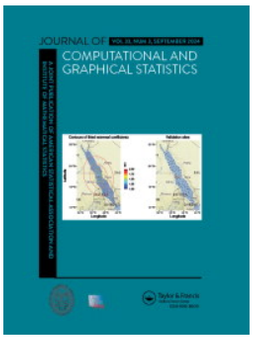


Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Nonparametric Additive Models for Billion Observations

Mengyu Li, Jingyi Zhang & Cheng Meng

To cite this article: Mengyu Li, Jingyi Zhang & Cheng Meng (19 Mar 2024): Nonparametric Additive Models for Billion Observations, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2024.2319684

To link to this article: https://doi.org/10.1080/10618600.2024.2319684

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 00, NO. 0, 1–16
https://doi.org/10.1080/10618600.2024.2319684

Nonparametric Additive Models for Billion Observations

Mengyu Li[a], Jingyi Zhang*[b], and Cheng Meng[c]

[a]Institute of Statistics and Big Data, Renmin University of China, Beijing, China; [b]Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China; [c]Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China

ABSTRACT

The nonparametric additive model (NAM) is a widely used nonparametric regression method. Nevertheless, due to the high computational burden, classic statistical techniques for fitting NAMs are not well-equipped to handle massive data with billions of observations. To address this challenge, we develop a scalable element-wise subset selection method, referred to as Core-NAM, for fitting penalized regression spline based NAMs. Specifically, we first propose an approximation of the penalized least squares estimation, based on which we develop an efficient variant of generalized cross-validation (GCV) to select the smoothing parameter and approximate the Bayesian confidence intervals for statistical inference. Theoretically, we show that the proposed estimator approximately minimizes an upper bound of the estimation mean squared error. Moreover, we provide a non-asymptotic approximation guarantee for the proposed estimator and establish the asymptotic optimality of the proposed variant of GCV. Extensive simulations demonstrate the superior accuracy and efficiency of the Core-NAM method. We also apply the proposed method to a total column ozone dataset containing nearly one billion observations, and the results indicate a speed-up by almost a thousand times with comparable performance compared to the full data approach. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2023
Accepted January 2024

KEYWORDS
Coresets; Penalized least squares; Regression spline; Subset selection

1. Introduction

Nonparametric additive models (NAMs) (Stone 1985; Hastie and Tibshirani 1990; Wood 2017) are a class of models that aim to capture the complex relationship between a response variable and covariates as a sum of smooth functions of the covariates. Compared to classical linear models, NAMs relax the strict linear assumption and possess better flexibility. Moreover, NAMs overcome the so-called "curse of dimensionality" that impedes the estimation of multivariate nonparametric models (Stone 1985).

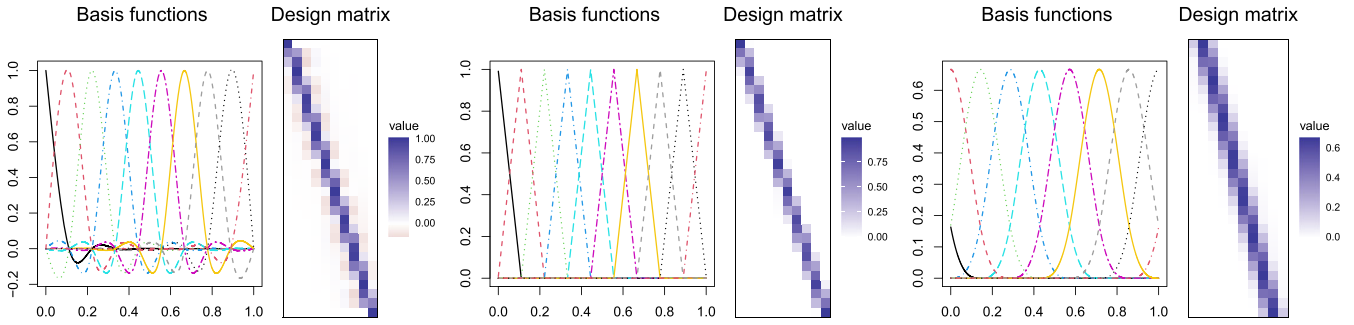
Despite the effectiveness, fitting NAMs using conventional methods can be computationally expensive, particularly when the sample size is considerable. For instance, given a sample with n observations, the time complexity of iterative backfitting algorithms (Breiman and Friedman 1985; Buja, Hastie, and Tibshirani 1989) is $O(n^2)$ per iteration, and that of penalized least squares is $O(n^3)$ using smoothing splines (Wahba 1990) or $O(nq^2)$ using regression splines (Wood and Augustin 2002), where q is the number of basis functions.

To address this computational bottleneck, many scalable methods have been developed for nonparametric regression, including but not limited to matrix decomposition (Wood, Goude, and Shaw 2015; Wood et al. 2017), parallelization (Wood, Goude, and Shaw 2015; Wood et al. 2017), and covariate discretization (Helwig and Ma 2016; Wood et al. 2017; Zhang et al. 2018). Furthermore, various basis selection approaches to approximate smoothing splines (Ma, Huang, and Zhang 2015; Ma et al. 2017; Meng et al. 2020, 2022; Diao et al. 2023) and efficient smoothing parameter selection methods (Wood 2004; Helwig and Ma 2015; Wood, Pya, and Säfken 2016; Sun, Zhong, and Ma 2021) have been proposed for NAMs or general smooth models. The variable selection problem has also been investigated extensively for NAMs (Meier, van de Geer, and Bühlmann 2009; Huang, Horowitz, and Wei 2010; Marra and Wood 2011; Fan, Feng, and Song 2011; Scheipl, Fahrmeir, and Kneib 2012; Dai, Lyu, and Li 2023). We refer to Perperoglou et al. (2019); Wood (2020) for recent reviews. Nevertheless, existing research on nonparametric regression has thus far only used datasets comprising at most 10 million observations as examples (Wood et al. 2017; Yang, Yao, and Zhao 2023), and more efficient approaches for analyzing super large-scale data with billion observations are still meager.

One solution to improve the computational efficiency is to fit the model to a subset of observations. Such a line of work is called the coresets approach, also referred to as subsampling or subset selection, which has been widely used for dealing with massive data of huge n. By effectively selecting representative observations from the full sample, the coresets not only greatly reduce computational burden (Ma and Sun 2015; Wang, Yang, and Stufken 2019; Wu et al. 2023; Zhang et al. 2023), but also play an essential role in measurement-constrained problems (Wang, Yu, and Singh 2017; Meng et al. 2021) and privacy-preserving

CONTACT Cheng Meng *chengmeng@ruc.edu.cn* Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China.
*Joint first author.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.



(a) Cubic regression spline. (b) P-spline (with the order of B-spline basis being 1). (c) P-spline (with the order of B-spline basis being 3).

Figure 1. Illustration of the locality of basis functions (left) and the sparsity of design matrices (right). Both the n observations and q knots are grid points evenly spaced on $[0, 1]$, where $n = 30$ and $q = 10$. Each subfigure corresponds to a specific choice of spline.

settings (Wang, Balle, and Kasiviswanathan 2019; Balle, Barthe, and Gaboardi 2020). Another possibility to reduce the computational cost is the online methods (Schifano et al. 2016; Kong and Xia 2019; Xue and Yao 2022; Yang and Yao 2023; Yang, Yao, and Zhao 2023), which are beyond the scope of this article.

A variety of randomized and deterministic coresets methods for regression have been proposed in the past decade. Although existing coresets approaches have proven to be effective theoretically and empirically, their proper use is mostly restricted to parametric regression, such as linear models (Dasgupta et al. 2009; Boutsidis, Drineas, and Magdon-Ismail 2013; Ma and Sun 2015; Meng et al. 2017; Dereziński, Warmuth, and Hsu 2018; Wang, Yang, and Stufken 2019; Ma et al. 2020; Wang et al. 2021), generalized linear models (Wang, Zhu, and Ma 2018; Ai et al. 2021; Yu et al. 2022), quantile regression (Wang and Ma 2021; Ai et al. 2021), streaming time series (Xie et al. 2019; Xie, Bai, and Ma 2023), among others. We refer to Li and Meng (2021); Yu, Ai, and Ye (2023) for a comprehensive overview. A few exceptions to this trend include several model-free coresets methods without requiring explicit model assumptions (Mak and Joseph 2018; Joseph and Mak 2021; Dai, Song, and Wang 2023), as well as a recently developed independence-encouraging subsampling method (Zhang et al. 2023) designed for NAMs.

However, considering that widely used basis functions (e.g., the cubic regression spline basis and the B-spline basis) are typically local (Wood 2017), that is, each basis function only has relatively large values over a small interval (whose size depends on q) and is close (or equal) to zero in the remaining domain, it follows that the induced design matrix is often numerically (or strictly) sparse[1] (as illustrated in Figure 1). For such design matrices of high sparsity, the coresets methods relying on row-wise selection may become inefficient. More precisely, when the design matrix is sparse, the selected row-wise subset also tends to be sparse, in which the near zero elements have almost no impact on downstream matrix computations, leading to inefficient results. To overcome this obstacle, Li et al. (2023) proposed a novel element-wise subset selection method for sparse design matrices in classical linear models, named "core-elements." In this article, by realizing the inherent sparsity of the design matrix arising from basis evaluations, we propose an efficient and scalable approach for approximating penalized least squares estimation in additive models.

Major contributions. We summarize our contributions as follows. *First, we design a novel* Core*-NAM method for NAMs with improved scalability.* By selecting core-elements from the full sample and leveraging sparse matrix operations, the proposed method efficiently approximates the penalized least squares estimation and chooses the smoothing parameter through a core-elements generalized cross-validation (Core-GCV) within constantly low time and space. *Second, we establish the theoretical properties of the proposed method.* We show the asymptotic optimality of our proposed Core-GCV procedure; that is, under certain regularity conditions, minimizing the Core-GCV score is asymptotically equivalent to minimizing the "golden criterion," loss function, in the sense of expectation. Moreover, we provide a relative error bound for the core-elements estimation. *Third, we conduct numerical experiments on billion observations.* Extensive simulations demonstrate the effectiveness of our method in model fitting, prediction, and Bayesian confidence interval approximation. In addition, we consider a newly released total column ozone (TCO) dataset (Bodeker et al. 2023), which records daily ozone measurements spanning the period from 1978 to 2019, containing nearly a billion observations. Compared with the full data approach, the proposed Core-NAM method requires only 0.11% of the running time to obtain almost as accurate predictions, demonstrating a decent tradeoff between speed and accuracy.

The remainder of this article is organized as follows. We begin by introducing the additive model and core-elements in Section 2. In Section 3, we develop the main algorithm for additive models, and its theoretical properties are discussed in Section 4. We evaluate the performance of the proposed estimator through extensive synthetic and real-world datasets in Sections 5 and 6, respectively. The technical proofs and additional numerical results are provided in the supplementary materials. R codes to reproduce the numerical results in this article are provided at this link: *https://github.com/Mengyu8042/Core-NAM*.

[1] Intuitively, "numerically sparse" is a relaxation of "sparse" that allows many small (but nonzero) elements; see Gupta and Sidford (2018) and Carmon et al. (2020) for a detailed definition.

2. Background

Here we summarize the notation used throughout the article. We adopt the common convention of using uppercase boldface letters for matrices, lowercase boldface letters for vectors, and regular font for scalars. For a vector $\boldsymbol{x}$, we denote its ℓ_p norm by $\|\boldsymbol{x}\|_p$ and abbreviate its Euclidean norm (i.e., ℓ_2 norm) by $\|\boldsymbol{x}\|$. The spectral norm and Frobenius norm of a matrix $\mathbf{X}$ are denoted as $\|\mathbf{X}\|_2$ and $\|\mathbf{X}\|_F$, respectively. Besides, we use $\kappa(\mathbf{X})$ to represent the condition number of $\mathbf{X}$, that is, the ratio of the largest singular value to the smallest singular value.

2.1. Nonparametric Additive Model

For iid data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$, the additive model takes the form[2]

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i \text{ with } f(\boldsymbol{x}_i) = \sum_{j=1}^{p} f_j(x_{ij}), \tag{1}$$

where y_i are the responses, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ contain p covariates, $f_j(\cdot)$ are smooth functions of the jth covariate, and ε_i are random noises with zero mean and constant variance σ^2.

Standard spline-based approaches to estimate the model (1) employ either smoothing splines (Reinsch 1967; Wahba 1990; Green and Silverman 1993; Gu 2013) or penalized regression splines (Eilers and Marx 1996; Wood and Augustin 2002; Wood 2003), among which the former class requires as many parameters as the sample size n, while the latter one uses $q \ll n$ parameters. In practice, the latter is often preferable for computational convenience when the size of n is considerable. Therefore, we focus on the penalized regression splines in the following.

Penalized regression splines. To estimate the smooth function $f(\cdot)$, each $f_j(\cdot)$ is assumed to have a representation

$$f_j(x) = \sum_{k=1}^{q} b_{jk}(x)\beta_{jk}, \tag{2}$$

where $b_{j1}(\cdot), \ldots, b_{jq}(\cdot)$ are q chosen basis functions defined on a sequence of q knots, and $\beta_{j1}, \ldots, \beta_{jq}$ are unknown parameters. Substituting (2) into (1) yields a linear model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \ \ \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2\mathbf{I}). \tag{3}$$

Here, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ is the response vector, $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_p) \in \mathbb{R}^{n \times pq}$ is a design matrix arising from basis functions evaluated at $\{\boldsymbol{x}_i\}_{i=1}^n$, that is, the (i,k)th component of $\mathbf{X}_j \in \mathbb{R}^{n \times q}$ is $b_{jk}(x_{ij})$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top)^\top \in \mathbb{R}^{pq}$ is a vector of parameters with $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jk})^\top \in \mathbb{R}^q$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$ is the noise vector. Without loss of generality, we assume that $n \gg d := pq$ and p is fixed.

One key choice in (2) is the basis dimension q (i.e., the number of knots) (Perperoglou et al. 2019). To prevent overfitting and under-fitting, a hugely popular approach to facilitate the choice of q is to use a relatively large number of knots and control the model smoothness by adding a wiggliness penalty to the least squares objective (Eilers and Marx 1996; Wood 2017; Perperoglou et al. 2019). Such a method is called the penalized least squares (PLS) criterion, which is to minimize the objective

$$\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta}, \tag{4}$$

where $\mathbf{S}_j \in \mathbb{R}^{pq \times pq}$ are known penalty matrices such that $\boldsymbol{\beta}^\top \mathbf{S}_j \boldsymbol{\beta}$ measure the wiggliness of $f_j(\cdot)$, and $\lambda_j \geq 0$ are smoothing parameters that balance between model fitting and model smoothness, for $j = 1, \ldots, p$. For example, a prevalent class of penalized regression splines are P-splines (Eilers and Marx 1992, 1996, 2010), in which the basis functions are B-spline bases, and the nonzero block of $\mathbf{S}_j$ equals $\mathbf{D}_j^\top \mathbf{D}_j$, where $\mathbf{D}_j$ is the matrix representation of the δth difference operator Δ^δ ($\delta \in \mathbb{Z}_+$).

Smoothing parameter selection. Once the smoothing parameter $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$ is given, the objective (4) is readily minimized to obtain the PLS estimation[3]

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = (\mathbf{X}^\top\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{X}^\top\boldsymbol{y}, \tag{5}$$

where $\mathbf{S}_{\boldsymbol{\lambda}} = \sum_{j=1}^{p} \lambda_j \mathbf{S}_j$. Some popular criteria to choose the "optimal" value of $\boldsymbol{\lambda}$ include Mallows' C_L (Mallows 1973), cross-validation (CV) or generalized cross-validation (GCV) (Stone 1974; Wahba and Wold 1975; Craven and Wahba 1978; Golub, Heath, and Wahba 1979), restricted maximum likelihood (REML) (Wahba 1985; Wood 2011; Gu 2013), and more. Considering the desirable properties of GCV (Craven and Wahba 1978; Li 1987; Gu and Wahba 1991; Gu 2013; Patil et al. 2021), we focus on this criterion, that is, to minimize the GCV score

$$\mathcal{V}(\boldsymbol{\lambda}) = \frac{n\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|^2}{[n - \mathrm{Tr}(\mathbf{H}_{\boldsymbol{\lambda}})]^2} \tag{6}$$

with respect to (w.r.t.) the smoothing parameter, where $\mathrm{Tr}(\mathbf{H}_{\boldsymbol{\lambda}})$ is the effective degrees of freedom of the model, and $\mathbf{H}_{\boldsymbol{\lambda}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{X}^\top$ is the hat matrix, satisfying that $\mathbf{H}_{\boldsymbol{\lambda}}\boldsymbol{y} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$. Note that calculating either $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ or $\mathcal{V}(\boldsymbol{\lambda})$ involves a computing time of the order $O(nq^2)$, so the overall time complexity of solving $\min_{\boldsymbol{\lambda}} \mathcal{V}(\boldsymbol{\lambda})$ is $O(\zeta nq^2)$, where ζ is the number of iterations required for the optimization problem to converge. In the face of extremely large datasets, the computational cost of the order $O(nq^2)$ for a single iteration may be prohibitively large, let alone to calculate it iteratively. Consequently, computation is a major bottleneck for applying additive models on massive data.

2.2. Core-Elements for Linear Models

Recognizing the limitations of coresets strategies when applied to sparse design matrices, Li et al. (2023) proposed a core-elements method for approximating the ordinary least squares (OLS) estimation in linear models (3).

Let $\mathbf{X}^* \in \mathbb{R}^{n \times d}$ be a sparse sketch of the design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. In particular, given a sampling budget $s \in \mathbb{Z}_+$ (i.e.,

[2] To simplify the presentation, the model (1) only includes the main-effects, but it should be emphasized that the proposed method also allows for incorporating interaction terms between covariates, for example, $f_{jk}(x_{ij}, x_{ik})$.

[3] Without loss of generality, we assume that $\mathbf{X}^\top\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}}$ is positive definite. Otherwise, it is straightforward to replace the inverse with a generalized inverse.

the number of selected elements), let $\mathbf{P} \in \mathbb{R}^{n\times d}$ be a binary matrix such that its elements involve s ones and $(nd - s)$ zeros. The sketch $\mathbf{X}^*$ then is formulated as $\mathbf{X}^* = \mathbf{P} \odot \mathbf{X}$, where $\odot$ represents the Hadamard product, that is, element-wise product. It is assumed that both $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^{*\top}\mathbf{X}$ are nonsingular.

Starting from a general estimation based on $\mathbf{X}^*$, taking the form $\widetilde{\boldsymbol{\beta}}(\mathbf{D}) = \mathbf{D}\mathbf{X}^{*\top}\boldsymbol{y}$, where $\mathbf{D} \in \mathbb{R}^{d\times d}$ is a scaling matrix to be determined, the core-elements estimation proposed by Li et al. (2023) is motivated by two fundamental properties: unbiasedness and effectiveness. First, to ensure $\widetilde{\boldsymbol{\beta}}(\mathbf{D})$ is unbiased to the true parameter $\boldsymbol{\beta}$, the scaling matrix $\mathbf{D}$ is set to be $(\mathbf{X}^{*\top}\mathbf{X})^{-1}$. Subsequently, an upper bound for the variance of the estimation

$$\widetilde{\boldsymbol{\beta}} = (\mathbf{X}^{*\top}\mathbf{X})^{-1}\mathbf{X}^{*\top}\boldsymbol{y} \tag{7}$$

is derived and approximately minimized, leading to the principle of core-elements selection; that is, the sparse sketch $\mathbf{X}^*$, containing at most s nonzero elements, keeps the elements with the top $\lfloor s/d \rfloor$ largest absolute values in each column of $\mathbf{X}$ while setting the remaining elements to zero. The proposed core-elements estimation has theoretical approximation guarantees w.r.t. the full sample OLS estimation, and it exhibits superior performance over mainstream competitors in empirical experiments, particularly when applied to sparse design matrices.

3. Method

In Section 2.2, we introduced the core-elements approach designed for sparse matrices in classical linear models. Encouragingly, the design matrices in NAMs exhibit natural sparsity (as discussed in Section 1), making it reasonable to extend the idea of core-elements to accelerate the estimation of NAMs. Nonetheless, this extension is not straightforward mainly for two reasons. First, compared to the OLS estimation, PLS involves an additional penalty term, necessitating a new form of core-elements estimator with theoretical guarantees. Second, the main computational bottleneck in NAMs is the selection of smoothing parameters, which is significantly more time-consuming than computing the PLS estimation itself. Therefore, developing a fast GCV approach based on core-elements is of utmost importance.

In this section, we present our main algorithm. We first develop the core-elements estimation for NAMs and introduce the principle to select core-elements motivated by approximately minimizing the mean squared error (MSE). Based on the proposed estimator, we further design a core-elements GCV procedure to select the smoothing parameter efficiently.

Core-elements estimation. Inspired by the formulation (7), we propose an approximation for the PLS estimation (5) based on the sparse sketch $\mathbf{X}^*$, taking the form

$$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = (\mathbf{X}^{*\top}\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{X}^{*\top}\boldsymbol{y}. \tag{8}$$

Similar to that in (4), here we assume that $\mathbf{X}^{*\top}\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}}$ is of full rank. Starting on the formulation (8), our goal is to find a sketch $\mathbf{X}^*$ that approximately minimizes the MSE of $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$, defined as $\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) = \mathbb{E}(\|\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}\|^2)$. Although the MSE has a closed form (see supplementary materials for details), directly minimizing it poses a significant challenge. To surmount the obstacle, we provide an upper bound for $\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$ in Proposition 1 and aim to minimize this upper bound instead.

Proposition 1. Let $\mathbf{L} = \mathbf{X} - \mathbf{X}^*$. A Taylor expansion of $\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$ around the point $\mathbf{X}^* = \mathbf{X}$ yields the following upper bound,

$$\begin{aligned}&\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\\ &\le \underbrace{\sigma^2\{[1+O(\gamma_0)](\|\mathbf{X}\mathbf{D}\|_F^2 + \|\mathbf{D}\|_2^2\|\mathbf{L}\|_F^2) + O(\gamma_0)\,\mathrm{Tr}(\mathbf{D})\}}_{\text{Upper bound of variance}}\\ &+ \underbrace{[1+O(\gamma_0)]\|\mathbf{D}\mathbf{S}_{\boldsymbol{\lambda}}\boldsymbol{\beta}\|^2}_{\text{Upper bound of square bias}},\end{aligned} \tag{9}$$

where $\mathbf{D} = (\mathbf{X}^\top\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}$, and the spectral radius $\gamma_0 = \|\mathbf{D}\mathbf{L}^\top\mathbf{X}\|_2$ is assumed to be smaller than one to guarantee the convergence of the matrix series.

When $\mathbf{X}^* = \mathbf{X}$, it can be shown that the upper bound in (9) equals $\mathrm{MSE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$. Otherwise, this upper bound decreases as $\|\mathbf{L}\|_F$ and the remainder γ_0 decreases. The γ_0 can be further bounded by

$$\gamma_0 \le \|\mathbf{D}\|_2\|\mathbf{X}\|_2\|\mathbf{L}\|_2 \le \|\mathbf{D}\|_2\|\mathbf{X}\|_2\left(d\max_{j\in\{1,\dots,d\}}\mathbf{L}^{(j)\top}\mathbf{L}^{(j)}\right)^{1/2},$$

where $\mathbf{L}^{(j)}$ denotes the jth column of $\mathbf{L}$. Such an inequality indicates that a smaller value of the maximum column norm of $\mathbf{L}$ is associated with a smaller γ_0. As a result, to minimize the upper bound of $\mathrm{MSE}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$, we need to keep both $\|\mathbf{L}\|_F$ and the column norms of $\mathbf{L}$ as small as possible. This leads to a core-elements selection criterion aligning with that in Li et al. (2023): given the sampling budget s, the sketch $\mathbf{X}^*$ is constructed by retaining $\lfloor s/d \rfloor$ elements with the top largest absolute values w.r.t. each column of $\mathbf{X}$ and zeroing out the remaining. Intuitively, such $\mathbf{L}$ has the approximately minimum column norm respecting every column. Hence, both the values of $\|\mathbf{L}\|_F$ and $\|\mathbf{L}\|_2$ are approximately minimized, resulting in a relatively small upper bound of MSE in Proposition 1.

Core-elements GCV. Next, we consider the smoothing parameter selection based on core-elements. Let r_s be the number of rows containing nonzero elements in $\mathbf{X}^*$. By extracting these r_s rows from $\mathbf{X}^*$ and $\mathbf{X}$, and corresponding elements from $\boldsymbol{y}$, we denote $\mathbf{X}_s^* \in \mathbb{R}^{r_s\times d}$, $\mathbf{X}_s \in \mathbb{R}^{r_s\times d}$, and $\boldsymbol{y}_s \in \mathbb{R}^{r_s}$, respectively. Similar to (6), we define the objective of core-elements GCV (Core-GCV) as

$$\mathcal{V}_s(\boldsymbol{\lambda}) = \frac{r_s\|\boldsymbol{y}_s - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|^2}{[r_s - \mathrm{Tr}(\mathbf{H}_{\boldsymbol{\lambda}}^*)]^2}, \tag{10}$$

where the hat matrix $\mathbf{H}_{\boldsymbol{\lambda}}^*$ is defined as $\mathbf{H}_{\boldsymbol{\lambda}}^* = \mathbf{X}_s(\mathbf{X}_s^{*\top}\mathbf{X}_s + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{X}_s^{*\top}$. Note that the core-elements estimation (8) can also be written as $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} = (\mathbf{X}_s^{*\top}\mathbf{X}_s + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{X}_s^{*\top}\boldsymbol{y}_s$, so the condition $\mathbf{H}_{\boldsymbol{\lambda}}^*\boldsymbol{y}_s = \mathbf{X}_s\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ is also satisfied here.

Combining the above mentioned procedures, Algorithm 1 summarizes the Core-NAM method for additive models. This algorithm is applicable to a broad spectrum of basis functions, such as cubic regression splines and P-splines for univariate cases, and tensor product bases for interaction modeling. The basis functions are located on q evenly spaced knots spanning the domain. Furthermore, we investigate smoothness-adaptive knot selection, with methodology and simulation details available in the supplementary materials.

Algorithm 1 CORE-NAM method for additive models

1: **Input:** $\mathbf{X} = (X_{ij}) \in \mathbb{R}^{n\times d}$ with $d = pq$, $\boldsymbol{y} \in \mathbb{R}^n$, $r = \lfloor s/d \rfloor \in \mathbb{Z}_+$
2: Construct the sparse sketch $\mathbf{X}^*$: $O(\text{nnz}(\mathbf{X}))$
   a) Initialize $\mathbf{P} = (P_{ij}) = \mathbf{0}_{n\times d}$
   b) For $j = 1, \ldots, d$:
      Let $\mathcal{I} = \{i_1, \ldots, i_r\}$ be an index set, such that $\{|X_{i_k j}|\}_{k=1}^r$ are the largest r elements among $\{|X_{ij}|\}_{i=1}^n$
      Let $P_{ij} = 1$, for $i \in \mathcal{I}$
   c) $\mathbf{X}^* = \mathbf{P} \odot \mathbf{X}$, where $\odot$ represents the Hadamard product
3: Compute $\widetilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \mathcal{V}_s(\boldsymbol{\lambda})$, where $\mathcal{V}_s(\boldsymbol{\lambda})$ is defined by (10) $O(\zeta r q^2 + \zeta q^3)$
4: **Return** $\widetilde{\boldsymbol{\beta}}_{\widetilde{\boldsymbol{\lambda}}} = (\mathbf{X}^{*\top}\mathbf{X} + \mathbf{S}_{\widetilde{\boldsymbol{\lambda}}})^{-1}\mathbf{X}^{*\top}\boldsymbol{y}$ $O(rq^2 + q^3)$

Computational cost. For Step 2 in Algorithm 1, constructing the sketch $\mathbf{X}^*$ by using a partition-based selection algorithm requires $O(\text{nnz}(\mathbf{X}))$ computational time (Musser 1997; Martinez 2004; Wang, Yang, and Stufken 2019), which can be easily paralleled. For Steps 3 and 4, because each column of $\mathbf{X}^*$ contains at most r nonzero elements, calculating $\mathbf{X}^{*\top}\mathbf{X}$ takes $O(rq^2)$ time by using sparse matrix representations and operations (Bates and Eddelbuettel 2013). The computational cost of $\widetilde{\boldsymbol{\beta}}$ or $\mathcal{V}_s(\boldsymbol{\lambda})$ is thus of the order $O(rq^2 + q^3)$. Therefore, the overall time complexity of Algorithm 1 is $O(\text{nnz}(\mathbf{X}) + \zeta rq^2 + \zeta q^3)$, where ζ is the number of iterations to solve the optimization problem in Step 3. In addition, the memory cost of Algorithm 1 is $O(\text{nnz}(\mathbf{X}) + r_s q)$. Once the core-elements are prepared, the estimation only requires constant low time and space when r and q are fixed.

The CORE-NAM approach can also be used to efficiently approximate the Bayesian confidence interval (Wahba 1983; Wood 2006b); see the supplementary materials for details.

4. Theoretical Results

In this section, we first show the asymptotic optimality of the proposed CORE-GCV criterion. Then, we demonstrate that the core-elements estimation achieves the $(1+\epsilon)$-approximation w.r.t. the full sample estimation (5). Technical proofs are provided in supplementary material.

4.1. Optimality of Core-Elements GCV

Despite the "golden criterion" to choose $\boldsymbol{\lambda}$ is to minimize the loss function

$$\mathcal{L}_s(\boldsymbol{\lambda}) = \frac{1}{r_s}\|\mathbf{X}_s\boldsymbol{\beta} - \mathbf{X}_s\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|^2, \tag{11}$$

this cannot be done directly because $\mathcal{L}_s(\cdot)$ involves the unknown $\boldsymbol{\beta}$. However, Theorem 1 shows that minimizing the expected CORE-GCV score is asymptotically equivalent to minimizing the expected loss function.

Theorem 1. As $s \to \infty$ and $q = o(s)$, under the conditions:

(i) $r_s^{-1}\,\text{Tr}(\mathbf{H}_{\boldsymbol{\lambda}}^*) \to 0$;
(ii) $[r_s^{-1}\,\text{Tr}(\mathbf{H}_{\boldsymbol{\lambda}}^*)]^2/[r_s^{-1}\,\text{Tr}(\mathbf{H}_{\boldsymbol{\lambda}}^*\mathbf{H}_{\boldsymbol{\lambda}}^{*\top})] \to 0$,

the $\mathbb{E}[\mathcal{V}_s(\boldsymbol{\lambda})]$ always has a minimum $\widetilde{\boldsymbol{\lambda}}$ such that the "inefficiency" of CORE-GCV defined by

$$I^* = \frac{\mathbb{E}[\mathcal{L}_s(\widetilde{\boldsymbol{\lambda}})]}{\min_{\boldsymbol{\lambda}}\mathbb{E}[\mathcal{L}_s(\boldsymbol{\lambda})]}$$

tends to one.

Theorem 1 indicates that the loss when $\boldsymbol{\lambda}$ is estimated by minimizing $\mathcal{V}_s(\boldsymbol{\lambda})$ is close to the minimum possible loss, in the sense of expectation. The assumption $q = o(s)$ implies $r_s \to \infty$ as $s \to \infty$. Conditions (i) and (ii) are commonly used in GCV literature (Golub, Heath, and Wahba 1979; Gu and Ma 2005; Gu 2013; Xu, Shang, and Cheng 2019), and can be respectively satisfied when $\|\mathbf{H}_{\boldsymbol{\lambda}}^*\|_2 = o(r_s/q)$ and $\kappa(\mathbf{H}_{\boldsymbol{\lambda}}^*) = o(\sqrt{r_s/q})$. Such conditions are mild, only requiring the spectral norm (resp. condition number) of $\mathbf{H}_{\boldsymbol{\lambda}}^*$ to grow more slowly than r_s/q (resp. $\sqrt{r_s/q}$).

Remark 1. Since the loss function $\mathcal{L}_s(\cdot)$ is defined on a subset of observations, the corresponding optimality in Theorem 1 is indeed "local optimality." The "global optimality" w.r.t. $\mathcal{L}(\boldsymbol{\lambda}) = n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|^2$ can also be similarly achieved by defining

$$\widetilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}}\mathbb{E}[\mathcal{V}_s'(\boldsymbol{\lambda})], \quad \text{where } \mathcal{V}_s'(\boldsymbol{\lambda}) = \frac{n\|\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|^2}{[n - \text{Tr}(\mathbf{H}_{\boldsymbol{\lambda}}^*)]^2}.$$

However, the time and space complexities of computing $\mathcal{V}_s'(\boldsymbol{\lambda})$ grow linearly with n, not as scalable as $\mathcal{V}_s(\boldsymbol{\lambda})$. In practice, $\mathcal{V}_s(\boldsymbol{\lambda})$ and $\mathcal{V}_s'(\boldsymbol{\lambda})$ usually have similar empirical performance, while the former requires significantly less time. Hence, we choose $\mathcal{V}_s(\boldsymbol{\lambda})$ as the CORE-GCV score.

4.2. Approximation Guarantee

Theorem 2 provides a non-asymptotic relative error bound for the proposed core-elements estimation $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$.

Theorem 2. Recall that $\mathbf{X}^*$ is the sparse sketch of $\mathbf{X}$ and $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ is defined by (8). When $\mathbf{X}^*$ satisfies $\|\mathbf{X} - \mathbf{X}^*\|_2 \le \epsilon'\|\mathbf{X}\|_2$ with

$$0 < \epsilon' \le \frac{1}{c}\left[1 + \frac{c+1}{(\sqrt{1+\epsilon} - 1)\,\text{RSSE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})}\right]^{-1}, \tag{12}$$

we have

$$\|\boldsymbol{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|^2 \le (1+\epsilon)\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|^2.$$

In (12), $c = \|\mathbf{X}\|_2^2\|(\mathbf{X}^\top\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\|_2$ and $\text{RSSE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) = \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|/\|\boldsymbol{y}\|$ is the relative sum of squares error (RSSE) of the full sample estimation $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$.

Theorem 2 indicates that to achieve the $(1+\epsilon)$-approximation, Algorithm 1 requires a sketch $\mathbf{X}^*$ such that the ratio $\|\mathbf{X} - \mathbf{X}^*\|_2/\|\mathbf{X}\|_2$ is $O(\epsilon^{1/2})$. Intuitively, such a result also indicates that when the predictor matrix $\mathbf{X}$ becomes (numerically) sparser, fewer elements are required to select to achieve the same relative error. The upper bound in (12) also depends on the value of c and $\text{RSSE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$, in which $c \le \kappa^2(\mathbf{X})$ because $\mathbf{S}_{\boldsymbol{\lambda}}$ is positive semidefinite. In particular, to achieve the $(1+\epsilon)$-relative error, a larger ϵ' is admitted when the condition number of $\mathbf{X}$ or the signal-to-noise ratio of additive models is smaller. The following remark discusses the relationship between ϵ' and the number of selected elements s in a specific case.

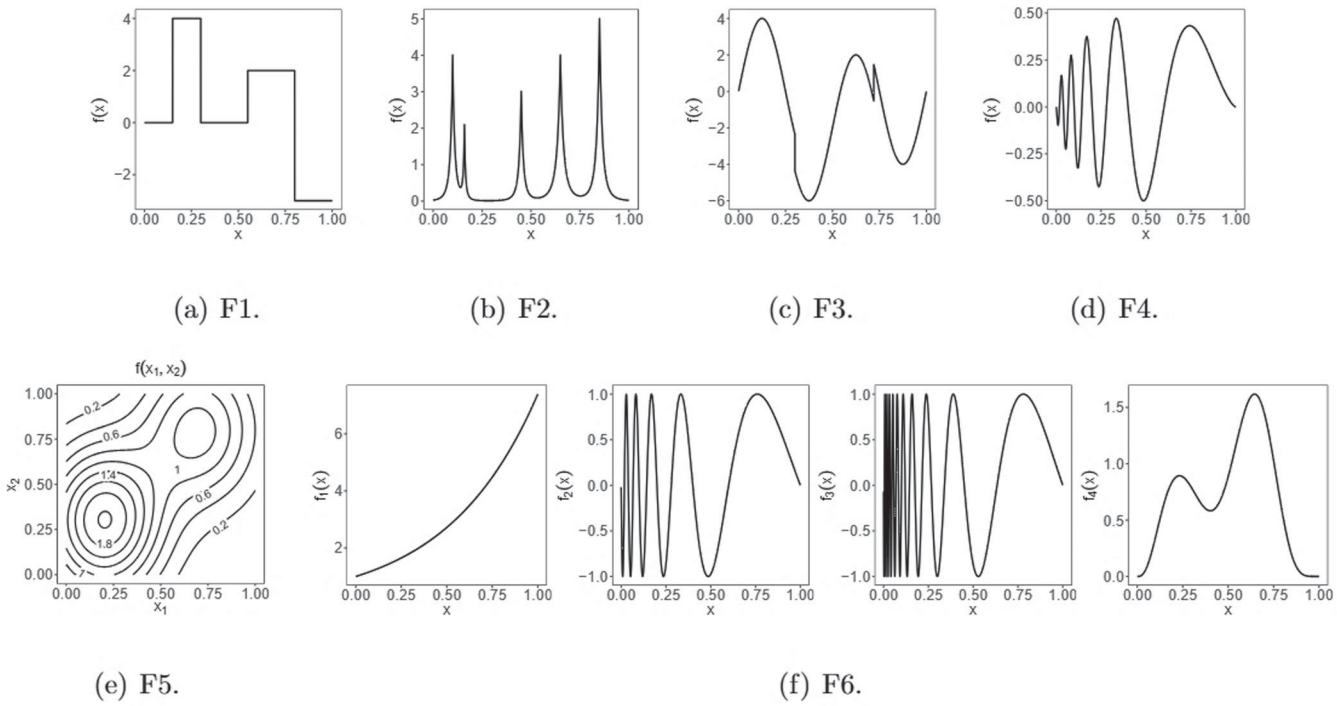


Figure 2. Illustration of the true functions.

Remark 2. Suppose the observations are n grid points on $[0, 1]$ for each of the p covariates, and each covariate is represented using q evenly spaced B-spline basis functions with the order being k ($0 \le k < q$); the examples of $k = 1$ and $k = 3$ are depicted in Figure 1(b) and (c), respectively. Then, the design matrix $\mathbf{X}$ is sparse with $(k + 1)/q \times 100\%$ nonzero elements equally distributed on each column. If the subsample parameter $r = \lfloor s/d \rfloor$ in Algorithm 1 satisfies $r < (k+1)n/q$ and

$$\frac{r}{n} \ge \frac{k+1}{q} - \frac{\epsilon'^2 \|\mathbf{X}\|_2^2}{npq}, \tag{13}$$

then the sketch matrix $\mathbf{X}^*$ achieves the condition $\|\mathbf{X} - \mathbf{X}^*\|_2 \le \epsilon' \|\mathbf{X}\|_2$ in Theorem 2. Moreover, under the conditions of Remark 2, it holds that $\|\mathbf{X}\|_2 = O(\sqrt{(k+1)np})$ when q is fixed. Then, the inequality (13) indicates that Algorithm 1 needs to select $(1 - c\epsilon'^2)(k + 1)/q \times 100\%$ elements to achieve the $(1 + \epsilon)$-relative error for a constant $c > 0$. Such a result tells us a smaller proportion of elements need to be selected if the number of basis functions q is larger. This is reasonable because a larger q corresponds to a sparser design matrix $\mathbf{X}$ for locally spaced basis functions; see Section 5 for details.

5. Simulation Studies

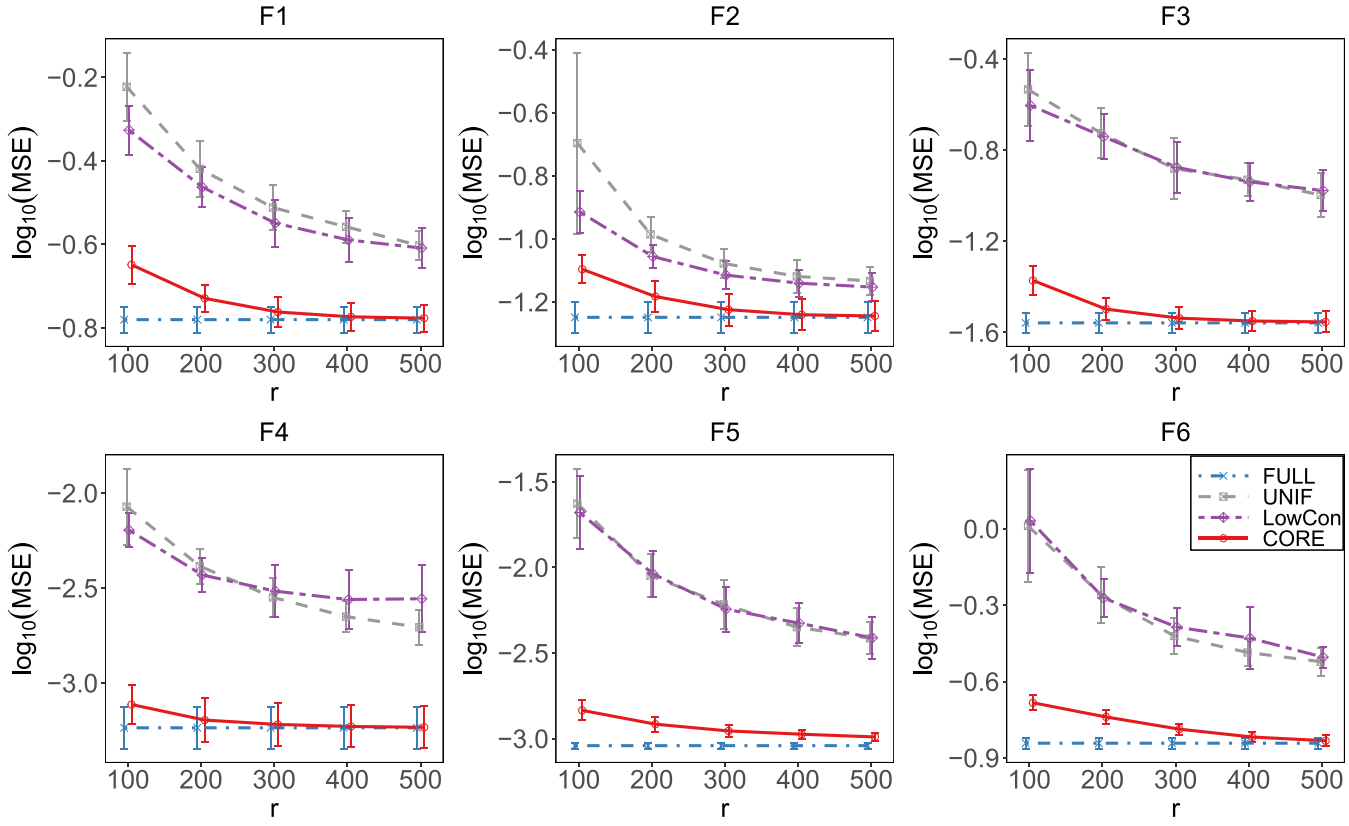
In this section, we evaluate the performance of the CORE-NAM method using synthetic data. We compare Algorithm 1 (CORE) with the full sample approach (FULL) and two row-sampling methods, including the uniform subsample (UNIF) and the LowCon method (LowCon) (Meng et al. 2021), which selects a subsample approximating a space-filling design via nearest neighbor search. To make a fair comparison, we select r rows for the row-sampling methods and $s = rd$ elements for the proposed CORE-NAM approach, such that the number of selected elements from $\mathbf{X}$ is equal. All experiments are implemented on a server with 256GB RAM and 64 cores Intel® Xeon® Gold 5218 CPU.

5.1. Accuracy and Efficiency

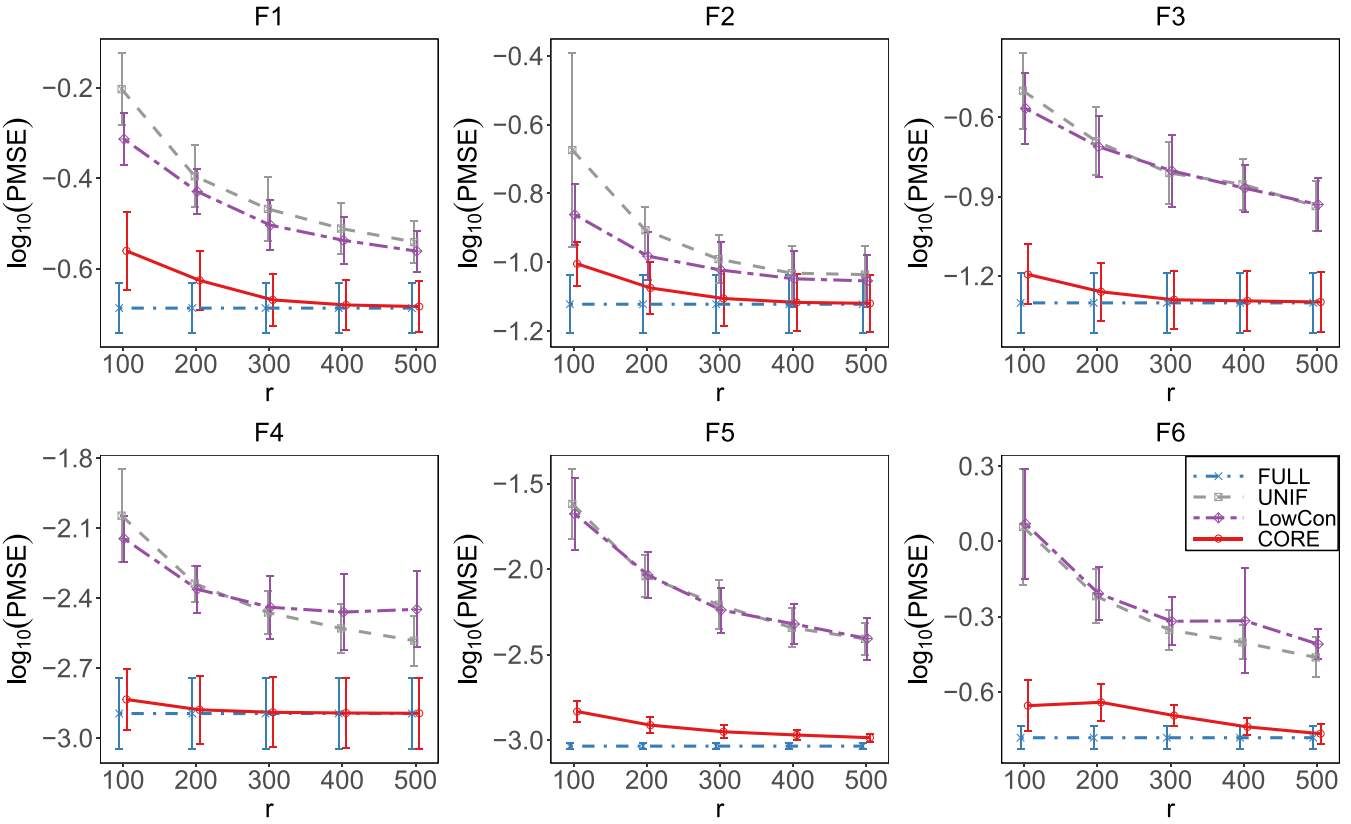
We generate iid observations with sample size n and dimension p uniformly on $[0, 1]^p$ according to (1). The settings of the function $f(\cdot)$ include four univariate functions (**F1–F4**) and two multivariate functions (**F5–F6**), which are listed as follows and illustrated in Figure 2.

- **F1.** $f(x) = \sum_j h_j K(x - t_j)$ with $K(x) = [1 + \text{sgn}(x)]/2$, where $\text{sgn}(\cdot)$ is the sign function, $(h_j) = (4, -4, 2, -5)$, and $(t_j) = (0.15, 0.3, 0.55, 0.8)$.
- **F2.** $f(x) = \sum_j h_j K[(x - t_j)/w_j]$ with $K(x) = 1/(1+|x|^4)$, where $(h_j) = (4, 2, 3, 4, 5)$, $(t_j) = (0.1, 0.16, 0.45, 0.65, 0.85)$, and $(w_j) = (0.04, 0.02, 0.04, 0.06, 0.05)$.
- **F3.** $f(x) = 4\sin(4\pi x) - \text{sgn}(x - 0.3) - \text{sgn}(0.72 - x)$.
- **F4.** $f(x) = \sqrt{x(1-x)}\sin[2\pi(1 + \delta)/(x + \delta)]$, where $\delta = 0.2$.
- **F5.** $f(x_1, x_2) = 0.75\exp[-(x_1 - 0.2)^2/\sigma_1^2 - (x_2 - 0.3)^2/\sigma_2^2]/(\pi\sigma_1\sigma_2) + 0.45\exp[-(x_1 - 0.7)^2/\sigma_1^2 - (x_2 - 0.8)^2/\sigma_2^2]/(\pi\sigma_1\sigma_2)$, where $\sigma_1 = 0.3$ and $\sigma_2 = 0.4$.
- **F6.** $f(x_1, x_2, x_3, x_4) = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4)$, where $f_1(x) = \exp(2x)$, $f_2(x) = \sin[2\pi(1 + \delta_1)/(x + \delta_1)]$ with $\delta_1 = 0.1$, $f_3(x) = \sin[2\pi(1 + \delta_2)/(x + \delta_2)]$ with $\delta_2 = 0.1$, and $f_4(x) = 10^5 x^{11}(1 - x)^6 + 10^3 x^3(1 - x)^{10}$.

Note that **F1–F4** are classical smoothness-inhomogeneous examples proposed by Donoho and Johnstone (1994, 1995), named *Blocks*, *Bumps*, *HeaviSine*, and *Doppler*, respectively. The functions in **F5–F6** are also commonly used in nonparametric regression literature (Wood 2006b; Meng et al. 2020; Sun, Zhong, and Ma 2021). Additional analyses on higher-dimensional

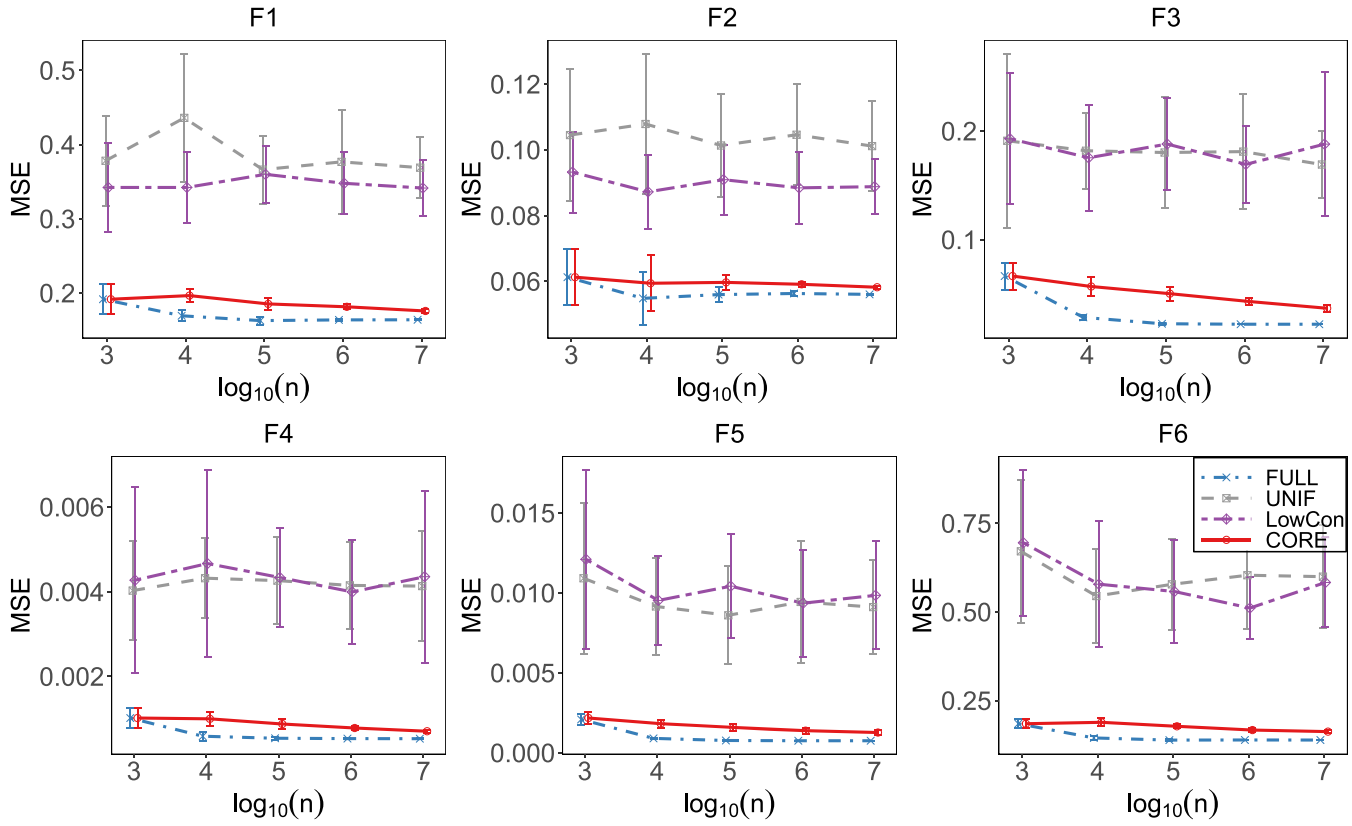


(a) Fitting errors for **F1–F6**.

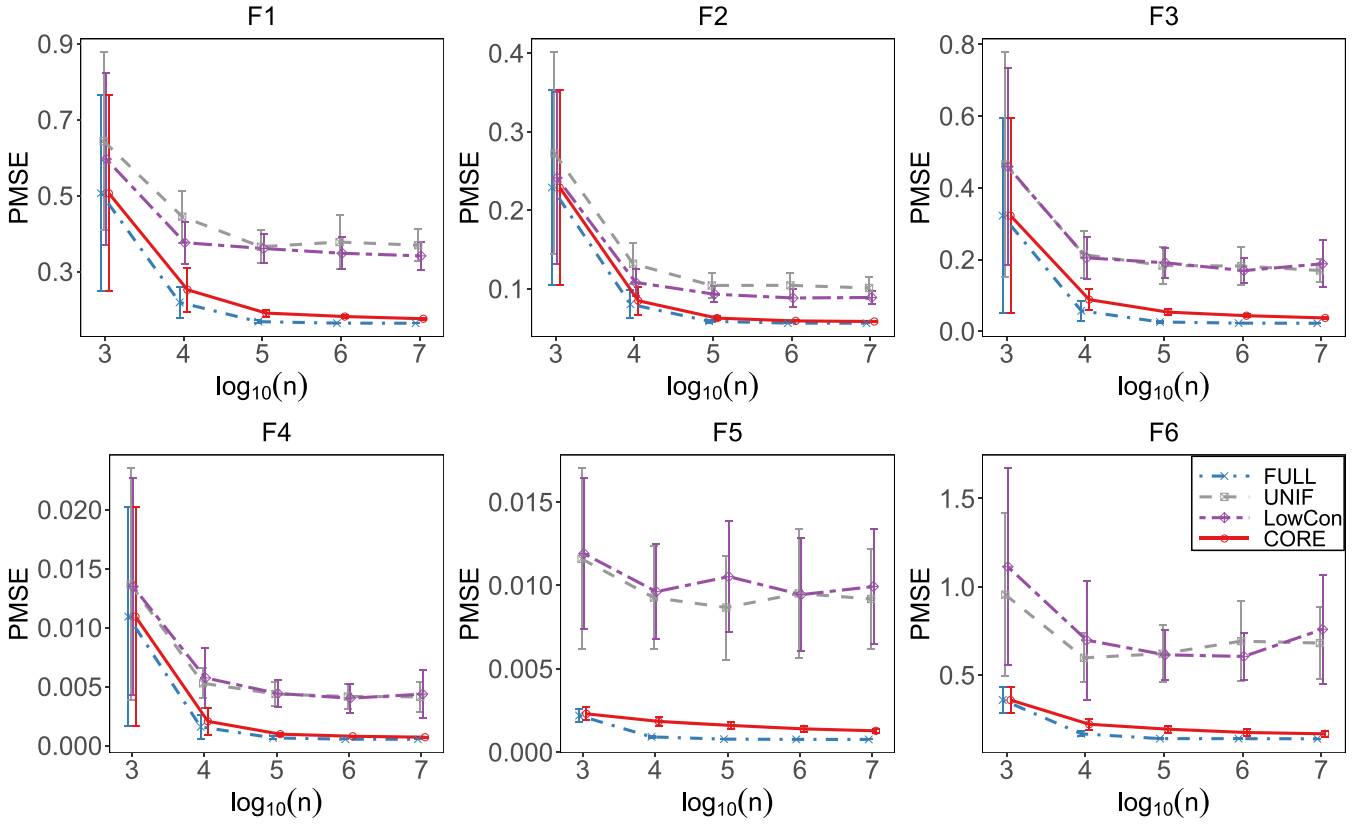


(b) Prediction errors for **F1–F6**.

Figure 3. Comparison of different approaches w.r.t. MSE and PMSE under fixed n and increasing r. The vertical bar represents the standard deviation obtained from multiple replicates.



(a) Fitting errors for **F1–F6**.



(b) Prediction errors for **F1–F6**.

Figure 4. Comparison of different approaches w.r.t. MSE and PMSE under increasing n and fixed r. The vertical bar represents the standard deviation obtained from multiple replicates.

Table 1. Comparison of different approaches w.r.t. computational time (seconds) for $n \in \{10^3, 10^5, 10^7\}$ and $r = 200$.

	(a) **F1**.				(b) **F5**.				(c) **F6**.			
Methods	FULL	UNIF	LowCon	CORE	FULL	UNIF	LowCon	CORE	FULL	UNIF	*LowCon*	CORE
$n = 10^3$	0.07	0.07	0.08	0.11	2.70	0.08	0.10	0.17	3.70	1.16	1.23	2.83
$n = 10^5$	0.64	0.10	0.11	0.15	3.14	0.11	0.13	0.35	6.13	1.31	1.49	3.43
$n = 10^7$	75.98	0.22	6.54	0.24	69.98	0.11	9.50	0.50	267.54	2.07	13.98	4.87

NOTE: The average time obtained from multiple replicates is reported.
*FULL is implemented using the `bam` function in the R package `mgcv` (Wood 2017).

Table 2. Time composition (seconds) of different approaches for **F6** under $n \in \{10^3, 10^5, 10^7\}$ and $r = 200$.

	(a) Subdata selection.				(b) Smoothing parameter selection.				(c) Model estimation.			
Methods	FULL	UNIF	*LowCon*	CORE	*FULL*	UNIF	LowCon	CORE	*FULL*	UNIF	LowCon	CORE
$n = 10^3$	–	0.00	0.00	0.00	3.68	1.15	1.22	2.82	0.02	0.01	0.01	0.01
$n = 10^5$	–	0.01	0.08	0.03	5.79	1.29	1.40	3.38	0.34	0.01	0.01	0.02
$n = 10^7$	–	0.02	12.01	1.04	234.65	2.04	1.96	3.81	32.89	0.01	0.01	0.02

NOTE: The average time obtained from multiple replicates is reported.

functions are presented in the supplementary materials. The signal-to-noise ratio, defined as SNR $= \text{var}[f(\boldsymbol{x})]/\sigma^2$, is set to be 5.

We use the cubic regression spline with the number of basis functions being $q = 40$ for **F1–F4** and $q = 30$ for **F6**, and we use the tensor product spline (Wood 2006a, 2017) with $q = 6 \times 6$ to fit the interactions in **F5**. The smoothing parameter $\boldsymbol{\lambda}$ is selected by minimizing the CORE-GCV score for the proposed approach. For FULL and the row-sampling methods (i.e., UNIF and LowCon), $\boldsymbol{\lambda}$ is estimated by minimizing the GCV score respectively over the full sample and subsample following Chen and Zhang (2022).

We evaluate their performance using two measures. One is the fitting error measured by mean squared error MSE $= \sum_{i=1}^{n}[\hat{f}(\boldsymbol{x}_i) - f(\boldsymbol{x}_i)]^2/n$, where $\hat{f}$ is the estimator obtained by different approaches. The other is the prediction error measured by prediction mean squared error PMSE $= \sum_{\boldsymbol{x}\in\mathcal{X}_{test}}[\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})]^2/|\mathcal{X}_{test}|$, where the testing set $\mathcal{X}_{test}$ contains about 10^6 evenly spaced grid points on $[0, 1]^p$. In particular, the testing set size $|\mathcal{X}_{test}| = \lfloor 10^{6/p} \rfloor^p$. One hundred replicates are implemented for each experiment.

First, we fix $n = 10^4$ and see the effect of the subsample size r. Figure 3 shows the log-transformed MSE and PMSE versus $r \in \{100, 200, \ldots, 500\}$. The error when using the full sample is a constant w.r.t. r and is included for comparison. In Figure 3, we observe all subdata-based methods improve as r increases, with our proposed CORE approach consistently performing the best. More importantly, the performance of CORE can be comparable to that of using the full sample for both fitting and prediction, even when only 4% of the elements are selected. Such superiority can also be observed from the comparison of fitted curves, which are relegated to supplementary materials.

Next, we investigate the performance of different methods under increasing full sample size n. We take $n \in \{10^3, 10^4, \ldots, 10^7\}$ with $r = 200$ and plot the corresponding MSE and PMSE in Figure 4. The results again show that the CORE-NAM outperforms the row-subsampling approaches and has similar performance to full data under all circumstances. Notably, we observe that both MSE and PMSE of the CORE method decrease as the size n grows, even with a fixed subsample size of $r = 200$.

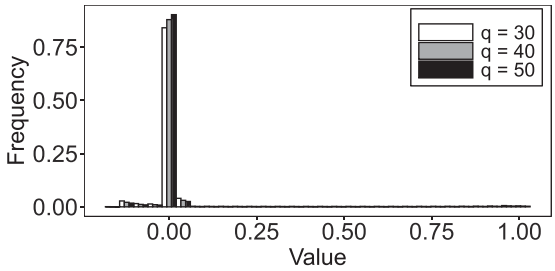


Figure 5. Histogram of the design matrix $\mathbf{X}$ for the basis dimension $q \in \{30, 40, 50\}$ with the sample size $n = 10^4$.

To see the computational efficiency of the proposed CORE-NAM approach, we report the computing time (in seconds) of various methods in Table 1. The reported time includes the steps of subdata selection, smoothing parameter selection, and model parameter estimation. The results under **F2–F4** are similar to those under **F1** and thus are omitted. From Table 1, we observe that when n is enormous, the CORE approach computes as fast as the naive UNIF method while comparing favorably to the LowCon method, also being much more efficient than the full data method. Taking the case of $n = 10^7$, for instance, CORE can speed up the full sample estimation about a hundred times. For a more comprehensive understanding of computational efficiency, we further provide the time composition for each method in Table 2. It can be seen that the computational load of the CORE method is primarily dominated by the smoothing parameter selection, and this is attributed to its iterative process in updating model parameters and computing the CORE-GCV score. Overall, combining the observations in Figure 4 and Table 1 demonstrates that the proposed CORE-NAM algorithm is suitable for dealing with large-scale data analysis.

5.2. Sensitivity Analysis

As discussed in Section 4, the proposed method is influenced by the design matrix sparseness and the noise level. We empirically analyze how these two factors impact the performance of CORE-NAM. We fix the full data size $n = 10^4$ and set the subdata size $r \in \{100, 200, \ldots, 500\}$, same to those in Figure 3. We control

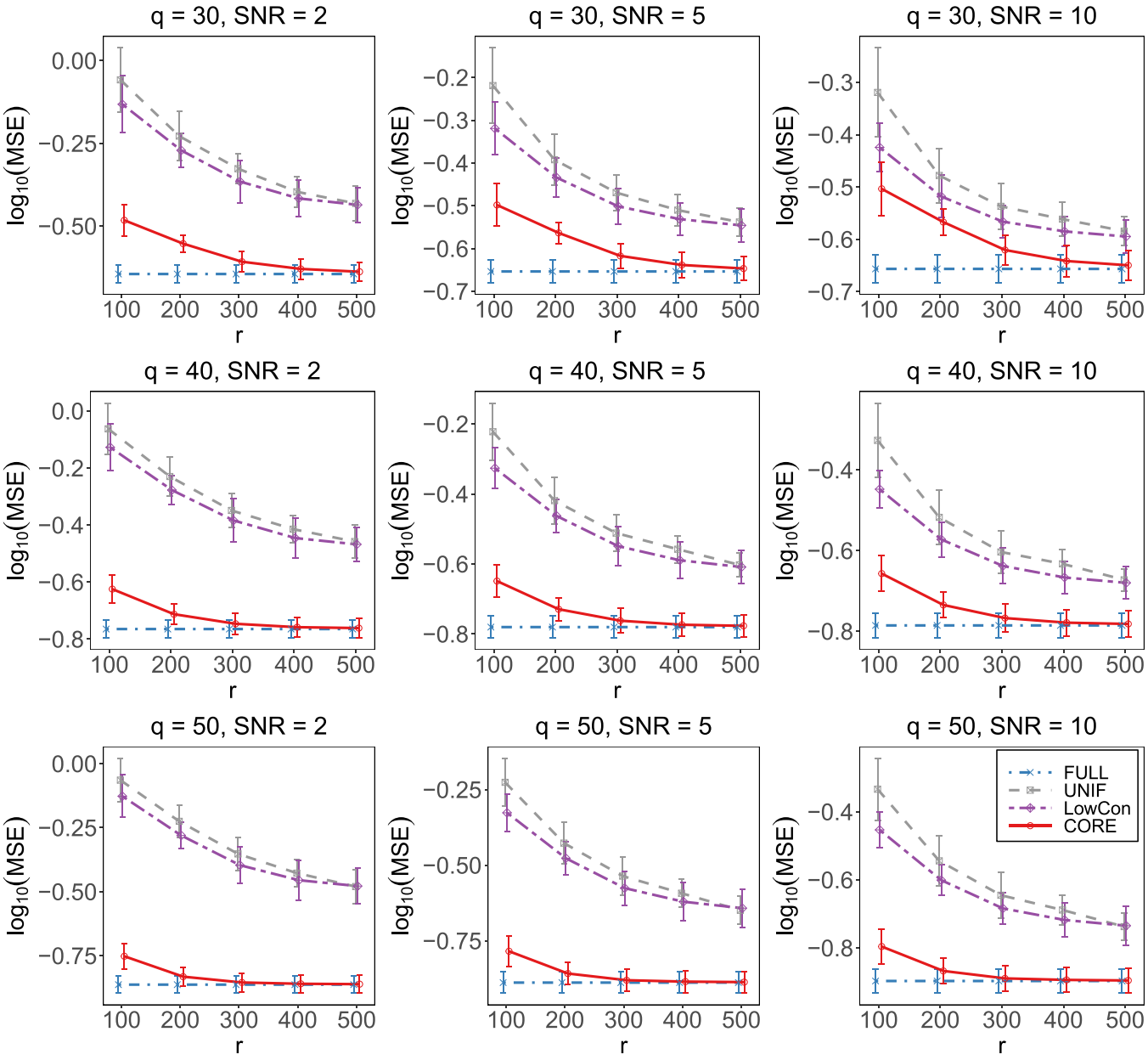


Figure 6. Impact of the sparseness and noise levels on MSE for **F1**. Each row corresponds to a particular degree of sparsity ($q \in \{30, 40, 50\}$) and each column represents a different noise level (SNR $\in \{2, 5, 10\}$). Vertical bars represent the standard deviations.

the sparseness and noise level by setting the basis dimension q and the signal-to-noise ratio, that is, $q \in \{30, 40, 50\}$ and SNR $\in \{2, 5, 10\}$, respectively.

We first illustrate the numerical sparsity of $\mathbf{X}$ and its relationship with q in Figure 5, taking the cubic regression splines as an example. Although there are only a few exactly zero elements in $\mathbf{X}$, Figure 5 shows that most elements are very close to zero and can be neglected. The numerical sparsity of $\mathbf{X}$ is mainly determined by the basis dimension q. Specifically, a larger value of q leads to a higher proportion of elements in $\mathbf{X}$ approaching zero. For instance, when $q \in \{30, 40, 50\}$, the proportions of elements in $\mathbf{X}$ smaller than 10^{-6} are 71.5%, 77.8%, and 81.8%, respectively.

Figures 6 and 7 show the log-transformed MSE and PMSE, respectively, for different combinations of q and SNR. The results indicate the advantage of Core-NAM becomes more apparent as the basis dimension q increases or the SNR decreases. Specifically, for a larger q or smaller SNR, Core-NAM only needs to select a smaller proportion of elements from $\mathbf{X}$ to perform as well as the full data. These findings are consistent with our theoretical results. We also observe that the CORE approach yields satisfactory estimation over a wide range of q and SNR, demonstrating its robustness to these hyperparameters. Such success can be attributed to the fact that the Core-NAM can effectively use the sparsity structure of the design matrix and can extract critical information from it.

5.3. *Optimality of* Core-GCV

We verify the asymptotic optimality of the proposed core-elements GCV by calculating the empirical inefficiency

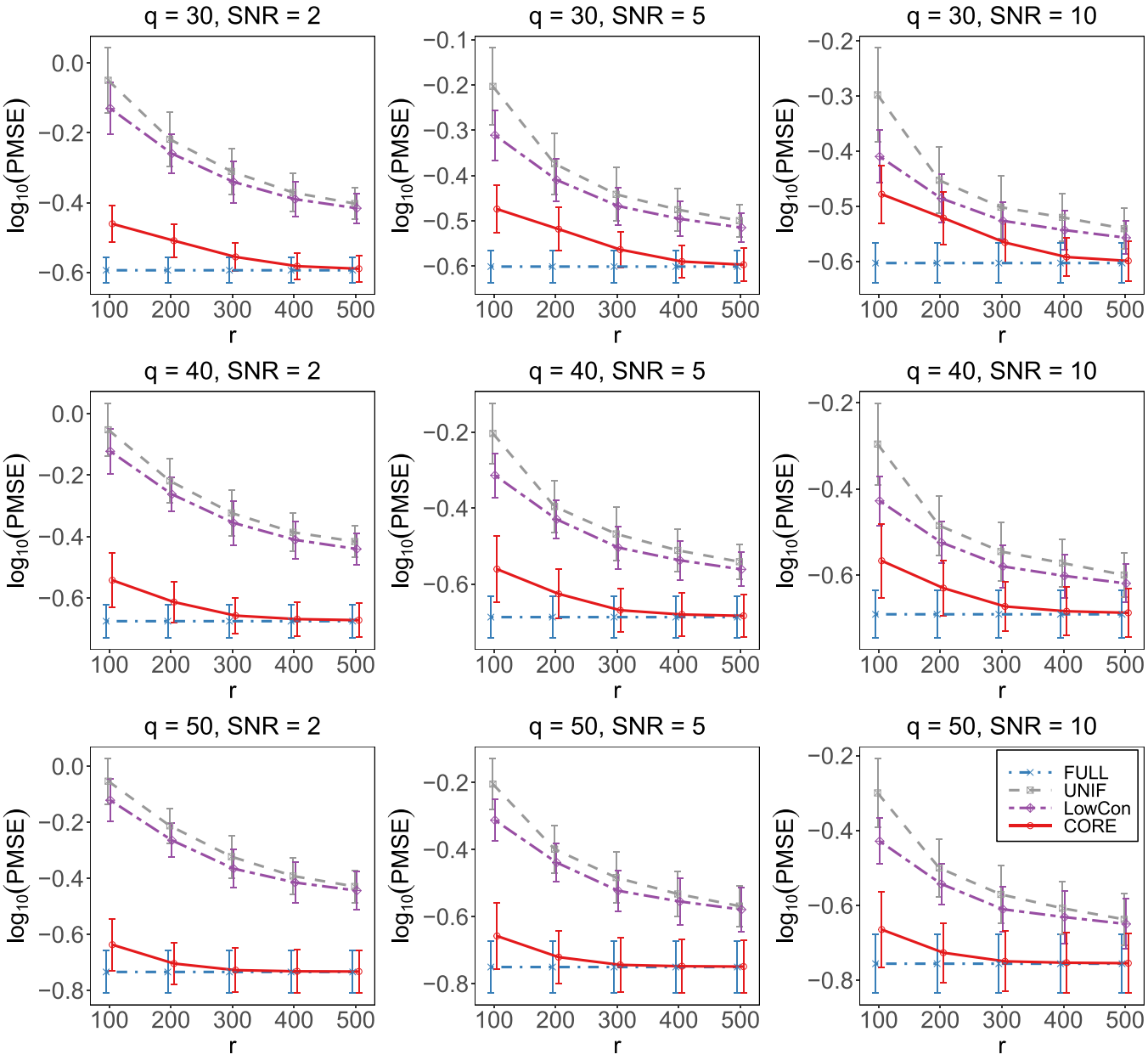


Figure 7. Impact of the sparseness and noise levels on PMSE for **F1**. Each row corresponds to a particular degree of sparsity ($q \in \{30, 40, 50\}$) and each column represents a different noise level (SNR $\in \{2, 5, 10\}$). Vertical bars represent the standard deviations.

$\mathcal{L}_s(\widetilde{\boldsymbol{\lambda}})/\min_{\boldsymbol{\lambda}} \mathcal{L}_s(\boldsymbol{\lambda})$, where $\widetilde{\boldsymbol{\lambda}}$ is selected by minimizing the Core-GCV score (10) and $\mathcal{L}_s(\cdot)$ is the loss function (11).

To satisfy the conditions in Theorem 1, we increase both the full data size n and the subdata size r, that is, $n \in \{10^3, 10^{3.5}, 10^4, 10^{4.5}\}$ and $r \in \{100, 200, 400, 800\}$. The empirical inefficiencies against increasing n for **F1–F6** are shown in Figure 8. It can be observed that the inefficiency gradually converges to the optimal value of one as n increases, which is in great agreement with Theorem 1.

6. Real Data Example

Ozone is a crucial component of the earth's atmosphere that protects the biosphere from dangerous solar ultraviolet radiation and maintains the atmospheric equilibrium. Total column ozone (TCO), a measurement of the total amount of atmospheric ozone in a given column, has been identified as one of the fifty essential climate variables by the Global Climate Observing System (Bodeker et al. 2021).

In this real data example, we aim to predict ozone levels and examine the ozone geographical distribution using the NIWA-BS Total Column Ozone Database collected by Bodeker et al. (2023). The database contains near-global daily ozone measurements at the resolution of 1.25° longitude by 1° latitude spanning from October 31, 1978 to December 31, 2019, with the data size of $n \approx 8 \times 10^8$. Following the model structures in Wood et al. (2017) and Meng et al. (2020), we fit a nonparametric additive model of the form

$$\log(\texttt{tco}_i) = f_1(\texttt{year}_i) + f_2(\texttt{doy}_i) + f_3(\texttt{lat}_i, \texttt{lon}_i) + \epsilon_i.$$

Here, the response $\log(\texttt{tco}_i)$ is the log-transformed TCO in Dobson units (DU); the predictors are year ($\texttt{year}_i$), day of year

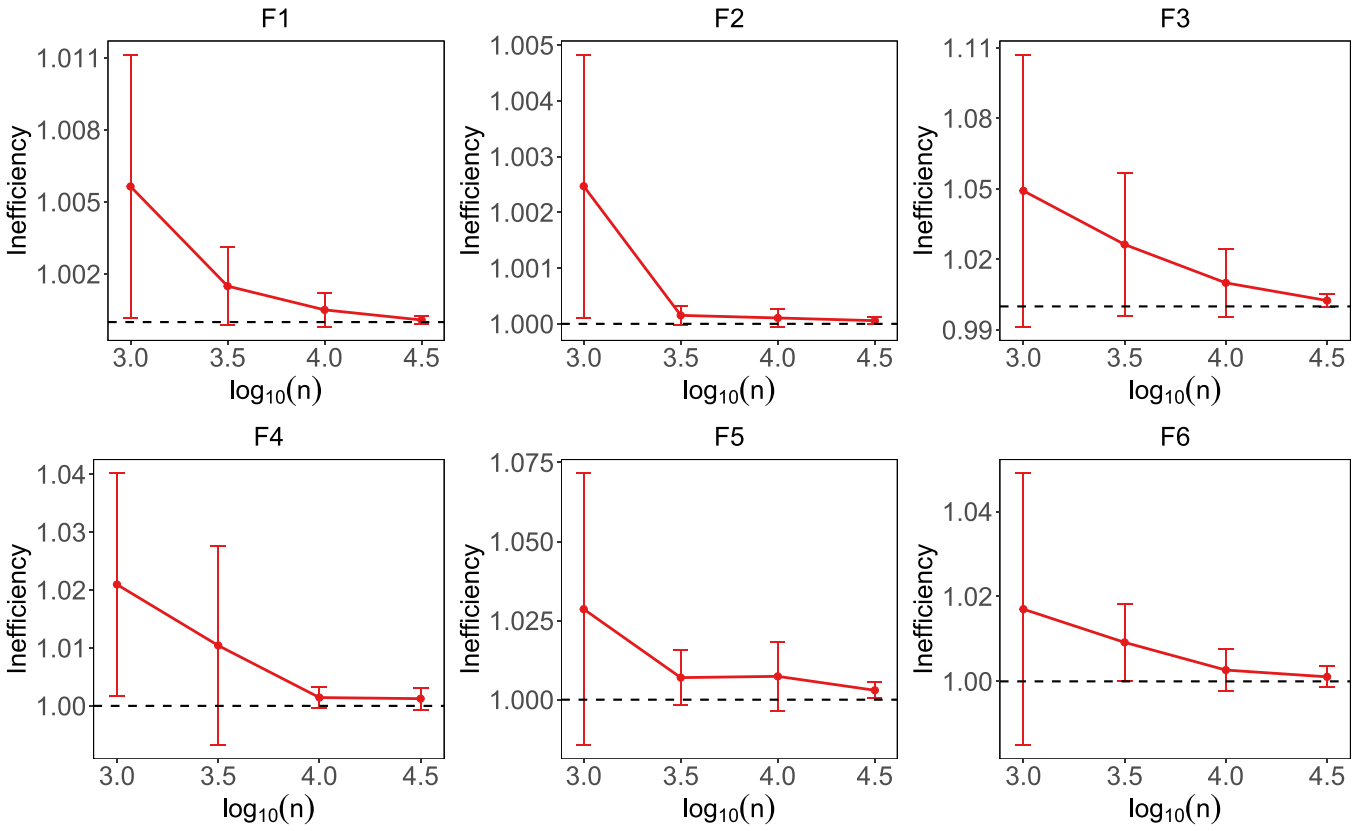


Figure 8. Empirical inefficiency versus increasing n. The horizontal dashed line represents the optimal value of one. The vertical bar represents the standard deviation obtained from multiple replicates.

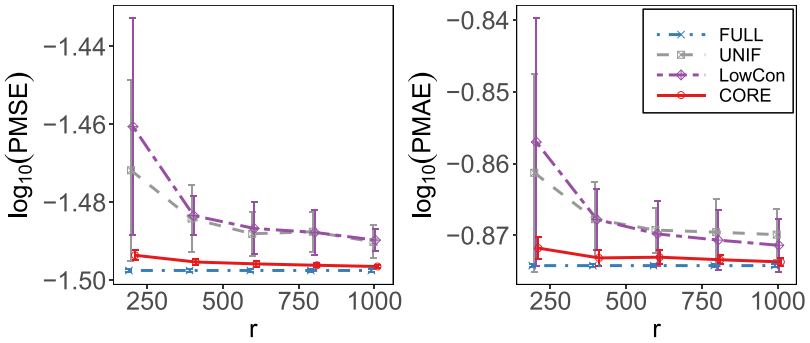


Figure 9. Comparison of different approaches w.r.t. PMSE (left) and PMAE (right) for the ozone dataset. The vertical bar represents the standard deviation obtained from multiple replicates.

(doy$_i$), latitude (lat$_i$), and longitude (lon$_i$); and ϵ_i represents the random noise. We use cubic regression splines ($q = 20$), cyclic cubic regression splines ($q = 20$), and tensor product smooths ($q = 6 \times 6$) to fit the functions f_1, f_2, f_3, respectively. Considering that the true model is unknown, we evaluate the performance of different methods via PMSE, with the training and test sets randomly partitioned according to the ratio of 100:1. We also consider a more robust evaluation metric, prediction mean absolute error PMAE $= \sum_{\boldsymbol{x}\in\mathcal{X}_{test}} |\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})|/|\mathcal{X}_{test}|$, to address potential outliers in real-world data. For a fair comparison, the subsample size is taken to be $r \in \{200, 400, \ldots, 1000\}$ rows, or $s = rd$ elements equivalently. The smoothing parameter $\boldsymbol{\lambda}$ is set in the same way as that in simulations.

Table 3. Comparison of different approaches w.r.t. computational time (seconds) for the ozone dataset.

Methods	FULL	UNIF	LowCon	CORE
$r = 500$	–	0.830	974.873	8.054
$r = 1000$	–	1.795	989.582	18.910
n	16926.709	–	–	–

NOTE: The average time obtained from multiple replicates is reported.
*FULL is implemented using the `bam` function in the R package `mgcv` (Wood 2017).

The prediction errors of subdata-based approaches against increasing subsample sizes are presented in Figure 9, with the error using full data also plotted for comparison. Agree with the results in Figure 3, the proposed CORE approach

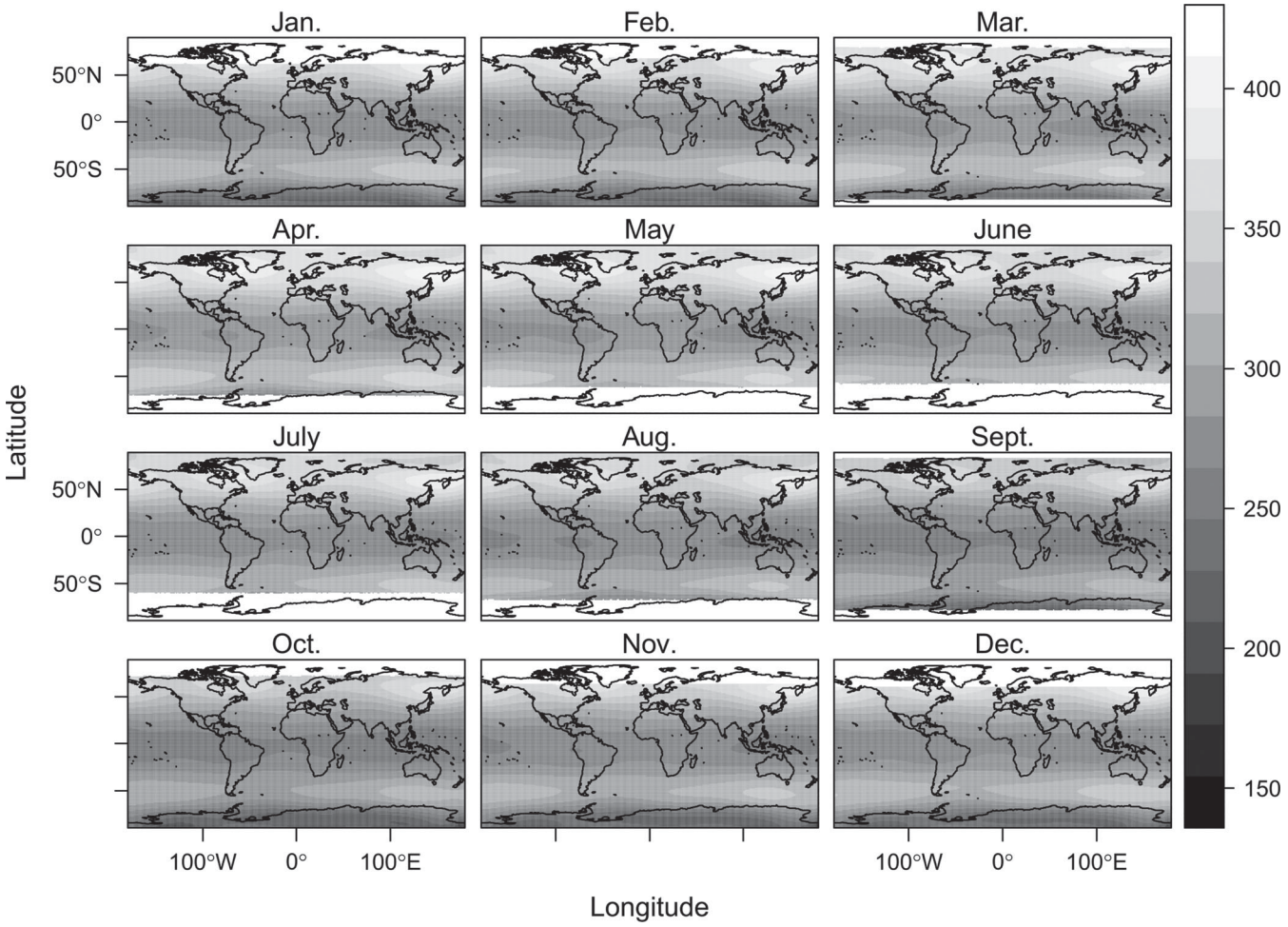


Figure 10. Monthly average predicted total column ozone value (DU) for each month in the year of 2018, obtained from the Core-NAM approach with $r = 2000$. Each subfigure represents a different month.

outperforms all other subdata-based methods and is comparable to the full sample prediction, even when the selected number of elements only accounts for less than 0.0002% of the entire design matrix. Table 3 reports the computational time for the ozone dataset, from which we can see CORE only requires 0.11% running time compared to FULL, indicating the significant efficiency of the proposed method in handling massive data.

We illustrate the prediction results of the Core-NAM method to investigate the global distribution of ozone. Figure 10 shows the monthly average predicted TCO for each month in 2018. It is seen that TCO peaks in the mid-latitude to high-latitude regions in the northern hemisphere and the mid-latitude regions in the southern hemisphere, while the minimum TCO is in the polar regions. We also observe seasonal variability; that is, at northern mid-latitudes, ozone amounts become larger in winter and early spring and smaller in summer and fall. Such observations conform to natural laws (Stolarski 2003) and can be elucidated by the Brewer–Dobson circulation theory (Butchart 2014).

Further, we present the predicted TCO across years to study long-term evolution. In Figure 11, the top panels show TCO for earlier years (1980–1982), while the bottom panels show recent counterparts (2016–2018). We observe that the contemporary ozone levels near the South Polar are less than half the past. These large Antarctic ozone losses are popularly known as the Antarctic ozone hole (Newman 2003), whose development is caused by the pollution with chemicals containing chlorine and bromine.

7. Concluding Remarks

To address the computational challenges in NAMs, we propose a novel and scalable Core-NAM method for efficient model fitting, prediction, and statistical inference. The proposed approach has theoretical guarantees and behaves superiorly to competitors in extensive numerical experiments, showing excellent capability in super large-scale data analysis.

Future works plan to extend the Core-NAM method to a broader range of scenarios, such as the generalized additive model (Wood 2017), the smoothing spline ANOVA model (Gu 2013), and distributed learning (Li and Zhao 2022). Considering that the smoothing spline ANOVA model can also be estimated via penalized least squares with smoothing parameters selected by GCV, it is promising to extend the proposed core-elements estimation and core-elements CGV method to accelerate the corresponding procedures. However, a critical challenge

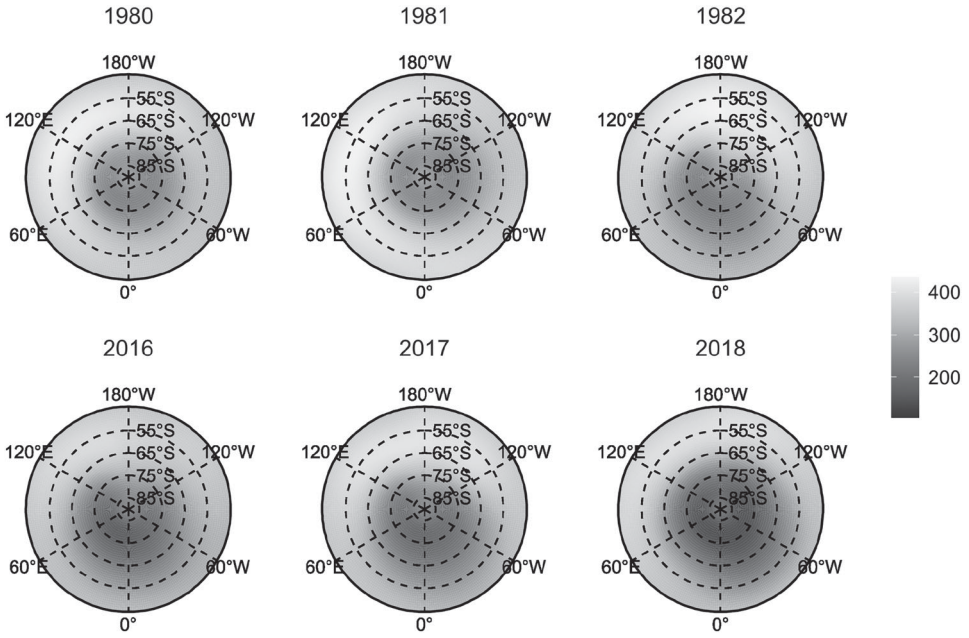


Figure 11. Monthly average predicted total column ozone value (DU, latitude 45°S to 90°S) during October in specific years, obtained from the Core-NAM approach with $r = 2000$. Each subfigure represents a different year.

in smoothing splines is the computational burden arising from the large number of basis functions, which equals the sample size. This highlights the need for exploring strategies to sparsify these basis functions efficiently. Additionally, facilitating with the developed powerful computing tool, we are interested in the applications of nonparametric regression in massive real-world data, solving critical scientific problems and advancing our understanding in various domains.

Supplementary Materials

Appendix: contains complete proofs of theoretical results and additional experiments to evaluate the performance of the proposed method. (appendix.pdf, a pdf file)

Code: contains R code that implements the proposed method and reproduces the numerical results. A readme file describing the contents is included. (code.zip, a zip file)

Acknowledgments

The authors appreciate the Editor, Associate Editor, and two anonymous reviewers for their constructive comments that helped improve the work.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work is supported by Beijing Municipal Natural Science Foundation No. 1232019, National Natural Science Foundation of China Grant No. 12101606, No. 12101606, National Key R&D Program of China No. 2021YFA1001300, Renmin University of China research fund program for young scholars, and the Outstanding Innovative Talents Cultivation Funded Programs 2021 of Renmin University of China.

ORCID

Mengyu Li http://orcid.org/0000-0002-5286-7525
Jingyi Zhang http://orcid.org/0000-0002-3147-8838
Cheng Meng http://orcid.org/0000-0002-7111-0966

References

Ai, M., Wang, F., Yu, J., and Zhang, H. (2021), "Optimal Subsampling for Large-Scale Quantile Regression," *Journal of Complexity*, 62, 101512. [2]

Ai, M., Yu, J., Zhang, H., and Wang, H. (2021), "Optimal Subsampling Algorithms for Big Data Regressions," *Statistica Sinica*, 31, 749–772. [2]

Balle, B., Barthe, G., and Gaboardi, M. (2020), "Privacy Profiles and Amplification by Subsampling," *Journal of Privacy and Confidentiality*, 10, 1–32. [2]

Bates, D., and Eddelbuettel, D. (2013), "Fast and Elegant Numerical Linear Algebra Using the RcppEigen Package," *Journal of Statistical Software*, 52, 1–24. [5]

Bodeker, G. E., Nitzbon, J., Lewis, J., Schwertheim, A., Tradowsky, J. S., and Kremser, S. (2023), "NIWA-BS Total Column Ozone Database v3.4.1," available at: *https://zenodo.org/records/7447660*. [2,11]

Bodeker, G. E., Nitzbon, J., Tradowsky, J. S., Kremser, S., Schwertheim, A., and Lewis, J. (2021), "A Global Total Column Ozone Climate Data Record," *Earth System Science Data*, 13, 3885–3906. [11]

Boutsidis, C., Drineas, P., and Magdon-Ismail, M. (2013), "Near-Optimal Coresets for Least-Squares Regression," *IEEE Transactions on Information Theory*, 59, 6880–6892. [2]

Breiman, L., and Friedman, J. H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80, 580–598. [1]

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," *The Annals of Statistics*, 17, 453–510. [1]

Butchart, N. (2014), "The Brewer–Dobson Circulation," *Reviews of Geophysics*, 52, 157–184. [13]

Carmon, Y., Jin, Y., Sidford, A., and Tian, K. (2020), "Coordinate Methods for Matrix Games," in *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*, pp. 283–293. IEEE. [2]

Chen, Y., and Zhang, N. (2022), "Optimal Subsampling for Large Sample Ridge Regression," arXiv preprint arXiv:2204.04776. [9]

Craven, P., and Wahba, G. (1978), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method

of Generalized Cross-Validation," *Numerische Mathematik*, 31, 377–403. [3]

Dai, W., Song, Y., and Wang, D. (2023), "A Subsampling Method for Regression Problems based on Minimum Energy Criterion," *Technometrics*, 65, 192–205. [2]

Dai, X., Lyu, X., and Li, L. (2023), "Kernel Knockoffs Selection for Nonparametric Additive Models," *Journal of the American Statistical Association*, 118, 2158–2170. [1]

Dasgupta, A., Drineas, P., Harb, B., Kumar, R., and Mahoney, M. W. (2009), "Sampling Algorithms and Coresets for l_p Regression," *SIAM Journal on Computing*, 38, 2060–2078. [2]

Dereziński, M., Warmuth, M. K., and Hsu, D. J. (2018), "Leveraged Volume Sampling for Linear Regression," in *Advances in Neural Information Processing Systems* (Vol. 31), pp. 2510–2519. [2]

Diao, H., Ai, M., Tian, Y., and Yu, J. (2023), "Efficient Basis Selection for Smoothing Splines via Rotated Lattices," *Stat*, 12, e581. [1]

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455. [6]

——— (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90, 1200–1224. [6]

Eilers, P. H., and Marx, B. D. (1992), "Generalized Linear Models with P-Splines," in *Advances in GLIM and Statistical Modelling: Proceedings of the GLIM 92 Conference and the 7th International Workshop on Statistical Modelling*, pp. 72–77, Springer. [3]

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

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing with B-splines and Penalties," *Statistical Science*, 11, 89–121. [3]

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557. [1]

Golub, G. H., Heath, M., and Wahba, G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–223. [3,5]

Green, P. J., and Silverman, B. W. (1993), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Boca Raton, FL: Chapman and Hall/CRC. [3]

Gu, C. (2013), *Smoothing Spline ANOVA Models* (2nd ed.), New York: Springer. [3,5,13]

Gu, C., and Ma, P. (2005), "Optimal Smoothing in Nonparametric Mixed-Effect Models," *The Annals of Statistics*, 33, 1357–1379. [5]

Gu, C., and Wahba, G. (1991), "Minimizing GCV/GML Scores with Multiple Smoothing Parameters via the Newton Method," *SIAM Journal on Scientific and Statistical Computing*, 12, 383–398. [3]

Gupta, N., and Sidford, A. (2018), "Exploiting Numerical Sparsity for Efficient Learning: Faster Eigenvector Computation and Regression," in *Advances in Neural Information Processing Systems* (Vol. 31), pp. 5274–5283. [2]

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton, FL: Chapman & Hall/CRC. [1]

Helwig, N. E., and Ma, P. (2015), "Fast and Stable Multiple Smoothing Parameter Selection in Smoothing Spline Analysis of Variance Models with Large Samples," *Journal of Computational and Graphical Statistics*, 24, 715–732. [1]

——— (2016), "Smoothing Spline ANOVA for Super-Large Samples: Scalable Computation via Rounding Parameters," *Statistics and Its Interface*, 9, 433–444. [1]

Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313. [1]

Joseph, V. R., and Mak, S. (2021), "Supervised Compression of Big Data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 14, 217–229. [2]

Kong, E., and Xia, Y. (2019), "On the Efficiency of Online Approach to Nonparametric Smoothing of Big Data," *Statistica Sinica*, 29, 185–201. [2]

Li, K.-C. (1987), "Asymptotic Optimality for C_p, C_l, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975. [3]

Li, M., Yu, J., Li, T., and Meng, C. (2023), "Core-Elements for Classical Linear Regression," arXiv preprint arXiv:2206.10240. [2,3,4]

Li, T., and Meng, C. (2021), "Modern Subsampling Methods for Large-Scale Least Squares Regression," *International Journal of Cyber-Physical Systems*, 2, 1–28. [2]

Li, M., and Zhao, J. (2022), "Communication-Efficient Distributed Linear Discriminant Analysis for Binary Classification," Statistica Sinica, 32, 1343–1361. [13]

Ma, P., Huang, J. Z., and Zhang, N. (2015), "Efficient Computation of Smoothing Splines via Adaptive Basis Sampling," *Biometrika*, 102, 631–645. [1]

Ma, P., and Sun, X. (2015), "Leveraging for Big Data Regression," *Wiley Interdisciplinary Reviews: Computational Statistics*, 7, 70–76. [1,2]

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. [1]

Ma, P., Zhang, X., Xing, X., Ma, J., and Mahoney, M. (2020), "Asymptotic Analysis of Sampling Estimators for Randomized Numerical Linear Algebra Algorithms," in *International Conference on Artificial Intelligence and Statistics*, pp. 1026–1035, PMLR. [2]

Mak, S., and Joseph, V. R. (2018), "Support Points," *The Annals of Statistics*, 46, 2562–2592. [2]

Mallows, C. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675. [3]

Marra, G., and Wood, S. N. (2011), "Practical variable Selection for Generalized Additive Models," *Computational Statistics & Data Analysis*, 55, 2372–2387. [1]

Martinez, C. (2004), "Partial Quicksort," in *Proc. 6th ACMSIAM Workshop on Algorithm Engineering and Experiments and 1st ACM-SIAM Workshop on Analytic Algorithmics and Combinatorics*, pp. 224–228. [5]

Meier, L., van de Geer, S., and Bühlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [1]

Meng, C., Wang, Y., Zhang, X., Mandal, A., Zhong, W., and Ma, P. (2017), "Effective Statistical Methods for Big Data Analytics," in *Handbook of Research on Applied Cybernetics and Systems Science*, pp. 280–299, IGI Global. [2]

Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., and Ma, P. (2021), "Lowcon: A Design-based Subsampling Approach in a Misspecified Linear Model," *Journal of Computational and Graphical Statistics*, 30, 694–708. [1,6]

Meng, C., Yu, J., Chen, Y., Zhong, W., and Ma, P. (2022), "Smoothing Splines Approximation Using Hilbert Curve Basis Selection," *Journal of Computational and Graphical Statistics*, 31, 802–812. [1]

Meng, C., Zhang, X., Zhang, J., Zhong, W., and Ma, P. (2020), "More Efficient Approximation of Smoothing Splines via Space-Filling Basis Selection," *Biometrika*, 107, 723–735. [1,6,11]

Musser, D. R. (1997), "Introspective Sorting and Selection Algorithms," *Software: Practice and Experience*, 27, 983–993. [5]

Newman, P. A. (2003), "An Introduction to Stratospheric Ozone," in *Stratospheric Ozone: An Electronic Textbook*. NASA's Goddard Space Flight Center Atmospheric Chemistry and Dynamics Branch (Code 916), available at: *http://www.ccpo.odu.edu/SEES/ozone/oz_class.htm*. [13]

Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. (2021), "Uniform Consistency of Cross-Validation Estimators for High-Dimensional Ridge Regression," in *International Conference on Artificial Intelligence and Statistics*, pp. 3178–3186, PMLR. [3]

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., and Schmid, M. (2019), "A Review of Spline Function Procedures in R," *BMC Medical Research Methodology*, 19, 1–16. [1,3]

Reinsch, C. H. (1967), "Smoothing by Spline Functions," *Numerische Mathematik*, 10, 177–183. [3]

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012), "Spike-and-Slab Priors for Function Selection in Structured Additive Regression Models," *Journal of the American Statistical Association*, 107, 1518–1532. [1]

Schifano, E. D., Wu, J., Wang, C., Yan, J., and Chen, M.-H. (2016), "Online Updating of Statistical Inference in the Big Data Setting," *Technometrics*, 58, 393–403. [2]

Stolarski, R. (2003), "Stratospheric Ozone Variability," in *Stratospheric Ozone: An Electronic Textbook*. NASA's Goddard Space Flight Center

Atmospheric Chemistry and Dynamics Branch (Code 916), available at: *http://www.ccpo.odu.edu/SEES/ozone/oz_class.htm*. [13]

Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, 13, 689–705. [1]

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Series B, 36, 111–133. [3]

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. [1,6]

Wahba, G. (1983), "Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society*, Series B, 45, 133–150. [5]

Wahba, G. (1985), "A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem," *The Annals of Statistics*, 13, 1378–1402. [3]

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

Wahba, G., and Wold, S. (1975), "A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation," *Communications in Statistics-Theory and Methods*, 4, 1–17. [3]

Wang, H., and Ma, Y. (2021), "Optimal Subsampling for Quantile Regression in Big Data," *Biometrika*, 108, 99–112. [2]

Wang, H., Yang, M., and Stufken, J. (2019), "Information-Based Optimal Subdata Selection for Big Data Linear Regression," *Journal of the American Statistical Association*, 114, 393–405. [1,2,5]

Wang, H., Zhu, R., and Ma, P. (2018), "Optimal Subsampling for Large Sample Logistic Regression," *Journal of the American Statistical Association*, 113, 829–844. [2]

Wang, L., Elmstedt, J., Wong, W. K., and Xu, H. (2021), "Orthogonal Subsampling for Big Data Linear Regression," *The Annals of Applied Statistics*, 15, 1273–1290. [2]

Wang, Y., Yu, A. W., and Singh, A. (2017), "On Computationally Tractable Selection of Experiments in Measurement-Constrained Regression Models," *The Journal of Machine Learning Research*, 18, 5238–5278. [1]

Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P. (2019), "Subsampled Rényi Differential Privacy and Analytical Moments Accountant," in *International Conference on Artificial Intelligence and Statistics* (Vol. 89), pp. 1226–1235, PMLR. [2]

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

——— (2004), "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models," *Journal of the American Statistical Association*, 99, 673–686. [1]

——— (2006a), "Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models," *Biometrics*, 62, 1025–1036. [9]

——— (2006b), "On Confidence Intervals for Generalized Additive Models based on Penalized Regression Splines," *Australian & New Zealand Journal of Statistics*, 48, 445–464. [5,6]

——— (2011), "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models," *Journal of the Royal Statistical Society*, Series B, 73, 3–36. [3]

——— (2017), *Generalized Additive Models: An Introduction with R* (2nd ed.), London: Chapman and Hall/CRC. [1,2,3,9,12,13]

——— (2020), "Inference and Computation with Generalized Additive Models and their Extensions," *Test*, 29, 307–339. [1]

Wood, S. N., and Augustin, N. H. (2002), "GAMs with Integrated Model Selection Using Penalized Regression Splines and Applications to Environmental Modelling," *Ecological Modelling*, 157, 157–177. [1,3]

Wood, S. N., Goude, Y., and Shaw, S. (2015), "Generalized Additive Models for Large Data Sets," *Journal of the Royal Statistical Society*, Series C, 64, 139–155. [1]

Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. (2017), "Generalized Additive Models for Gigadata: Modeling the U.K. Black Smoke Network Daily Data," *Journal of the American Statistical Association*, 112, 1199–1210. [1,11]

Wood, S. N., Pya, N., and Säfken, B. (2016), "Smoothing Parameter and Model Selection for General Smooth Models," *Journal of the American Statistical Association*, 111, 1548–1563. [1]

Wu, S., Cheng, H., Cai, J., Ma, P., and Zhong, W. (2023), "Subsampling in Large Graphs Using Ricci Curvature," in *The Eleventh International Conference on Learning Representations*. [1]

Xie, R., Bai, S., and Ma, P. (2023), "Optimal Sampling Designs for Multi-Dimensional Streaming Time Series with Application to Power Grid Sensor Data," *The Annals of Applied Statistics*, 17, 3195–3215. [2]

Xie, R., Wang, Z., Bai, S., Ma, P.,, and Zhong, W. (2019), "Online Decentralized Leverage Score Sampling for Streaming Multidimensional Time Series," in *International Conference on Artificial Intelligence and Statistics*, pp. 2301–2311. [2]

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. [5]

Xue, D., and Yao, F. (2022), "Dynamic Penalized Splines for Streaming Data," *Statistica Sinica*, 32, 1363–1380. [2]

Yang, Y., and Yao, F. (2023), "Online Estimation for Functional Data," *Journal of the American Statistical Association*, 118, 1630–1644. [2]

Yang, Y., Yao, F., and Zhao, P. (2023), "Online Smooth Backfitting for Generalized Additive Models," *Journal of the American Statistical Association*, 1–29. [1,2]

Yu, J., Ai, M., and Ye, Z. (2023), "A Review on Design Inspired Subsampling for Big Data," *Statistical Papers*, 1–44. [2]

Yu, J., Wang, H., Ai, M., and Zhang, H. (2022), "Optimal Distributed Subsampling for Maximum Quasi-Likelihood Estimators with Massive Data," *Journal of the American Statistical Association*, 117, 265–276. [2]

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

Zhang, J., Meng, C., Yu, J., Zhang, M., Zhong, W., and Ma, P. (2023), "An Optimal Transport Approach for Selecting a Representative Subsample with Application in Efficient Kernel Density Estimation," *Journal of Computational and Graphical Statistics*, 32, 329–339. [1]

Zhang, Y., Wang, L., Zhang, X., and Wang, H. (2023),"'Independence-Encouraging Subsampling for Nonparametric Additive Models," arXiv preprint arXiv:2302.13441. [2]