldsymbol{x}_i, i = 1, \ldots, n$, suppose $\boldsymbol{x}_i$ belongs to the block c_j,
map $\boldsymbol{x}_i$ to the center of the block $c_{j'} = H_k^{-1}(c_j)$.

Step 4. Draw a histogram for the mapped points with C bins;
let $\widetilde{C}$ be the number of non-empty bins.

Step 5. Randomly select roughly $\lceil q/\widetilde{C} \rceil$ number of data points from each non-empty bin;
let $\{\boldsymbol{x}_i^*\}_{i=1}^q$ to denote the selected ones.

Step 6. Minimize the objective function (2) over the effective subspace
$\mathcal{H}_E = \mathcal{N}_J \oplus \text{span}\{R_J(\boldsymbol{x}_i^*, \cdot), i = 1, \ldots, q\}$.

Figure 3 gives an illustration of Algorithm 1, in which all data points are shown in Fig. 3(a). We set $k = d = 2$ in Algorithm 1, resulting in $2^{2 \times 2} = 16$ blocks in Fig. 3(b), denoted by $c_1, \ldots, c_{16}$, respectively. All the data points are then mapped to the center of $c_{j'}$s, $j = 1, \ldots, 16$, as shown in Fig. 3(c). Let the subsample size $q = 8$. We then draw $C = 8$ bins for the histogram in Fig. 3(c), resulting in $\widetilde{C} = 8$ non-empty bins. We then randomly select $\widetilde{C}/q = 1$ data point from each bin. The selected data points are labeled as black triangles in both Fig. 3(c) and Fig. 3(d). Note that each bin in Fig. 3(c) is associated with a "meta-block", as illustrated in Fig. 3(d). As a result, when $\widetilde{C} = q$, Algorithm 1 ensures none of the two selected data points lie in the same meta-block, and thus the selected subsample tend to be equally-spaced.

The choice of k is a key to Algorithm 1, while the performance of Algorithm 1 is not sensitive to the choice of k, as long as k is not too small. The reasons are as follows. Recall that for dimension d, the curve H_{k_0} with respect to (w.r.t.) a positive integer k_0 partitions

the interval $[0,1]$ into 2^{dk_0} blocks, denoted by $\{c'_j\}_{j=1}^{2^{dk_0}}$. One nice property of the Hilbert space-filling curve is that, suppose one data point $\boldsymbol{x}$ is mapped into a block using the curve H_{k_0}, then for any $k \geq k_0$, the curve H_k always map $\boldsymbol{x}$ into the same block, despite the fact that its position within the block may vary. Note that Algorithm 1 draws a histogram for the mapped data points with C bins and randomly selects several data points from each of the non-empty bins. As a result, when C is properly chosen and is fixed, the value of $k \geq k_0$ does not affect such a histogram, and thus does not affect the result of Algorithm 1.

The computational cost for Algorithm 1 mainly resides in Step 2 and Step 6. It can be shown that the computational cost for Step 2 is of the order $O(n)$, which is negligible compared to the computational cost for Step 5, which is of the order $O(nq^2)$. In sum, analogous to other basis selection methods, the overall computational cost for Algorithm 1 is at the order of $O(nq^2)$.

4 Convergence Rates for Function Estimation

Let f_x be the probability density function of the predictors defined on Ω. We require Ω to be bounded, and without loss of generality, we assume $\Omega \subseteq [0,1]^d$. Let $V(g) = \int_{\Omega} g^2 f_X \mathrm{d}\boldsymbol{x}$. A function $f(x)$ defined on Ω is said to be Lipschitz continuous, if for any $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, there exist a constant M such that $|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq M\|\boldsymbol{x} - \boldsymbol{y}\|$, where $\|\cdot\|$ is the Euclidean norm. We introduce some essential regularity conditions in the following.

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

- Condition 2. For some $\beta > 0$ and $r > 1$, $\rho_v > \beta v^r$ for sufficiently large v;

- Condition 3. For all μ and v, $\mathrm{var}\{\phi_v(\boldsymbol{x})\phi_\mu(\boldsymbol{x})\} \leq B$, where ϕ_v, ϕ_μ are the eigenfunctions associated with V and J in $\mathcal{H}$, B denotes a positive constant.

- Condition 4. For all μ and v, $\phi_\mu(\boldsymbol{x})\phi_v(\boldsymbol{x}) \in L^2(\Omega)$, and is Lipschitz continuous.

- Condition 5. Assume that $\max\{(qn_i)/n\}_{i=1}^C = O_p(1)$, where n_i is the number of observations in the ith bin.

- Condition 6. As $n \to \infty, q^{1+2/d} = O(n)$.

Condition 1 implies that there exist a sequence of eigenfunctions $\phi_v \in \mathcal{H}$ and the associated sequence of eigenvalues $\rho_v \uparrow \infty$ satisfying $V(\phi_v, \phi_\mu) = \delta_{v\mu}$ and $J(\phi_v, \phi_\mu)) = \rho_v \delta_{v\mu}$, where $\delta_{v\mu}$ is the Kronecker delta. The growth rate of ρ_v is closely related to the convergence rate of smoothing spline estimates (Gu, 2013). Condition 1 can be verified under some special cases when the eigenfunctions are available in explicit forms. Consider $J(\eta) = \int_0^1 \eta'' \mathrm{d}x$, where η is a periodic function on $[0,1]$. The eigenfunctions ϕ_vs are the sine and cosine functions in such a case, and thus Condition 1 holds naturally. We refer to Section 9.1 of Gu (2013) for more details on the construction of the eigenfunctions. In general, Conditions 1–3 are widely-used in the asymptotic analysis for smoothing spline estimates, and we refer to Gu (2013); Ma et al. (2015a) for more technical discussion of these conditions. Condition 4 is satisfied naturally for various choices of eigenfunctions. Condition 5 naturally holds for the regular sampling in $[0,1]^d$. Moreover, Condition 5 prevents some extreme cases of the probability density function f_x. For example, when the one-dimensional data points $\left\{H_k^{-1}(x_i)\right\}_{i=1}^n$ follow the Dirac delta function, one has $\max\{(qn_i)/n\}_{i=1}^C = qn/n = q$, which is in conflict with Condition 5. Finally, Condition 6 naturally holds when the number of basis q is not too large. For example, when $d \leq 2$, Condition 6 holds when $q = O(n^{1/2})$.

Recall that $\{\boldsymbol{x}_j^*\}_{j=1}^q$ is a subsample selected by the proposed algorithm. Moreover, C and $\tilde{C}$ are the number of bins and the number of non-empty bins in Step 4 of Algorithm

1, respectively. For brevity, throughout this section, we assume only one data point is selected from each non-empty bin; that is, we assume $q = \tilde{C}$. The extensions to more general cases where $q/\tilde{C} = O(1)$ are straightforward. We let $\tilde{n}_j$ to denote the number of data points within the bin that x_j^* lies in. Consider the estimator $\sum_{j=1}^q \{\tilde{n}_j/n\}\phi_\nu(x_j^*)\phi_\mu(x_j^*)$. Intuitively, such an estimator can be regarded as the mean estimator of the stratified sampling. This is because, for $j = 1, ..., q\{\tilde{n}_j/n\}\phi_\nu(x_j^*)\phi_\mu(x_j^*)$ can be regarded as the sample mean of the jth strata. The following lemma gives the convergence rate of the selected subsample in terms of numerical integration. All the proofs throughout this section are relegated to the Supplementary Material.

Lemma 4.1. *Under Conditions 4–6, for all μ and ν, $\sum_{j=1}^q \{\tilde{n}_j/n\}\phi_\nu(x_j^*)\phi_\mu(x_j^*)$ is an asymptotically unbiased estimate for $\int_\Omega \phi_\nu(x)\phi_\mu(x)f_X(x)\mathrm{d}x$. Furthermore, when $\Omega \subseteq [0,1]^d$ we have*

$$\left\{\int_\Omega \phi_\nu(x)\phi_\mu(x)f_X(x)\mathrm{d}x - \sum_{j=1}^q \{\tilde{n}_j/n\}\phi_\nu(x_j^*)\phi_\mu(x_j^*)\right\}^2 = O_p(q^{-1-2/d}).$$

Lemma 4.1 shows the advantage of $\{x_i^*\}_{i=1}^q$ over a randomly selected subsample $\{x_i^+\}_{i=1}^q$. To be specific, for all μ and ν, as a direct consequence of Condition 3, which assumes that the variance of $\phi_\nu(x)\phi_\mu(x)$ is finite, we have

$$\mathbb{E}\left[\int_\Omega \phi_\nu(x)\phi_\mu(x)f_X(x)\mathrm{d}x - \frac{1}{q}\sum_{j=1}^q \phi_\nu(x_j^+)\phi_\mu(x_j^+)\right]^2 = O(q^{-1}).$$

Consequently, Lemma 4.1 suggests that one can approximate the integration $\int_\Omega \phi_\nu(x)\phi_\mu(x)f_X(x)\mathrm{d}x$ more effectively, by calculating $\sum_{j=1}^q (n_j/n)\phi_\nu(x_j^*)\phi_\mu(x_j^*)$ instead of $\sum_{j=1}^q \phi_\nu(x_j^+)\phi_\mu(x_j^+)/q$. Lemma 1 paves the way for our main theorem below.

Theorem 4.1. *Suppose $\sum_i \rho_i^p V(\eta_0, \phi_i)^2 < \infty$ for some $p \in [1, 2]$. Under Conditions 1–6, as $\lambda \to 0$ and $q^{1+2/d}\lambda^{2/r} \to \infty$, 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-basis smoothing spline estimator $\hat{\eta}$ has the convergence rate $(V + \lambda J)(\hat{\eta} - \eta_0) = O_p\{n^{-pr/(pr+1)}\}$. Theorem 1 thus states that the proposed estimator $\hat{\eta}_E$ achieves the identical convergence rate as the full-basis estimator. In particular, the convergence rate of the full-basis estimator gives a lower bound for all the estimators that are based on a subset of the basis functions. According to Gu and Kim

(2002); Ma et al. (2015a), and Meng et al. (2020b), all these proposed estimators have the sample convergence rate as the full-basis estimator $\hat{\eta}$, under different conditions.

We emphasize that the goal of Theorem 1 is not to demonstrate that the proposed estimator enjoys a more superior convergence rate. Instead, Theorem 1 indicates that to achieve such a convergence rate, the proposed estimator requires a relatively smaller q, compared with other estimators. In particular, both the uniform basis selection method (Gu and Kim, 2002) and the adaptive basis selection method (Ma et al., 2015a) require $q = O\{n^{2/(pr+1)+\delta}\}$ for an arbitrary small positive number δ. While for the proposed method, combining the condition $q^{1+2/d}\lambda^{2/r} \to \infty$ and $\lambda \asymp n^{-r/(pr+1)}$ in Theorem 1 yields, an essential choice of q should satisfy $q = O\left[n^{2d/\{(pr+1)(d+2)\}+\delta}\right]$, which is a smaller order of $O\{n^{2/(pr+1)+\delta}\}$. Although the estimator proposed in Meng et al. (2020b) only require $q = O\{n^{(1+\delta)/(pr+1)}\}$, their work assume the sample is uniformly generated from a hypercube, and such an assumption is not always achievable in practice. In the cases when the sample follows an arbitrary distribution, to the best of our knowledge, the proposed estimator is the one that requires the smallest order of q.

Consider the parameter p and q in Theorem 1. It is known that q is associated with the type of the spline, and p is closely associated with η_0. Both parameters have an impact on the convergence rate of the proposed estimator. According to Gu (2013), a common strategy is to set $p \in [1, 2]$ and $r \in [4 - \delta, 4]$ for cubic smoothing splines and tensor-product splines, in which case the size of q roughly lies in the interval $(O(n^{2d/\{9(d+2)\}}), O(n^{2d/\{5(d+2)\}}))$. We refer to Gu (2013) for more technical discussion on how to select p and r in practice.

5   Simulation Results

To show the effectiveness of the proposed smoothing spline estimator, we compare it with three mainstream competitors in terms of prediction accuracy. The competitors include the uniform basis selection method (Gu and Kim, 2002), the adaptive basis selection method (Ma et al., 2015a, 2017), and the space-filling basis selection method (Meng et al., 2020b). All the methods are implemented in R, and all the parameters are set as default.

We measure the performance for each method using the prediction mean squared error (MSE), defined as $\left[\sum_{i=1}^{n}\{\hat{\eta}_E(t_i) - \eta_0(t_i)\}^2\right]/n$, where $\{t_i\}_{i=1}^{n}$ is an independent testing dataset generate from the same probability density function as the training sample. Standard errors are calculated through a hundred replicates. In each replicate, we generate a synthetic training sample with $n = 2000$ from each of the following four probability density functions, and the sample is then scaled to $[0,1]^d$,

- D1: Uniform distribution on $[0,1]^d$;

- D2: A mixture t-distribution $(T_1, ..., T_d)$, where $\{T_i\}_{i=1}^{d}$ are independently generated from $t(10, -5)/2 + t(10,5)/2$;

- D3: A multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma_{ij} = 0.9^{|i-j|}$, $i, j = \{1, ..., d\}$.

- D4: A banana-shape distribution, which is generated by
  $\left(Z_1, Z_2 + \dfrac{z_1^2}{1.2}, \ldots, Z_d + \dfrac{z_1^2}{1.2}\right)$, where $(Z_1, Z_2, \ldots Z_d)$ is generated from the standard multivariate Gaussian distribution.

We consider four different regression functions, which are analogous to the functions considered in (Wood, 2003; Lin and Zhang, 2006):

- F1: A 2-d function $\sin(10 / (x_1 + x_2 + 0.15))$;

- F2: A 2-d function $h_1(x_1, x_2) + h_2(x_1, x_2)$, where $\sigma_1 = 0.1$, $\sigma_2 = 0.2$, and

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

$$h_2(t_1, t_2) = \{0.75/(\pi\sigma_1\sigma_2)\} \times \exp\left\{-(t_1 - 0.7)^2/\sigma_1^2 - (t_2 - 0.5)^2/\sigma_2^2\right\};$$

- F3: A 3-d function $\sin(\pi(x_1 + x_2 + x_3)/3) - x_1 - x_2^2$;

- F4: A 4-d function

$$x_1 + (2x_2 - 1)^2/2 + [\sin(10\pi x_3)/\{2 - \sin(10\pi x_3)\}]/3 +$$
$$\left\{0.1\sin(2\pi x_4) + 0.2\cos(4\pi x_4) + 0.3\sin(6\pi x_4)^2 + 0.4\cos(8\pi x_4)^3 + 0.5\sin(10\pi x_4)^2\right\}/4.$$

The signal-to-noise ratio, defined as $\mathrm{var}\{\eta(X)\}/\sigma^2$, is set to be two. We find the results show similar patterns with a large range of signal-noise-ratios. We set the number of basis q to be $\{20, 40, 60, 80, 100\}$. To combat the curse-of-dimensionality, we fit smoothing spline analysis of variance models with all main effects and two-way interactions.

Figure 4 shows the log prediction MSE versus different q under various settings. Each row represents a particular data distribution D1–D4, and each column represents a particular regression function F1–F4. We use solid lines to denote the proposed Hilbert basis selection method (HBS), dash-dotted lines to denote the adaptive basis selection method (ABS), dashed lines to denote the space-filling basis selection method (SBS), and dotted lines to denote the uniform basis selection method (UBS). The standard error bars are obtained from one hundred replicates. The results for the full-basis estimator is omitted here due to its high computation cost.

Three significant observations can be made from Fig. 4. We first observe that all the methods perform similarly, while the uniform basis selection method performs slightly worse, in the first row of Fig. 4, in which cases the observations are uniformly-distributed in a hypercube. Such an observation is consistent with the simulation results in Meng et al. (2020b), which suggests both the space-filling basis selection method and the adaptive basis selection method consistently outperform the uniform basis selection method. Nevertheless, in the lower three rows of Fig. 4, we observe the uniform basis selection method yields

decent performance occasionally. Such an observation suggests when the predictors do not follow the uniform distribution on a hypercube, none of the four basis selection methods consistently dominates the others.

Second, the MSE for the proposed estimator decreases faster than the other estimators as q increase. This observation is consistent with Theorem 1, which suggests the proposed estimator requires smaller q to achieve the identical convergence rate as the full-basis estimator.

Third, the proposed estimator may suffer from deteriorating performance when q is too small. We attributed such an observation to the fact that, when q is small, the proposed method tends to select a large proportion of basis functions corresponding to the data points that are close to the boundary. These basis functions may not have adequate benefits in terms of prediction. As q increases, the proportion of such basis functions decreases, and thus the proposed estimator achieves better performance. It is suggested in Gu (2014) to let q 30 for robust prediction in practice. According to such a suggestion and consider the cases when q 30 in Fig 4, we observe the proposed estimator outperforms the competitors in most of the settings.

6 Real data example

In petroleum refinery, a debutanizer column is used to separate butane from gasoline. Estimating the butane concentration in the bottom product of the debutanizer column is essential for improving the performance of the refining process. Of interest is to predict the butane concentration using the conditions of debutanizer columns and other related information, which are measured by the soft sensors in the petroleum refinery process. We consider a dataset with $n = 2,395$ and seven predictor variables, includes temperature, pressure, flow, and so on [1]. More details of this dataset can be found in Fortuna et al. (2007). The sample is first scaled to $[0,1]^7$ as a preprocessing step. Figure 5 shows histograms for each of the predictors in diagonal panels and scatter plots for each pair of the predictors in off-diagonal panels. We observe the sample in this dataset is extremely non-uniformly-distributed. The basis functions selected by the proposed Hilbert basis selection method and the uniform basis selection method are marked as black dots in the lower diagonal panels and the upper panels, respectively. Compare with the uniform basis selection method, we observe the proposed method selects the basis functions corresponding to the observations that are more equally-spaced.

We fitted the cubic tensor product smoothing spline analysis of variance model to the dataset, and we considered two different model settings,

- M1: additive model,

$$y_i = \eta_{\varnothing} + \sum_{j=1}^{7} \eta_j(x_{ij}) + \epsilon_i, \quad i = 1, \ldots, n;$$

---

[1]The dataset can be downloaded from https://home.isr.uc.pt/fasouza/datasets.html

- M2: by the preliminary model diagnostics (Gu, 2004), we considered the following functional ANOVA decomposition,

$$y_i = \eta_{\varnothing} + \sum_{j=1}^{7} \eta_j(x_{ij}) + \eta_{1,3}(x_{i1}, x_{i3}) + \eta_{1,5}(x_{i1}, x_{i5}) + \eta_{1,6}(x_{i1}, x_{i6})$$
$$+ \eta_{3,5}(x_{i3}, x_{i5}) + \epsilon_i, \quad i = 1, \ldots, n.$$

Here, the response y_i is the butane concentration of i-th observation, x_{ij} is the value of the j-th predictor of the i-th observation, $\eta_{\varnothing}$ is a constant function, $\{\eta_j\}_{j=1}^{7}$ are main effect functions, $\eta_{1,3}, \eta_{1,5}, \eta_{1,6}, \eta_{3,5}$ are two-way interaction functions of corresponding predictors, and ϵ_i's are i.i.d. normal errors with zero mean and unknown variance. We replicated the experiment one hundred times. To show the effectiveness of the proposed estimator, we compared it with the other three mainstream competitors, as mentioned in the previous section, in terms of the prediction MSE calculated on a holdout testing set. Figures 6 shows the log prediction MSE versus different q under two different model settings. Vertical bars represent the standard errors obtained from a hundred replicates. The horizontal lines represent the performance for the full sample estimator. We observe that the proposed estimator, labeled as solid lines, yields the second-best result for the smallest q considered here and the best result for other cases. We attribute such an observation to the fact that the proposed method selects the basis functions corresponding to roughly equally-spaced observations, resulting in a more effective estimation of the underlying regression function.

7. Discussion

In this paper, we proposed a novel basis selection method for smoothing splines approximation. Unlike the existing basis selection approaches, which mainly focus on the setting that the sample is uniformly distributed on a unit hypercube, the proposed method aims to provide an effective estimation that is adaptive to an arbitrary probability distribution of the sample. Motivated by importance sampling, we achieved the goal by carefully selecting a set of approximately equally-spaced observations, even when the sample is not uniformly-distributed. We proposed an efficient algorithm for identifying such observations by utilizing the Hilbert space-filling curve. The proposed estimator has the same convergence rate as the full-basis estimator when the number of basis q is roughly at the order of $O[n^{2d/\{(pr+1)(d+2)\}}]$. The superior performance of HBS over mainstream competitors was justified by various numerical experiments.

Our work is related to Meng et al. (2020b). In particular, Meng et al. (2020b) utilized space-filling design techniques, or low-discrepancy sequences, to identify an approximately equally-spaced subsample from "uniformly-distributed" sample, resulting in an efficient approximation. Our work extended their work to the non-uniform distribution setting and theoretically showed that the basis functions corresponding to an equally-spaced subsample still benefits the smoothing splines approximation.

The proposed method has the penitential to be applied to many large-sample applications, including but not limited to Gaussian process regression, kernel ridge regression, and low-

rank approximation of matrices. This work may speed up these techniques with theoretical guarantees. Some additional simulation results are provided in Supplementary Material.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

References

Ai Mingyao, Wang Fei, Yu Jun, and Zhang Huiming. Optimal subsampling for large-scale quantile regression. Journal of Complexity, 62:101512, 2021a.

Ai Mingyao, Yu Jun, Zhang Huiming, and Wang HaiYing. Optimal subsampling algorithms for big data regressions. Statistica Sinica, 31:749–772, 2021b.

Bader Michael. Space-filling curves: an introduction with applications in scientific computing Springer, 2012.

Bernstein Dennis S. Matrix mathematics: theory, facts, and formulas Princeton university press, 2009.

Drineas Petros, Magdon-Ismail Malik, Mahoney Michael W, and Woodruff David P. Fast approximation of matrix coherence and statistical leverage. Journal of Machine Learning Research, 13:3475–3506, 2012.

Eilers Paul HC and Marx Brian D. Splines, knots, and penalties. Wiley Interdisciplinary Reviews: Computational Statistics, 2(6):637–653, 2010.

Fortuna Luigi, Graziani Salvatore, Rizzo Alessandro, and Gabriella Xibilia Maria. Soft sensors for monitoring and control of industrial processes Springer Science & Business Media, 2007.

Geer Sara A and van de Geer Sara. Empirical Processes in M-estimation Cambridge university press, 2000.

Golub Gene H and Van Loan Charles F. Matrix computations JHU press, 2013.

Gu Chong. Model diagnostics for smoothing spline ANOVA models. Canadian Journal of Statistics, 32(4):347–358, 2004.

Gu Chong. Smoothing Spline ANOVA Models Springer Science & Business Media, 2013.

Gu Chong. Smoothing spline ANOVA models: R package gss. Journal of Statistical Software, 58 (5):1–25, 2014.

Gu Chong and Kim Young-Ju. Penalized likelihood regression: general formulation and efficient approximation. Canadian Journal of Statistics, 30(4):619–628, 2002.

Györfi László, Kohler Michael, Krzyzak Adam, and Walk Harro. A distribution-free theory of nonparametric regression Springer Science & Business Media, 2006.

Hastie Trevor. Pseudosplines. Journal of the Royal Statistical Society. Series B (Methodological), pages 379–396, 1996.

Hastie Trevor, Tibshirani Robert, and Friedman Jerome. The elements of statistical learning: data mining, inference, and prediction Springer Science & Business Media, 2009.

He Xuming, Shen Lixin, and Shen Zuowei. A data-adaptive knot selection scheme for fitting splines. IEEE Signal Processing Letters, 8(5):137–139, 2001.

Kahn Herman and Marshall Andy W. Methods of reducing sample size in Monte Carlo computations. Journal of the Operations Research Society of America, 1(5):263–278, 1953.

Li Tao and Meng Cheng. Modern subsampling methods for large-scale least squares regression. International Journal of Cyber-Physical Systems (IJCPS), 2(2):1–28, 2020.

Lin Yi and Helen Zhang Hao. Component selection and smoothing in multivariate nonparametric regression. The Annals of Statistics, 34(5):2272–2297, 2006.

Liu Jun S. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. Statistics and Computing, 6(2):113–119, 1996.

Liu Jun S. Monte Carlo Strategies in Scientific Computing Springer, 2008.

Liu Meimei, Shang Zuofeng, and Cheng Guang. How many machines can we use in parallel computing for kernel ridge regression? arXiv preprint arXiv:1805.09948, 2018.

Liu Meimei, Shang Zuofeng, and Cheng Guang. Nonparametric distributed learning under general designs. Electronic Journal of Statistics, 14(2):3070–3102, 2020.

Luo Zhen and Wahba Grace. Hybrid adaptive splines. Journal of the American Statistical Association, 92(437):107–116, 1997.

Ma Ping and Sun Xiaoxiao. Leveraging for big data regression. Wiley Interdisciplinary Reviews: Computational Statistics, 7(1):70–76, 2015.

Ma Ping, Huang Jianhua Z, and Zhang Nan. Efficient computation of smoothing splines via adaptive basis sampling. Biometrika, 102(3):631–645, 2015a.

Ma Ping, Mahoney Michael W, and Yu Bin. A statistical perspective on algorithmic leveraging. Journal of Machine Learning Research, 16:861–911, 2015b.

Ma Ping, Zhang Nan, Huang Jianhua Z, and Zhong Wenxuan. Adaptive basis selection for exponential family smoothing splines with application in joint modeling of multiple sequencing samples. Statistica Sinica, pages 1757–1777, 2017.

Ma Ping, Zhang Xinlian, Xing Xin, Ma Jingyi, and Mahoney Michael W. Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms The 23nd International Conference on Artificial Intelligence and Statistics, 2020.

Mahoney Michael W. Randomized algorithms for matrices and data. Foundations and Trends[textregistered] in Machine Learning, 3(2):123–224, 2011.

Meng Cheng, Wang Ye, Zhang Xinlian, Mandal Abhyuday, Ma Ping, and Zhong Wenxuan. Effective statistical methods for big data analytics. Handbook of Research on Applied Cybernetics and Systems Science, page 280, 2017.

Meng Cheng, Xie Rui, Mandal Abhyuday, Zhang Xinlian, Zhong Wenxuan, and Ma Ping. Lowcon: A design-based subsampling approach in a misspecified linear model. Journal of Computational and Graphical Statistics, pages 1–15, 2020a. [PubMed: 33013150]

Meng Cheng, Zhang Xinlian, Zhang Jingyi, Zhong Wenxuan, and Ma Ping. More efficient approximation of smoothing splines via space-filling basis selection. Biometrika, 107(3):723–735, 2020b. [PubMed: 32831354]

Reinsch Christian H. Smoothing by spline functions. Numerische mathematik, 10(3):177–183, 1967.

Ruppert David. Selecting the number of knots for penalized splines. Journal of computational and graphical statistics, 11(4):735–757, 2002.

Sagan Hans. Space-filling Curves Springer Science & Business Media, 2012.

Shang Zuofeng and Cheng Guang. Computational limits of a distributed algorithm for smoothing spline. The Journal of Machine Learning Research, 18(1):3809–3845, 2017.

Shang Zuofeng, Hao Botao, and Cheng Guang. Nonparametric Bayesian aggregation for massive data. Journal of machine learning research, 20(140):1–81, 2019.

Sklar Jeffrey C, Wu Junqing, Meiring Wendy, and Wang Yuedong. Nonparametric regression with basis selection from multiple libraries. Technometrics, 55(2):189–201, 2013.

Sun Xiaoxiao, Zhong Wenxuan, and Ma Ping. An asymptotic and empirical smoothing parameters selection method for smoothing spline anova models in large samples. Biometrika, 108(1): 149–166, 2021. [PubMed: 34294943]

Wahba Grace. Spline Models for Observational Data SIAM, 1990.

Wahba Grace and Craven Peter. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik, 31:377–404, 1978.

Wang Xiao, Shen Jinglai, and Ruppert David. On the asymptotics of penalized spline smoothing. Electronic Journal of Statistics, 5:1–17, 2011.

Wang Xiao, Du Pang, and Shen Jinglai. Smoothing splines with varying smoothing parameter. Biometrika, 100(4):955–970, 2013.

Wang Yuedong. Smoothing Splines: Methods and Applications CRC Press, 2011.

Wasserman Larry. All of nonparametric statistics Springer Science & Business Media, 2006.

Wood Simon N. Thin plate regression splines. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1):95–114, 2003.

Wood Simon N, Li Zheyuan, Shaddick Gavin, and Augustin Nicole H. Generalized additive models for gigadata: modeling the UK black smoke network daily data. Journal of the American Statistical Association, 112(519):1199–1210, 2017.

Xiao Luo, Li Yingxing, and Ruppert David. Fast bivariate P-splines: the sandwich smoother. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(3):577–599, 2013.

Xie Rui, Wang Zengyan, Bai Shuyang, Ma Ping, and Zhong Wenxuan. Online decentralized leverage score sampling for streaming multidimensional time series In The 22nd International Conference on Artificial Intelligence and Statistics, pages 2301–2311, 2019.

Xu Danqing and Wang Yuedong. Divide and recombine approaches for fitting smoothing spline models with large datasets. Journal of Computational and Graphical Statistics, 27(3):677–683, 2018.

Xu Ganggang, Shang Zuofeng, and Cheng Guang. Distributed generalized cross-validation for divide-and-conquer kernel ridge regression and its asymptotic optimality. Journal of computational and graphical statistics, 28(4):891–908, 2019.

Yang Yun, Pilanci Mert, and Wainwright Martin. Randomized sketches for kernels: fast and optimal nonparametric regression. The Annals of Statistics, 45(3):991–1023, 2017.

Yu Jun, Wang HaiYing, Ai Mingyao, and Zhang Huiming. Optimal distributed subsampling for maximum quasi-likelihood estimators with massive data. Journal of the American Statistical Association, page in press, 2020. doi:10.1080/01621459.2020.1773832.

Yuan Yuan, Chen Nan, and Zhou Shiyu. Adaptive B-spline knot selection using multi-resolution basis set. Iie Transactions, 45(12):1263–1277, 2013.

Zhang Hao Helen, Wahba Grace, Lin Yi, Voelker Meta, Ferris Michael, Klein Ronald, and Klein Barbara. Variable selection and model building via likelihood basis pursuit. Journal of the American Statistical Association, 99(467):659–672, 2004.

Zhang Jingyi, Jin Honghe, Wang Ye, Sun Xiaoxiao, Ma Ping, and Zhong Wenxuan. Smoothing spline ANOVA models and their applications in complex and massive datasets. Topics in Splines and Applications, 63, 2018a.

Zhang Xinlian, Xie Rui, and Ma Ping. Statistical leveraging methods in big data. In Handbook of Big Data Analytics, pages 51–74. Springer, 2018b.

Zhang Yuchen, Duchi John, and Wainwright Martin. Divide and conquer kernel ridge regression In Conference on learning theory, pages 592–617, 2013.

Zhang Yuchen, Duchi John, and Wainwright Martin. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. The Journal of Machine Learning Research, 16(1):3299–3340, 2015.

Zhong Wenxuan, Liu Yiwen, and Zeng Peng. A model-free variable screening method based on leverage score. Journal of the American Statistical Association, (just-accepted):1–36, 2021. [PubMed: 35757777]

Zhou Shanggang and Shen Xiaotong. Spatially adaptive regression splines and accurate knot selection schemes. Journal of the American Statistical Association, 96(453):247–259, 2001.

Zumbusch Gerhard. Parallel multilevel methods: adaptive mesh refinement and loadbalancing Springer Science & Business Media, 2012.
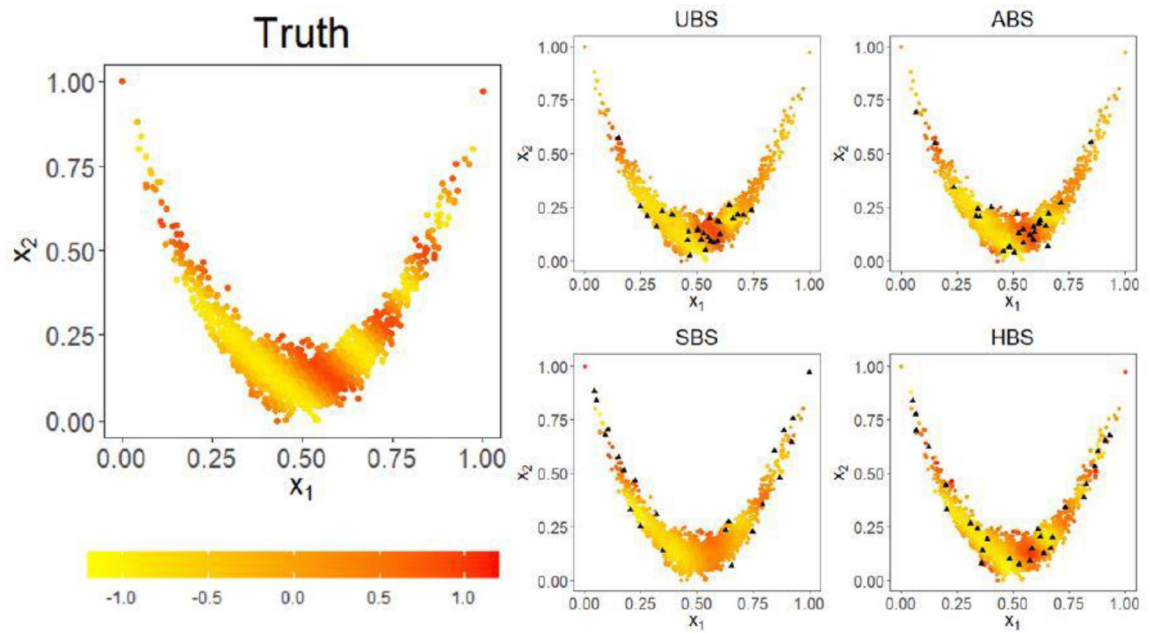
Fig. 1.

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 (UBS), the adaptive basis selection method (ABS), the space-filling basis selection method (SBS), and the proposed Hilbert basis selection method (HBS) are presented in the other panels, respectively. Black triangles are the selected basis functions. We observe that the proposed method outperforms the other methods in approximating the true function.
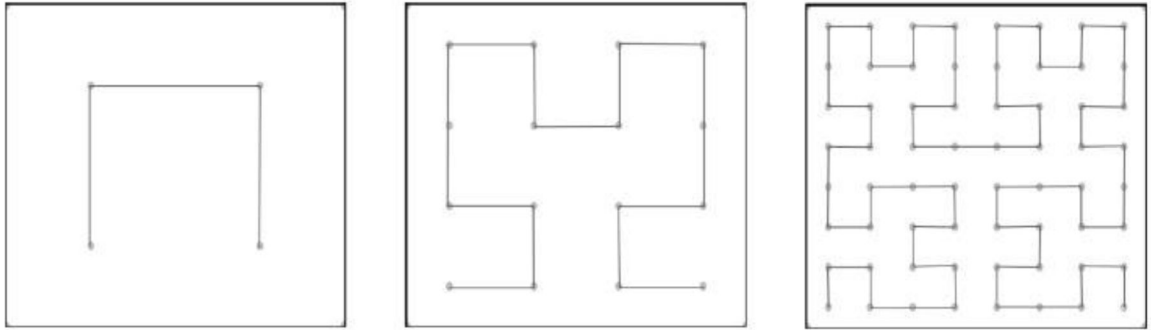
Fig. 2.
Illustration for the first three stages of the Hilbert's space filling curves when $d = 2$. From left to right, each panel shows H_1, H_2, and H_3, respectively.
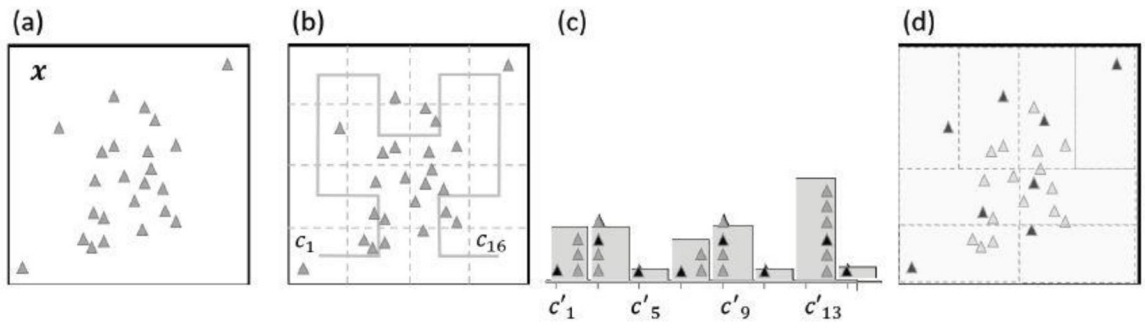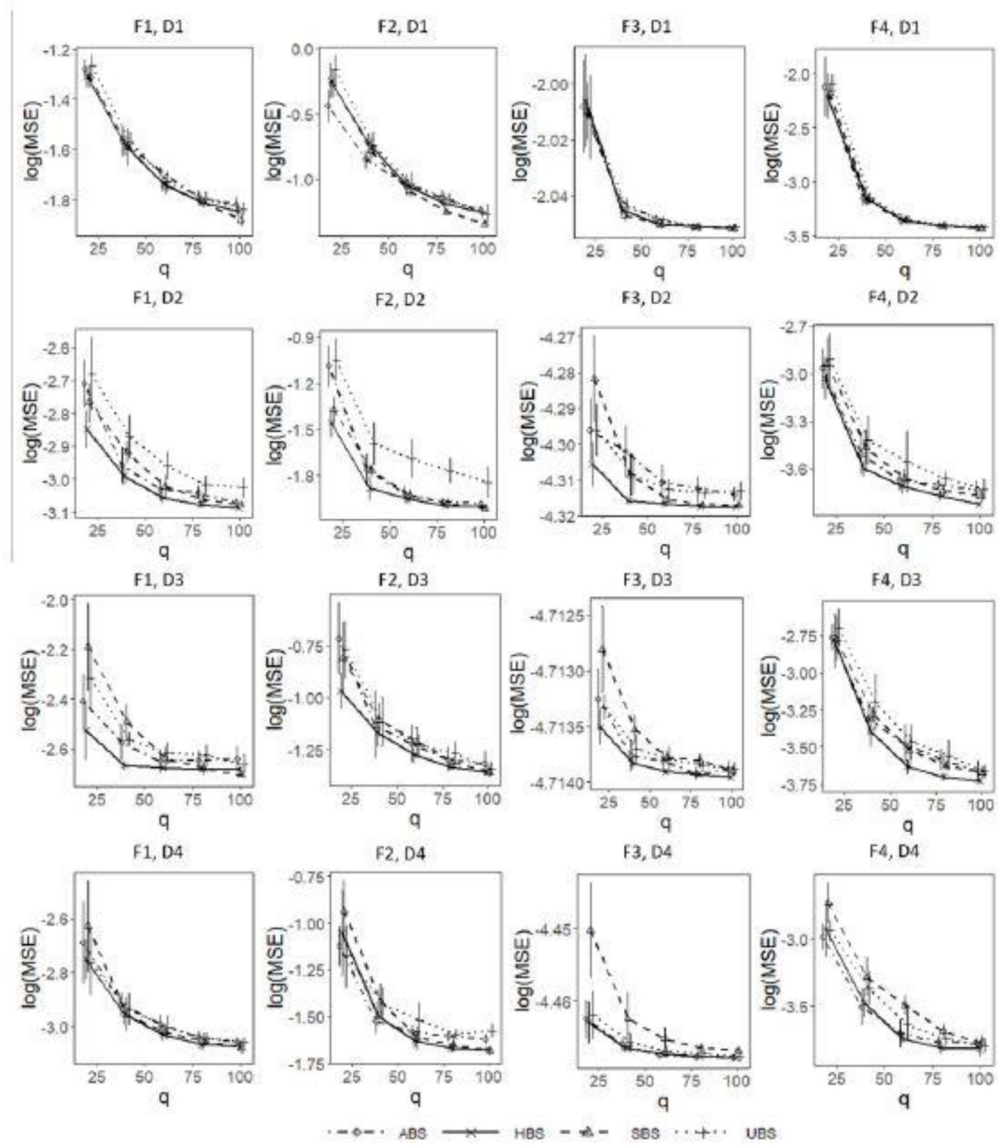
Fig. 3.
Illustration for Algorithm 1.

Fig. 4.
Simulation under different regression functions (from left to right) and different probability density functions for the predictors (from upper to lower). The prediction errors are plotted versus different q. Vertical bars represent the standard errors obtained from a hundred replicates.
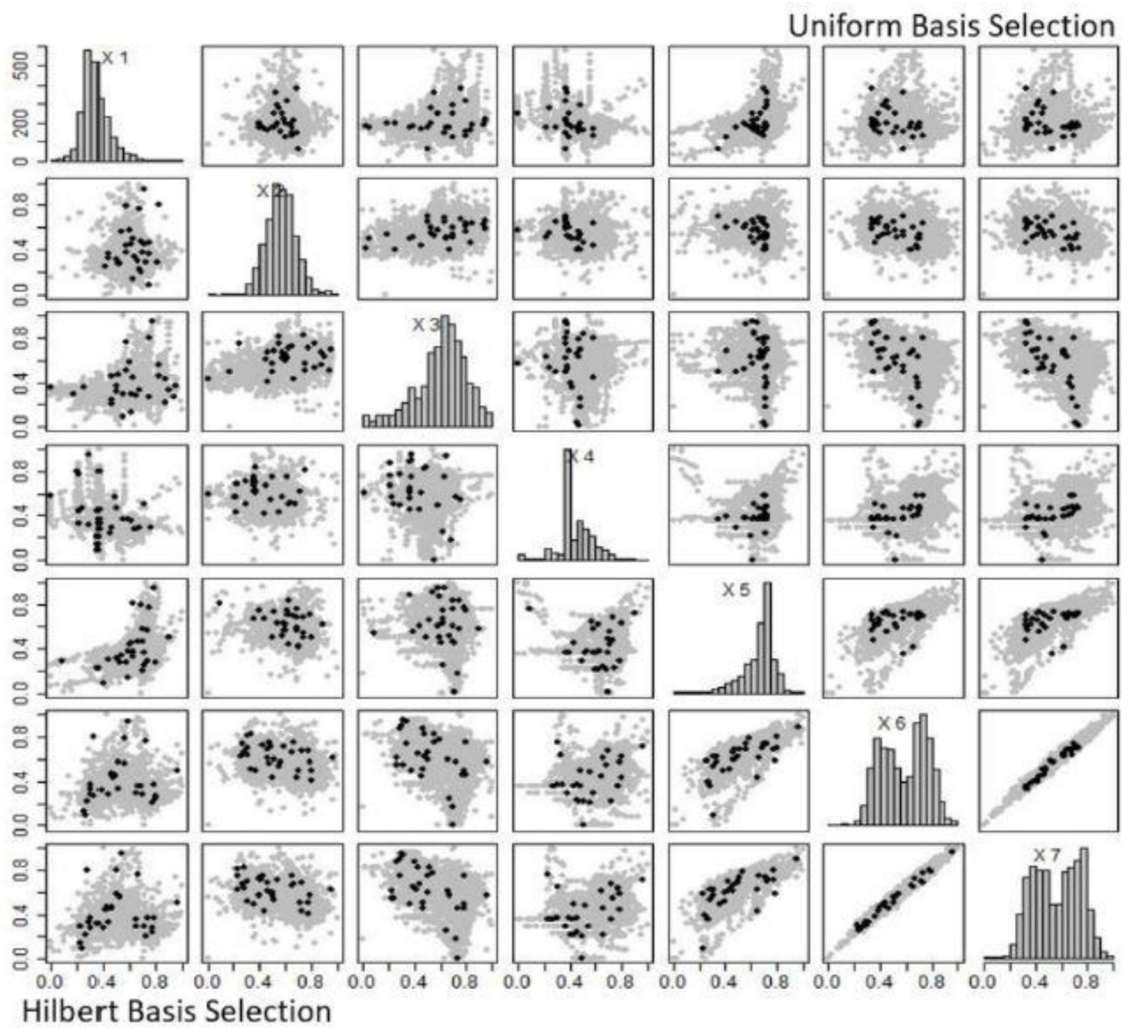
Fig. 5.
The diagonal panels show histograms for each of the predictors. The off-diagonal panels show the scatter plots corresponding to each pair of the predictors. The bases selected by the proposed method and the uniform basis selection method are shown in the lower diagonal panels and the upper diagonal panels, respectively. The black dots are the observations corresponding to the selected basis functions when $q = 40$.
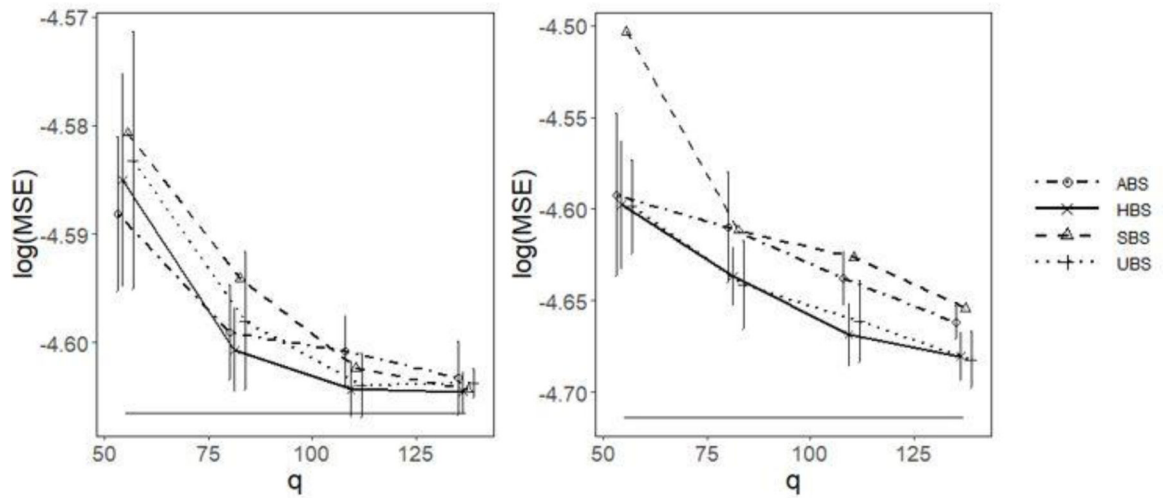
Fig. 6.
The left panel shows the log prediction MSE versus different q under model setting M1 for the debutanizer column dataset. The right panel shows the log prediction MSE versus different q under model setting M2. The horizontal lines represent the performance for the full sample estimator.