Smoothing Splines Approximation Using Hilbert Curve Basis Selection

Cheng Meng, Jun Yu, Yongkai Chen, Wenxuan Zhong & Ping Ma

View supplementary material

Published online: 12 Jan 2022.

Submit your article to this journal

Article views: 152

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

Check for updates

Smoothing Splines Approximation Using Hilbert Curve Basis Selection

Cheng Meng[a], Jun Yu[b], Yongkai Chen[c], Wenxuan Zhong[c], and Ping Ma[c]

[a]Institute of Statistics and Big Data, Renmin University of China, Beijing, China; [b]School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China; [c]Department of Statistics, University of Georgia, Athens, GA

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 hypercube. 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. Supplementary files for this article are available online.

1. Introduction

Smoothing spline estimators have been used pervasively in non-parametric regression models

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

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 iid 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 article.

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 use 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, Shen, and Shen (2001), Sklar et al. (2013), and Yuan, Chen, and Zhou (2013) considered the cases that the regression function has nonhomogeneous 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, Duchi, and Wainwright (2013) and Zhang, Duchi, and Wainwright (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, Shang, and Cheng (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, Shang, and Cheng (2018) and Liu, Shang, and Cheng (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, Hao, and Cheng 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 tradeoff 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, Li, and Ruppert (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, Huang, and Zhang (2015a) developed the uniform basis selection (UBS) method and the adaptive basis selection (ABS) 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 $d \leq pr + 1$. Their method aims to select approximately uniformly distributed observations by using space-filling designs or low-discrepancy sequences, resulting in a faster convergence rate compared with the UBS 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 hypercube 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 2000 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 Figure 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 UBS (Gu and Kim 2002), the ABS (Ma, Huang, and Zhang 2015a), and the space-filling basis selection (SBS) (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 Figure 1 show the heat maps of the spline estimates of all four basis selection methods, respectively. We observe the UBS method and the ABS 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 SBS method selects very few basis functions in the middle, resulting in degenerated performance in such an area. These observations
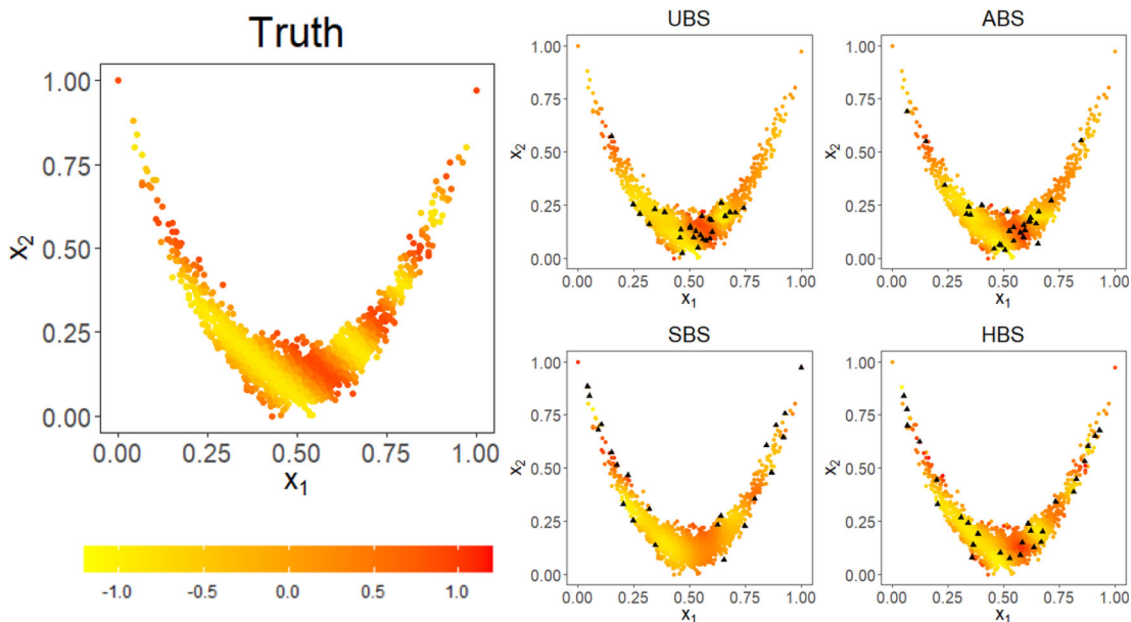

Figure 1. The leftmost panel shows the heat map for the true function. The heat maps for the spline estimates based on the UBS method, the ABS method, the SBS method, and the proposed HBS method are presented in the other panels, respectively. Black triangles are the selected basis functions. We observe that the proposed method outperforms the other methods in approximating the true function.

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 iid 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 iid 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 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 uses 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 UBS 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, Tibshirani, and Friedman 2009; Yang, Pilanci, and Wainwright 2017). Some simulation

results are provided in supplementary material to support this claim.

The rest of the article 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 using 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 Sections 5 and 6, respectively. Section 7 includes some discussion of the article. Proofs of the theorems are collected in the 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, Shen, and Ruppert 2011; Gu 2013; Wang, Du, and Shen 2013),

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

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 tradeoff 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, Shang, and Cheng (2019) generalized the GCV criterion to the setting of distributed learning. Recently, Sun, Zhong, and Ma (2021) proposed a more efficient approach for accelerating the calculation of λ.

In this article, we focus on minimizing the objective function (2) in a reproducing kernel Hilbert space, resulting in a smoothing spline estimate for η. Let $\mathcal{H} = \{\eta : J(\eta) < \infty\}$ be a reproducing kernel Hilbert space and $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ be the null space of $J(\eta)$. Let $\mathcal{N}_J$ be a m-dimensional linear subspace of $\mathcal{H}$ and $\{\xi_i\}_{i=1}^m$ be a set of basis for $\mathcal{N}_J$. Moreover, let $\mathcal{H}_J$ to denote the orthogonal complement of $\mathcal{N}_J$ in $\mathcal{H}$ such that $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$. It can be shown that $\mathcal{H}_J$ is a still a reproducing kernel Hilbert space, and we use $R_J(\cdot, \cdot)$ to denote the reproducing kernel of $\mathcal{H}_J$. Let $\mathbf{Y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ be the response vector, $\mathbf{S} \in \mathbb{R}^{n \times m}$ be a matrix where the (i, j)th element equals $\xi_j(\boldsymbol{x}_i)$, and $\mathbf{R} \in \mathbb{R}^{n \times n}$ 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 $\mathcal{H}$ 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 $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^{\mathrm{T}}$ 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

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

Although the solution of the minimization problem (3) has a closed form (Gu and Kim 2002; Ma, Huang, and Zhang 2015a),

the computational cost for calculating the solution is of the order $O(n^3)$, in a general case where $n \gg m$ and $d \geq 2$.

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, Mahoney, and Yu 2015b; Ma and Sun 2015; Meng et al. 2017; Zhang, Xie, and Ma 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, Liu, and Zeng 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 $\mathcal{H}_E = \mathcal{N}_J \oplus \text{span}\{R_J(x_i^*, \cdot), i = 1, \ldots, q\}$. Finally, the objective function (2) is minimized in the effective model space $\mathcal{H}_E$, 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 $\boldsymbol{\alpha}_E = (\alpha_1, \ldots, \alpha_m)^{\mathrm{T}}$ and $\boldsymbol{\beta}_E = (\beta_1, \ldots, \beta_q)^{\mathrm{T}}$ can be obtained through solving

$$(\hat{\boldsymbol{\alpha}}_E, \hat{\boldsymbol{\beta}}_E) = \underset{\boldsymbol{\alpha}_E \in \mathbb{R}^m, \, \boldsymbol{\beta}_E \in \mathbb{R}^q}{\text{argmin}} \frac{1}{n} (\mathbf{Y} - \mathbf{S}\boldsymbol{\alpha}_E - \mathbf{R}_*\boldsymbol{\beta}_E)^{\mathrm{T}}$$
$$\times (\mathbf{Y} - \mathbf{S}\boldsymbol{\alpha}_E - \mathbf{R}_*\boldsymbol{\beta}_E) + \lambda \boldsymbol{\beta}_E^{\mathrm{T}} \mathbf{R}_{**} \boldsymbol{\beta}_E, \quad (4)$$

where $\mathbf{R}_* \in \mathbb{R}^{n \times q}$ is a matrix where the (i, j)th element equals $R_J(x_i, x_j^*)$ and $\mathbf{R}_{**} \in \mathbb{R}^{q \times q}$ 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 hypercube. When such a condition is not met, they may suffer from deteriorating performance, as shown in Figure 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 $F : \Omega \rightarrow \mathbb{R}$ 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 $x_i, x_j \in \Omega$, $i, j \in \{1, \ldots, n\}$, 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 using 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 2013). Nowadays, space-filling curves have been widely used for computer graphics, approximately the 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 the 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 kth 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

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

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 cannot 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 using Hilbert space-filling curves, called the Hilbert basis selection (HBS) 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 preprocessing 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 nonempty bins
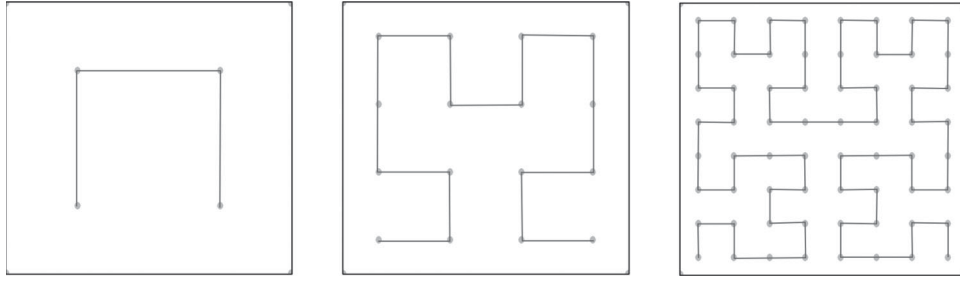
Figure 2. Illustration for the first three stages of the Hilbert's space filling curves when $d = 2$. From left to right, each panel shows H_1, H_2, and H_3, respectively.
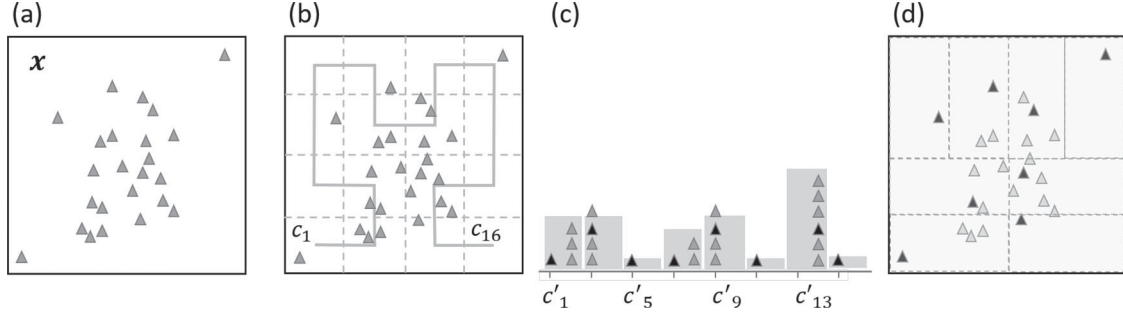

Figure 3. Illustration for Algorithm 1.

Algorithm 1 Hilbert basis selection method

Step 1. The sample $\{x_i\}_{i=1}^n$ 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 x_i, $i = 1, \ldots, n$, suppose x_i belongs to the block c_j, map 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 $\tilde{C}$ be the number of nonempty bins.

Step 5. Randomly select roughly $[q/\tilde{C}]$ number of data points from each nonempty bin; let $\{x_i^*\}_{i=1}^q$ to denote the selected ones.

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

---

and q be the subsample size. We then randomly select roughly $[q/\tilde{C}]$ data points from each nonempty 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.

Figure 3 gives an illustration of Algorithm 1, in which all data points are shown in Figure 3(a). We set $k = d = 2$ in Algorithm 1, resulting in $2^{2 \times 2} = 16$ blocks in Figure 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 Figure 3(c). Let the subsample size $q = 8$. We then draw $C = 8$ bins for the histogram in Figure 3(c), resulting in $\tilde{C} = 8$ nonempty bins. We then randomly select $\tilde{C}/q = 1$ data point from each bin. The selected data points are labeled as black triangles in both Figure 3(c) and 3(d). Note that each bin in Figure 3(c) is

associated with a "meta-block", as illustrated in Figure 3(d). As a result, when $\tilde{C} = q$, 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 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 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 nonempty 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 $V(g) = \int_\Omega g^2 f_X \mathrm{d}x$. A function $f(x)$ defined on Ω is said to be Lipschitz continuous, if for any $x, y \in \Omega$, 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 $r > 1$, $\rho_\nu > \beta\nu^r$ for sufficiently large ν;
- Condition 3. For all μ and ν, $\text{var}\{\phi_\nu(\boldsymbol{x})\phi_\mu(\boldsymbol{x})\} \leq B$, where ϕ_ν, ϕ_μ are the eigenfunctions associated with V and J in $\mathcal{H}$, B denotes a positive constant.
- Condition 4. For all μ and ν, $\phi_\mu(\boldsymbol{x})\phi_\nu(\boldsymbol{x}) \in L^2(\Omega)$, 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 $\phi_\nu \in \mathcal{H}$ 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 ρ_ν 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 $J(\eta) = \int_0^1 \eta''\,dx$, where η is a periodic function on $[0, 1]$. The eigenfunctions ϕ_νs 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, Huang, and Zhang (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 $\{H_k^{-1}(x_i)\}_{i=1}^n$ 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 $d \geq 2$, 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 nonempty bins in Step 4 of Algorithm 1, respectively. For brevity, throughout this section, we assume only one data point is selected from each nonempty 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 $\boldsymbol{x}_j^*$ lies in. Consider the estimator $\sum_{j=1}^q \{\tilde{n}_j/n\}\phi_\nu(\boldsymbol{x}_j^*)\phi_\mu(\boldsymbol{x}_j^*)$. Intuitively, such an estimator can be regarded as the mean estimator of the stratified sampling. This is because, for $j = 1, \ldots, q$, $\{\tilde{n}_j/n\}\phi_\nu(\boldsymbol{x}_j^*)\phi_\mu(\boldsymbol{x}_j^*)$ 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, $\sum_{j=1}^q \{\tilde{n}_j/n\}\phi_\nu(\boldsymbol{x}_j^*)\phi_\mu(\boldsymbol{x}_j^*)$ is an asymptotically unbiased estimate for $\int_\Omega \phi_\nu(\boldsymbol{x})\phi_\mu(\boldsymbol{x})f_X(\boldsymbol{x})d\boldsymbol{x}$. Furthermore, when $\Omega \subseteq [0, 1]^d$, we have

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

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

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

Consequently, Lemma 4.1 suggests that one can approximate the integration $\int_\Omega \phi_\nu(\boldsymbol{x})\phi_\mu(\boldsymbol{x})f_X(\boldsymbol{x})d\boldsymbol{x}$ more effectively, by calculating $\sum_{j=1}^q (n_j/n)\phi_\nu(\boldsymbol{x}_j^*)\phi_\mu(\boldsymbol{x}_j^*)$ instead of $\sum_{j=1}^q \phi_\nu(\boldsymbol{x}_j^+)\phi_\mu(\boldsymbol{x}_j^+)/q$. 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, Huang, and Zhang (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 UBS method (Gu and Kim 2002) and the ABS method (Ma, Huang, and Zhang 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 $q = O\{n^{2d/\{(pr+1)(d+2)\}+\delta}\}$, 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 UBS method (Gu and Kim 2002), the ABS method (Ma, Huang, and Zhang 2015a; Ma et al. 2017), and the SBS 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 $[\sum_{i=1}^{n} \{\hat{\eta}_E(t_i) - \eta_0(t_i)\}^2]/n$, 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 $(T_1, \ldots, T_d)$, 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 $\Sigma_{ij} = 0.9^{|i-j|}$, $i, j = \{1, \ldots, d\}$;
- D4: A banana-shape distribution, which is generated by $(Z_1, Z_2 + \frac{Z_1^2}{1.2}, \ldots, Z_d + \frac{Z_1^2}{1.2})$, 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

$$h_1(t_1, t_2) = \{0.75/(\pi \sigma_1 \sigma_2)\}$$
$$\times \exp\{-(t_1 - 0.2)^2/\sigma_1^2 - (t_2 - 0.3)^2/\sigma_2^2\},$$
$$h_2(t_1, t_2) = \{0.75/(\pi \sigma_1 \sigma_2)\}$$
$$\times \exp\{-(t_1 - 0.7)^2/\sigma_1^2 - (t_2 - 0.5)^2/\sigma_2^2\};$$

- F3: A 3-d function $\sin(\pi(x_1 + x_2 + x_3)/3) - x_1 - x_2^2$;
- F4: A 4-d function
$$x_1 + (2x_2 - 1)^2/2 + [\sin(10\pi x_3)/\{2 - \sin(10\pi x_3)\}]/3$$
$$+ \{0.1 \sin(2\pi x_4) + 0.2 \cos(4\pi x_4) + 0.3 \sin(6\pi x_4)^2$$
$$+ 0.4 \cos(8\pi x_4)^3 + 0.5 \sin(10\pi x_4)^3\}/4.$$

The signal-to-noise ratio, defined as $\text{var}\{\eta(X)\}/\sigma^2$, is set to be two. We find the results show similar patterns with a large range of signal-to-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 HBS method, dash-dotted lines to denote the ABS method, dashed lines to denote the SBS method, and dotted lines to denote the UBS method. 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 Figure 4. We first observe that all the methods perform similarly, while the UBS method performs slightly worse, in the first row of Figure 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 SBS method and the ABS method consistently outperform the UBS method. Nevertheless, in the lower three rows of Figure 4, we observe the UBS 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 $q \geq 30$ for robust prediction in practice. According to such a suggestion and consider the cases when $q \geq 30$ in Figure 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 $n = 2395$ and seven
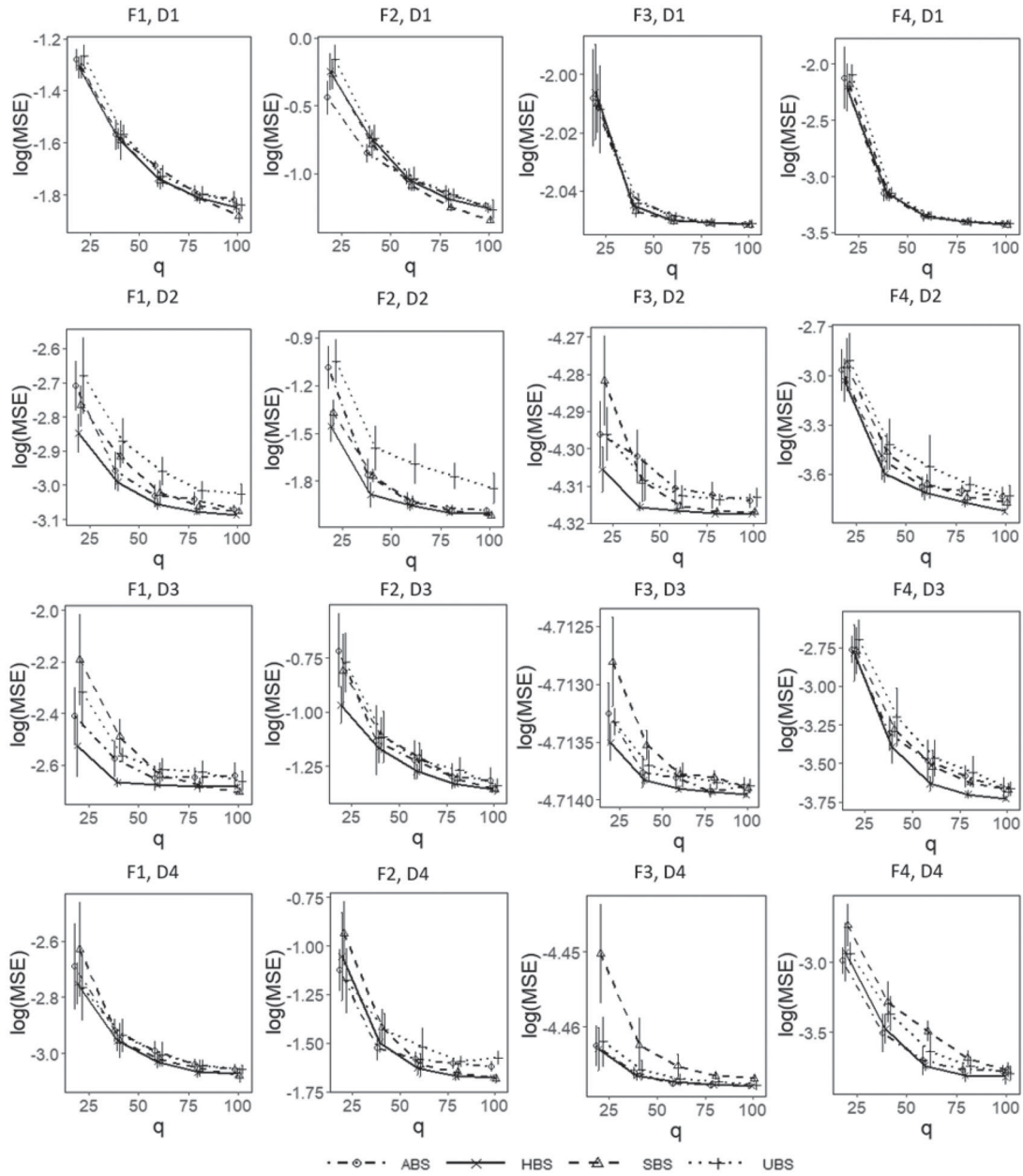
Figure 4. Simulation under different regression functions (from left to right) and different probability density functions for the predictors (from upper to lower). The prediction errors are plotted versus different q. Vertical bars represent the standard errors obtained from a hundred replicates.

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 scatterplots for each pair of the predictors in off-diagonal panels. We observe the sample in this dataset is extremely nonuniformly distributed. The basis functions selected by the proposed HBS method and the UBS method are marked as black dots in the lower diagonal panels and the upper panels, respectively. Compare with the UBS 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,

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

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

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

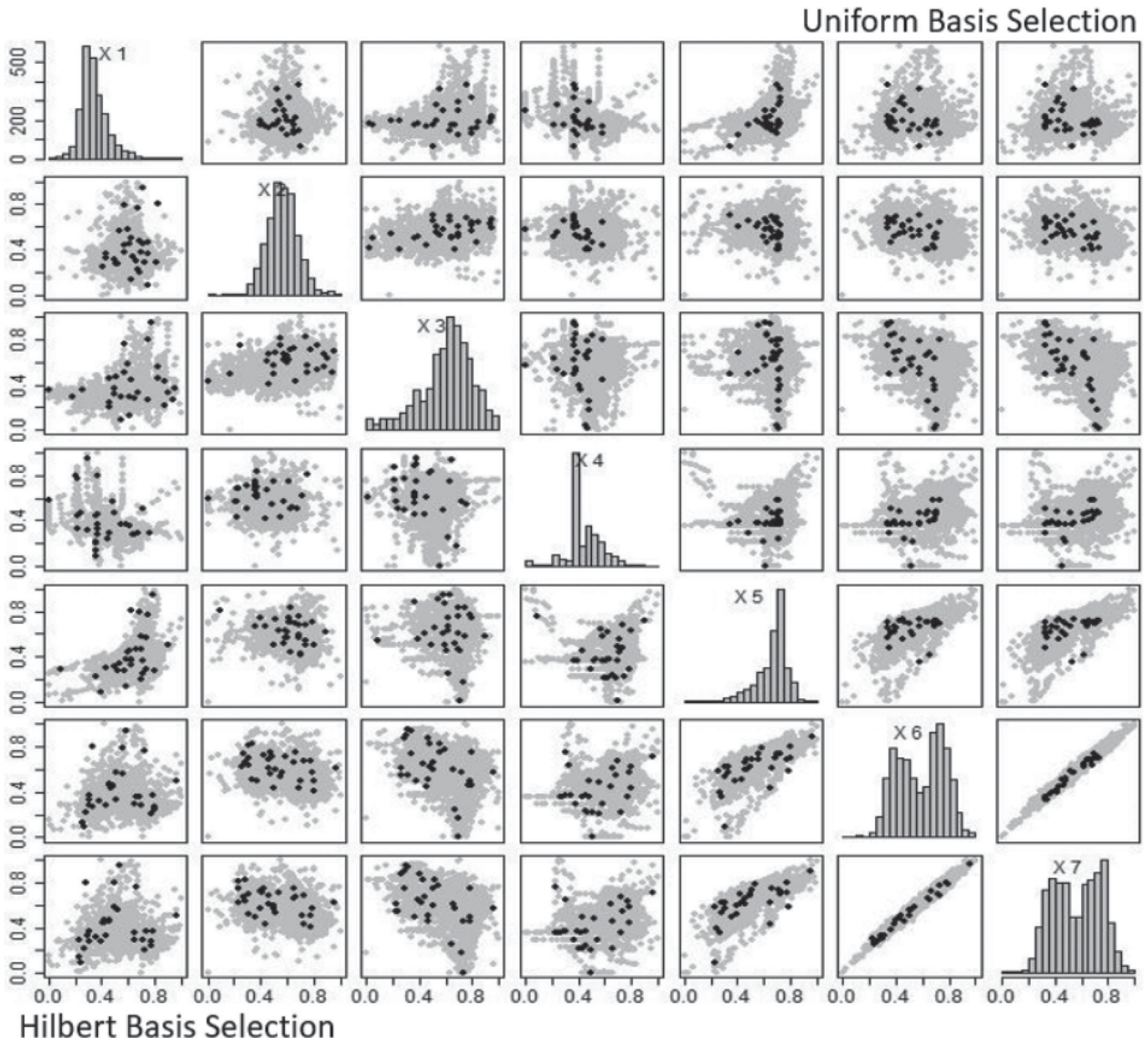[1]The dataset can be downloaded from *https://home.isr.uc.pt/~fasouza/datasets.html*

Figure 5. The diagonal panels show histograms for each of the predictors. The off-diagonal panels show the scatterplots 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 $q = 40$.
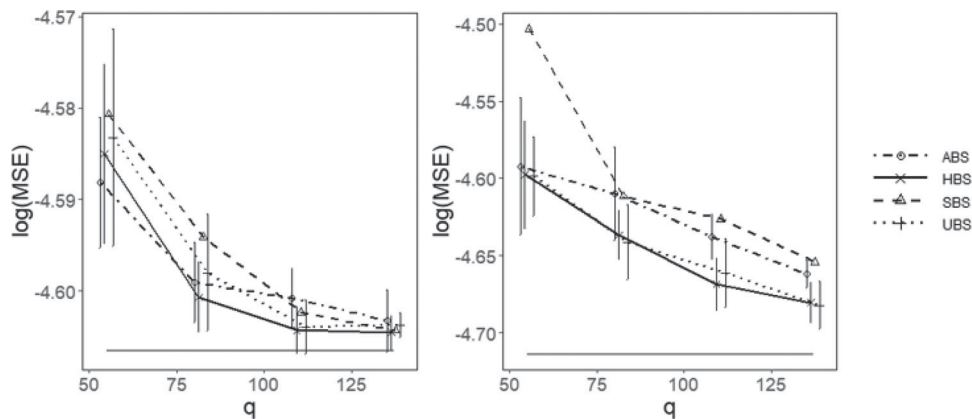

Figure 6. The left panel shows the log prediction MSE versus different q under model setting M1 for the debutanizer column dataset. The right panel shows the log prediction MSE versus different q under model setting M2. The horizontal lines represent the performance for the full sample estimator.

Here, the response y_i is the butane concentration of ith observation, x_{ij} is the value of the jth predictor of the ith observation, $\eta_\varnothing$ is a constant function, $\{\eta_j\}_{j=1}^7$ are main effect functions, $\eta_{1,3}, \eta_{1,5}, \eta_{1,6}$, and $\eta_{3,5}$ are two-way interaction functions of corresponding predictors, and ϵ_i's are iid 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 article, 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 using 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) used 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 nonuniform 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 the supplementary material.

ORCID

Ping Ma 🆔 http://orcid.org/0000-0002-5728-3596

References

Ai, M., Wang, F., Yu, J., and Zhang, H. (2021a), "Optimal Subsampling for Large-Scale Quantile Regression," *Journal of Complexity*, 62,101512. [4]

Ai, M., Yu, J., Zhang, H., and Wang, H. Y. (2021b), "Optimal Subsampling Algorithms for Big Data Regressions," *Statistica Sinica*, 31,749–772. [4]

Bader, M. (2013), *Space-Filling Curves: An Introduction With Applications in Scientific Computing*, Berlin: Springer-Verlag. [4]

Bernstein, D. S. (2009), *Matrix Mathematics: Theory, Facts, and Formulas*, Princeton, NJ: Princeton University Press. [1]

Drineas, P., Magdon-Ismail, M., Mahoney, M. W., and Woodruff, D. P. (2012), "Fast Approximation of Matrix Coherence and Statistical Leverage," *Journal of Machine Learning Research*, 13, 3475–3506. [4]

Eilers, P. H. C., and Marx, B. D. (2010), "Splines, Knots, and Penalties," *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 637–653. [3]

Fortuna, L., Graziani, S., Rizzo, A., and Xibilia, M. G. (2007), *Soft Sensors for Monitoring and Control of Industrial Processes*, Springer Science & Business Media. [8]

Geer, S. A., and van de Geer, S. (2000), *Empirical Processes in M-estimation*, Cambridge: Cambridge University Press. [3]

Golub, G. H., and Van Loan, C. F. (2013), *Matrix Computations*, Baltimore, MD: Johns Hopkins University Press. [1]

Gu, C. (2004), "Model Diagnostics for Smoothing Spline ANOVA Models," *Canadian Journal of Statistics*, 32, 347–358. [8]

——— (2013), *Smoothing Spline ANOVA Models*, New York: Springer Science & Business Media. [1,3,6,7]

——— (2014), "Smoothing Spline ANOVA Models: R Package gss," *Journal of Statistical Software*, 58, 1–25. [7]

Gu, C., and Kim, Y-J. (2002), "Penalized Likelihood Regression: General Formulation and Efficient Approximation," *Canadian Journal of Statistics*, 30, 619–628. [2,3,6,7]

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006), *A Distribution-Free Theory of Nonparametric Regression*, New York: Springer Science & Business Media. [3]

Hastie, T. (1996), "Pseudosplines," *Journal of the Royal Statistical Society*, Series B, 379–396. [1]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, New York: Springer. [3]

He, X., Shen, L., and Shen, Z. (2001), "A Data-Adaptive Knot Selection Scheme for Fitting Splines," *IEEE Signal Processing Letters*, 8, 137–139. [1]

Kahn, H., and Marshall, A. W. (1953), "Methods of Reducing Sample Size in Monte Carlo Computations," *Journal of the Operations Research Society of America*, 1, 263–278. [3]

Li, T., and Meng, C. (2020), "Modern Subsampling Methods for Large-Scale Least Squares Regression," *International Journal of Cyber-Physical Systems (IJCPS)*, 2, 1–28. [4]

Lin, Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Multivariate Nonparametric Regression," *The Annals of Statistics*, 34, 2272–2297. [7]

Liu, J. S. (1996), "Metropolized Independent Sampling With Comparisons to Rejection Sampling and Importance Sampling," *Statistics and Computing*, 6, 113–119. [3]

Liu, J. S. (2008), *Monte Carlo Strategies in Scientific Computing*, New York: Springer. [3]

Liu, M., Shang, Z., and Cheng, G. (2018), "How Many Machines Can We Use in Parallel Computing for Kernel Ridge Regression?" arXiv:1805.09948. [2]

Liu, M., Shang, Z., and Cheng, G. (2020), "Nonparametric Distributed Learning Under General Designs," *Electronic Journal of Statistics*, 14, 3070–3102. [2]

Luo, Z., and Wahba, G. (1997), "Hybrid Adaptive Splines," *Journal of the American Statistical Association*, 92, 107–116. [1]

Ma, P., and Sun, X. (2015), "Leveraging for Big Data Regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, 7, 70–76. [4]

Ma, P., Huang, J. Z., and Zhang, N. Efficient computation of smoothing splines via adaptive basis sampling. *Biometrika*, 102(3):631–645. [2,3,6,7]

Ping Ma, Michael W Mahoney, and Bin Yu. (2015a), "A Statistical Perspective on Algorithmic Leveraging," *Journal of Machine Learning Research*, 16, 861–911, 2015b. [4]

Ma, P., Zhang, N., Huang, J. Z., and Zhong, W. (2017), "Adaptive Basis Selection for Exponential Family Smoothing Splines With Application in Joint Modeling of Multiple Sequencing Samples," *Statistica Sinica*, 27, 1757–1777. [7]

Ma, P., Zhang, X., Xing, X., Ma, J., and Mahoney, M. W. (2020), "Asymptotic Analysis of Sampling Estimators for Randomized Numerical Linear Algebra Algorithms," *The 23nd International Conference on Artificial Intelligence and Statistics*, pp. 1026–1035. [4]

Mahoney, M. W. (2011), "Randomized Algorithms for Matrices and Data," *Foundations and Trends® in Machine Learning*, 3, 123–224. [4]

Meng, C., Wang, Y., Zhang, X., Mandal, A., Ma, P, and Zhong, W. (2017), "Effective Statistical Methods for Big Data Analytics," *Handbook of Research on Applied Cybernetics and Systems Science*, 280. [4]

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., and Ma, P. (2020a), "Lowcon: A Design-Based Subsampling Approach in a Misspecified Linear Model," *Journal of Computational and Graphical Statistics*, 1–15. [4]

Meng, C., Zhang, X., Zhang, J., Zhong, W., and Ma, P. (2020b), "More Efficient Approximation of Smoothing Splines Via Space-Filling Basis Selection," *Biometrika*, 107, 723–735. [2,6,7,10]

Reinsch, C. H. (1967), "Smoothing by Spline Functions," *Numerische mathematik*, 10, 177–183. [1]

Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735–757. [1]

Sagan, H. (2012), *Space-Filling Curves*, Springer Science & Business Media. [4]

Shang, Z., and Cheng, G. (2017), "Computational Limits of a Distributed Algorithm for Smoothing Spline," *The Journal of Machine Learning Research*, 18, 3809–3845. [2]

Shang, Z., Hao, B., and Cheng, G. (2019), "Nonparametric Bayesian Aggregation for Massive Data," *Journal of Machine Learning Research*, 20, 1–81. [2]

Sklar, J. C., Wu, J., Meiring, W., and Wang, Y. (2013), "Nonparametric Regression With Basis Selection From Multiple Libraries," *Technometrics*, 55, 189–201. [1]

Sun, X., Zhong, W., and Ma, P. (2021), "An Asymptotic and Empirical Smoothing Parameters Selection Method for Smoothing Spline ANOVA Models in Large Samples," *Biometrika*, 108, 149–166. [3]

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia, PA: SIAM. [1,3]

Wahba, G., and 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. [3]

Wang, X., Shen, J., and Ruppert, D. (2011), "On the Asymptotics of Penalized Spline Smoothing," *Electronic Journal of Statistics*, 5, 1–17. [3]

Wang, X., Du, P., and Shen, J. (2013), "Smoothing Splines With Varying Smoothing Parameter," *Biometrika*, 100, 955–970. [3]

Wang, Y. (2011), *Smoothing Splines: Methods and Applications*, Boca Raton: CRC Press. [1]

Wasserman, L. (2006), *All of Nonparametric Statistics*, Springer Science & Business Media. [3]

Wood, S. N. (2003), "Thin Plate Regression Splines," *Journal of the Royal Statistical Society*, Series B, 65, 95–114. [3,7]

Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. (2017), "Generalized Additive Models for Gigadata: Modeling the UK Black Smoke Network Daily Data," *Journal of the American Statistical Association*, 112, 1199–1210. [1]

Xiao, L., Li, Y., and Ruppert, D. (2013), "Fast Bivariate P-Splines: The Sandwich Smoother," *Journal of the Royal Statistical Society*, Series B, 75, 577–599. [2]

Xie, R., Wang, Z., Bai, S., Ma, P., and Zhong, W. (2019), "Online Decentralized Leverage Score Sampling for Streaming Multidimensional Time Series," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2301–2311. [4]

Xu, D., and Wang, Y. (2018), "Divide and Recombine Approaches for Fitting Smoothing Spline Models With Large Datasets," *Journal of Computational and Graphical Statistics*, 27, 677–683. [1,3]

Xu, G., Shang, Z., and Cheng, G. (2019), "Distributed Generalized Cross-Validation for Divide-and-Conquer Kernel Ridge Regression and Its Asymptotic Optimality," *Journal of Computational and Graphical Statistics*, 28, 891–908. [1,3]

Yang, Y., Pilanci, M., and Wainwright, M. (2017), "Randomized Sketches for Kernels: Fast and Optimal Nonparametric Regression," *The Annals of Statistics*, 45, 991–1023. [3]

Yu, J., Wang, H. Y., Ai, M., and Zhang, H. (2020), "Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators With Massive Data," *Journal of the American Statistical Association*, in press. doi: 10. 1080/01621459.2020.1773832. [4]

Yuan, Y., Chen, N., and Zhou, S. (2013), "Adaptive B-Spline Knot Selection Using Multi-Resolution Basis Set," *IIE Transactions*, 45, 1263–1277. [1]

Helen, H., Zhang, G. W., Lin, Y., Voelker, M., Ferris, M., Klein, R., and Klein, B. (2004), "Variable Selection and Model Building Via Likelihood Basis Pursuit," *Journal of the American Statistical Association*, 99, 659–672. [1]

Zhang, J., Jin, H., Wang, Y., Sun, X., Ma, P., and Zhong, W. (2018a), "Smoothing Spline ANOVA Models and Their Applications in Complex and Massive Datasets," *Topics in Splines and Applications*, 63, 63–82. [1]

Zhang, X., Xie, R., and Ma, P. (2018b), "Statistical Leveraging Methods in Big Data," in *Handbook of Big Data Analytics*, eds. W. Härdle, H. H. S. Lu and X. Shen, Springer, Cham, pp. 51–74. [4]

Zhang, Y., Duchi, J., and Wainwright, M. (2013), "Divide and Conquer Kernel Ridge Regression," in *Conference on Learning Theory*, Princeton, pp. 592–617. [1]

Zhang, Y., Duchi, J., and Wainwright, M. (2015), "Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm With Minimax Optimal Rates," *The Journal of Machine Learning Research*, 16, 3299–3340. [1]

Zhong, W., Liu, Y., and Zeng, P. (2021), "A Model-Free Variable Screening Method Based on Leverage Score," *Journal of the American Statistical Association*, 1–36. [4]

Zhou, S., and Shen, X. (2001), "Spatially Adaptive Regression Splines and Accurate Knot Selection Schemes," *Journal of the American Statistical Association*, 96, 247–259. [2]

Zumbusch, G. (2012), *Parallel Multilevel Methods: Adaptive Mesh Refinement and Loadbalancing*, Springer Science & Business Media. [4]