Approximate Kernel Selection via Matrix Approximation

Lizhong Ding, Shizhong Liao, Yong Liu, Li Liu, Fan Zhu, Yazhou Yao, Ling Shao, and Xin Gao

***Abstract*— Kernel selection is of fundamental importance for the generalization of kernel methods. This article proposes an approximate approach for kernel selection by exploiting the approximability of kernel selection and the computational virtue of kernel matrix approximation. We define approximate consistency to measure the approximability of the kernel selection problem. Based on the analysis of approximate consistency, we solve the theoretical problem of whether, under what conditions, and at what speed, the approximate criterion is close to the accurate one, establishing the foundations of approximate kernel selection. We introduce two selection criteria based on error estimation and prove the approximate consistency of the multilevel circulant matrix (MCM) approximation and Nyström approximation under these criteria. Under the theoretical guarantees of the approximate consistency, we design approximate algorithms for kernel selection, which exploits the computational advantages of the MCM and Nyström approximations to conduct kernel selection in a linear or quasi-linear complexity. We experimentally validate the theoretical results for the approximate consistency and evaluate the effectiveness of the proposed kernel selection algorithms.**

***Index Terms*— Approximate algorithms, approximate consistency, kernel matrix approximation, kernel selection.**

I. Introduction

FOR A FINITE sample of data, learning involves finding a function that yields good predictions on unknown data [1]. Learning is an ill-posed problem, and it is impossible to obtain a unique solution based only on data itself [2]. Additional assumptions are necessary to make learning possible. We call this set of assumptions the inductive bias [3]. The task of model selection involves determining the inductive bias, which is critical to the performance of learning algorithms. Kernel methods play an important role in the machine learning community and have recently been applied in multitask learning [4], k-means clustering [5], logistic regression [6], hypothesis testing [7], [8], and so on. For supervised kernel-based learning algorithms [9]–[12], the aim of model selection is to choose the kernel function and the regularization parameter. In this article, we focus on kernel selection, which has an essential influence on kernel-based learning [13]–[15].

Kernel selection is closely related to the generalization error of learning algorithms. The kernel with the smallest generalization error is usually regarded as the optimal kernel [16], [17]. However, we cannot directly compute the generalization error because the underlying distribution of the given data is often unknown. Using the theoretical upper bounds of the generalization error is a common strategy in kernel selection. The upper bounds are composed of the empirical error and the hypothesis space complexity [16]. Different measurements of the complexity constitute different kernel selection methods. These include the Rademacher complexity [17], the local Rademacher complexity [18], [19], maximal discrepancy [16], [17], covering number [20], and so on. Although kernel selection is closely linked to the generalization error, kernel selection criteria are not required to be unbiased estimates of the generalization error [21], [22]. The main requirement for kernel selection criteria is that they give an indication of the generalization error. Therefore, it is sufficient to compute approximate kernel selection criteria, which discriminates the good kernels from other candidates. We refer to this property as the approximability of kernel selection. On the other hand, the computational complexities of the existing kernel selection criteria are at least $O(l^2)$, where l is the number of samples. This kind of scalability is prohibitive for big data. Such considerations drive the study of this article.

By exploiting the approximability of the kernel selection problem and the computational virtue of kernel matrix approximation, we propose an approximate approach for kernel selection. We define approximate consistency, which measures the approximability of kernel selection. Then, with the notion of approximate consistency, we answer the theoretical question of approximate kernel selection: under what conditions and at what speed, the approximate criterion is close to the accurate one, if at all. We introduce two criteria defined by error estimation, as two cases for studying the approximate consistency of the multilevel circulant matrix (MCM) approximation and Nyström approximation. The results demonstrate

Manuscript received May 6, 2017; revised April 2, 2018, January 18, 2019, and October 12, 2019; accepted November 19, 2019. This work was supported in part by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award URF/1/4098-01-01 and Award BAS/1/1624-01-01, in part by the National Natural Science Foundation of China under Grant 61703396, in part by the National Natural Science Foundation of China under Grant 61673293, and in part by the CCF-Tencent Open Fund and Shenzhen Government under Grant GJHZ20180419190732022. *(Corresponding author: Xin Gao.)*

L. Ding, L. Liu, F. Zhu, and L. Shao are with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates.

S. Liao is with the School of Computer Science and Technology, Tianjin University, Tianjin 300350, China.

Y. Liu is with the Institute of Information Engineering, Chinese Academy of Sciences (CAS), Beijing 100093, China.

Y. Yao is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.

X. Gao is with the Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia (e-mail: xin.gao@kaust.edu.sa).

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2958922

the rationality of introducing matrix approximation into kernel selection. Based on theoretical findings on the approximate consistency, we design approximate kernel selection algorithms to conduct kernel selection, which alleviates the computational bottleneck issue faced by the accurate kernel selection procedures. The designed algorithms exploit the computational virtues of the MCM approximation and Nyström approximation to conduct kernel selection in $O(l\log(l))$ or $O(l)$ time complexity. We conduct experiments on benchmark and synthetic data to verify the theoretical findings for the approximate consistency and evaluate the effectiveness of the proposed approximate algorithms.

Section II introduces related work. In Section III, we discuss the relationship between this article and our previous works. Section IV presents two kernel selection criteria. Section V defines approximate consistency and analyzes the approximate consistency of the MCM and Nyström approximation. Section VI elaborates on the designs of the approximate algorithms for kernel selection. In Sections VII and VIII, we conduct experiments and conclude this article.

II. Related Work

Traditional kernel-based learning algorithms suffer from high time and space complexities due to their usage of the kernel matrix. Kernel matrix approximation can be adopted to effectively reduce the computational and storage burdens. Here, we introduce two types of kernel matrix approximation methods, MCM approximation and column sampling-based approximation. Using MCM approximation [23], rather than the original kernel matrix, we can approximately solve the eigensystems of the kernel matrix in a time complexity of $O(l\log(l))$ by employing a fast Fourier transform (FFT) [24]–[27]. Approximations based on column sampling have also been extensively studied, with representative methods, including the classical Nyström method with different kinds of sampling strategies [28]–[33], the modified Nyström method [34], [35], matrix least squares approximation [36], and CUR matrix decomposition [37].

Aside from their computational analyses, the kernel matrix approximation methods have also been theoretically studied [24], [26], [30]–[32], [34], [38]. Most existing theoretical analyses for kernel matrix approximation provide bounds for the discrepancy between the approximate matrix and the original kernel matrix for an appropriate norm (such as the Frobenius norm, spectral norm, or trace norm). However, these analyses are independent of the learning problem at hand and cannot reveal the impact of the approximation of kernel matrix on learning algorithms. Recent studies [26], [39]–[41] demonstrate the influence of kernel matrix approximation on the learned hypothesis, but none of these measure the influence of kernel matrix approximation on kernel selection. The existing theoretical analyses on kernel matrix approximation are not sufficient to justify the appositeness of introducing matrix approximation in kernel selection. This article defines approximate consistency to measure the discrepancy between the approximate criterion computed with the approximate matrix and the accurate criterion computed with the kernel matrix and also analyzes the convergence speed of the discrepancy for different matrix approximation algorithms.

III. Relations to Previous Works

The idea of approximate kernel selection was first proposed in [26]. To theoretically study the rationale behind approximate kernel selection, the notion of approximate consistency was defined [42], [43]. This article refines the definition in [42] and takes approximate consistency as a basic property of approximate kernel selection algorithms rather than matrix approximation algorithms. Besides refining the definition, we have the following additional contributions. First, we provide detailed explanations and theoretical analyses for MCM to demonstrate its utility in kernel selection. Second, we design novel approximate algorithms for kernel selection by exploiting the computational virtues of the MCM and Nyström approximation. Third, we provide theoretical analyses for the approximate kernel selection algorithms. Fourth, we provide extensive empirical evidence to support the theoretical findings for the approximate consistency.

From an algorithmic perspective, the MCM approximate kernel selection algorithm and the Nyström approximate kernel selection algorithm designed in this article are inspired by the computational skills in [26] and [44], respectively. However, there are two main differences between this article and [26] and [44]. First, no kernel selection criteria were given in [26] and [44], and kernel matrix approximations were adopted to accelerate the training of LSSVM. In this article, two kernel selection criteria based on error estimation are presented. The MCM and Nyström approximate kernel selection algorithms are designed for accelerating the computation of the kernel selection criteria. Second, the theoretical bounds given in [26] and [44] measured the discrepancy between the approximate hypothesis and the accurate one, quantifying the impact of the MCM and Nyström approximations on the hypothesis of LSSVM. In this article, approximate consistency is used to measure the discrepancy between the approximate criterion on the MCM or Nyström approximate matrix and the accurate one on the kernel matrix.

The previous work [45] and this article both adopt the ingenious algorithm proposed in [24] to generate MCMs. However, three points distinguish this article from [45]. First, [45] studied the kernel combination problem specifically for regression tasks, whereas this article studies the kernel selection problem for both classification and regression tasks. Second, no kernel selection criteria were given in [45], and the combination weights of base kernels were optimized using kernel ridge regression (KRR). In contrast, two kernel selection criteria are proposed in this article, and the optimal kernel is selected by minimizing the proposed criteria. Third, the approximation error bound given in [45] was for the discrepancy between the approximate and accurate hypotheses, measuring the impact of the approximation on the hypothesis of KRR. In this article, the approximate consistency determines the discrepancy between the approximate and accurate criteria, accessing the impact of the MCM or Nyström approximation on kernel selection criteria.

IV. Kernel Selection Criteria

The set of l labeled data points is denoted as $\mathcal{S} = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)) \in (\mathcal{X} \times \mathcal{Y})^l$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the input space and $\mathcal{Y}$ is the output domain. For the classification case, $\mathcal{Y} = \{-1, 1\}$, and for the regression case, $\mathcal{Y} = \mathbb{R}$. We assume that $|y| \leq M$ and $y \in \mathcal{Y}$, where M is a given constant. We consider the Mercer kernel κ in this article, which is a continuous, symmetric, and positive definite (SPD) function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$ [1]. The kernel matrix $\mathbf{K} = [\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^l$, defined on a finite set of inputs $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l\} \subseteq \mathcal{X}$, is SPD. We denote the reproducing kernel Hilbert space (RKHS) of the kernel κ as $\mathcal{H}_\kappa$ [46], which is defined as $\mathcal{H}_\kappa = \overline{\text{span}}\{\kappa(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathcal{X}\}$. For $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa}$ of $\mathcal{H}_\kappa$ is $\langle \kappa(\boldsymbol{x}, \cdot), \kappa(\boldsymbol{x}', \cdot) \rangle_{\mathcal{H}_\kappa} = \kappa(\boldsymbol{x}, \boldsymbol{x}')$.

Now, we provide two criteria for kernel selection. The first criterion is based on the regularized square loss $\mathcal{E}(f) = (1/l) \sum_{i=1}^l (f(\boldsymbol{x}_i) - y_i)^2 + \mu \|f\|_{\mathcal{H}_\kappa}^2$, where μ denotes the regularization parameter and $\|\cdot\|_{\mathcal{H}_\kappa}$ is the norm in $\mathcal{H}_\kappa$ induced by $\langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa}$. The optimal function is $f_\kappa = \arg\min_{f \in \mathcal{H}_\kappa} \mathcal{E}(f)$. By the representer theorem [47], we have $f_\kappa = \sum_{i=1}^l \alpha_i \kappa(\boldsymbol{x}_i, \cdot)$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^{\mathrm{T}} = (\mathbf{K} + \mu l \mathbf{I})^{-1} \boldsymbol{y}$, where $\boldsymbol{y} = (y_1 \ldots, y_l)^{\mathrm{T}}$ and $\mathbf{I}$ denotes the identity matrix. Therefore, $\|f_\kappa\|_{\mathcal{H}_\kappa}^2 = \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{K} \boldsymbol{\alpha} = \boldsymbol{y}^{\mathrm{T}} \mathbf{K}_\mu^{-1} \mathbf{K} \mathbf{K}_\mu^{-1} \boldsymbol{y}$, where $\mathbf{K}_\mu = \mathbf{K} + \mu l \mathbf{I}$. Denoting $\boldsymbol{f}_\kappa = (f_\kappa(\boldsymbol{x}_1), \ldots, f_\kappa(\boldsymbol{x}_l))^{\mathrm{T}}$, we have $\boldsymbol{f}_\kappa = \mathbf{K}\boldsymbol{\alpha} = \mathbf{K}\mathbf{K}_\mu^{-1} \boldsymbol{y}$, which implies $\boldsymbol{f}_\kappa - \boldsymbol{y} = \mathbf{K}\mathbf{K}_\mu^{-1} \boldsymbol{y} - \mathbf{K}_\mu \mathbf{K}_\mu^{-1} \boldsymbol{y} = -\mu l \mathbf{K}_\mu^{-1} \boldsymbol{y}$. Now

$$\begin{aligned} \mathcal{E}(f_\kappa) &= \frac{1}{l} (\boldsymbol{f}_\kappa - \boldsymbol{y})^{\mathrm{T}} (\boldsymbol{f}_\kappa - \boldsymbol{y}) + \mu \|f_\kappa\|_{\mathcal{H}_\kappa}^2 \\ &= \mu^2 \, l \boldsymbol{y}^{\mathrm{T}} \mathbf{K}_\mu^{-1} \mathbf{K}_\mu^{-1} \boldsymbol{y} + \mu \boldsymbol{y}^{\mathrm{T}} \mathbf{K}_\mu^{-1} \mathbf{K} \mathbf{K}_\mu^{-1} \boldsymbol{y} \\ &= \mu \boldsymbol{y}^{\mathrm{T}} \mathbf{K}_\mu^{-1} \boldsymbol{y} \end{aligned}$$

where $\mathcal{E}(f_\kappa)$ is the regularized empirical error of the optimal function f_κ. For a fixed regularization parameter μ, $\mathcal{E}(f_\kappa)$ only depends on the kernel κ. It is known that the kernel function κ has a one-to-one correspondence to RKHS $\mathcal{H}_\kappa$ [46]. Different kernels correspond to different RKHSs. In different RKHSs, different optimal functions are derived. We can select the optimal function f_κ, which makes $\mathcal{E}(f_\kappa)$ the smallest, from all optimal functions. We write

$$\mathcal{C}_{\text{ree}}(\mathbf{K}) = \mathcal{E}(f_\kappa) = \mu \boldsymbol{y}^{\mathrm{T}} \mathbf{K}_\mu^{-1} \boldsymbol{y} \tag{1}$$

where "ree" stands for "regularized empirical error." The optimal kernel can be found by $\kappa^* = \arg\min_{\kappa \in \mathcal{K}} \mathcal{C}_{\text{ree}}(\mathbf{K})$, where $\mathcal{K} = \{\kappa_1, \ldots, \kappa_N\}$ is a prescribed set of kernels.

In the following, we present the second kernel selection criterion. We consider the case where the observed output is corrupted by noise. Specifically, we always assume that $y_i = \dot{y}_i + \xi_i$, $1 \leq i \leq l$, where $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_l]^{\mathrm{T}}$ are random variables with finite covariance matrix $\mathbf{C}$ and mean 0. $\dot{\boldsymbol{y}} = [\dot{y}_1, \ldots \dot{y}_l]^{\mathrm{T}}$ is the underlying true output. Now, $\boldsymbol{f}_\kappa = \mathbf{K}\boldsymbol{\alpha} = \mathbf{K}\mathbf{K}_\mu^{-1} \boldsymbol{y}$ is referred to as an estimate of $\dot{\boldsymbol{y}}$. We have

$$\begin{aligned} \frac{1}{l} \mathbb{E}_{\boldsymbol{\xi}} \|\boldsymbol{f}_\kappa - \dot{\boldsymbol{y}}\|^2 &= \frac{1}{l} \|\mathbb{E}_{\boldsymbol{\xi}} \boldsymbol{f}_\kappa - \dot{\boldsymbol{y}}\|^2 + \frac{1}{l} \text{trace}(\text{var}_{\boldsymbol{\xi}}(\boldsymbol{f}_\kappa)) \\ &= \frac{1}{l} \|\mathbf{K}\mathbf{K}_\mu^{-1} \dot{\boldsymbol{y}} - \dot{\boldsymbol{y}}\|^2 + \frac{1}{l} \text{trace}\big(\mathbf{C}\mathbf{K}^2 \mathbf{K}_\mu^{-2}\big) \\ &= \underbrace{\mu^2 l \dot{\boldsymbol{y}}^{\mathrm{T}} \mathbf{K}_\mu^{-2} \dot{\boldsymbol{y}}}_{\text{bias}(\mathbf{K})} + \underbrace{\frac{1}{l} \text{trace}\big(\mathbf{C}\mathbf{K}^2 \mathbf{K}_\mu^{-2}\big)}_{\text{variance}(\mathbf{K})}. \end{aligned}$$

If $\mathbf{C}$ is equal to $\sigma^2 \mathbf{I}$, we have

$$\mathcal{C}_{\text{ipe}}(\mathbf{K}) = \underbrace{\mu^2 l \dot{\boldsymbol{y}}^{\mathrm{T}} \mathbf{K}_\mu^{-2} \dot{\boldsymbol{y}}}_{\text{bias}(\mathbf{K})} + \underbrace{\frac{\sigma^2}{l} \text{trace}\big(\mathbf{K}^2 \mathbf{K}_\mu^{-2}\big)}_{\text{variance}(\mathbf{K})} \tag{2}$$

where "ipe" stands for "in-sample prediction error." The optimal kernel $\kappa^* = \arg\min_{\kappa \in \mathcal{K}} \mathcal{C}_{\text{ipe}}(\mathbf{K})$.

V. Approximate Consistency

If we have a prescribed kernel set $\mathcal{K}$, a selection criterion $\mathcal{C}(\mathbf{K})$, and a matrix approximation algorithm $\mathcal{A}$, which generates the approximation $\widetilde{\mathbf{K}}$, the approximate kernel selection is designed to select the kernel κ^* as

$$\kappa^* = \arg\min_{\kappa \in \mathcal{K}} \; \mathcal{C}(\mathcal{A}(\mathcal{S}, \kappa)) = \arg\min_{\kappa \in \mathcal{K}} \; \mathcal{C}(\widetilde{\mathbf{K}}). \tag{3}$$

Let us look at the criteria $\mathcal{C}_{\text{ree}}(\mathbf{K})$ and $\mathcal{C}_{\text{ipe}}(\mathbf{K})$, presented in Section IV. Since the matrix inverse is required, the time complexities of $\mathcal{C}_{\text{ree}}(\mathbf{K})$ and $\mathcal{C}_{\text{ipe}}(\mathbf{K})$ are both $O(l^3)$, which is prohibitive for large-scale data. In Section VI, we will design approximate kernel selection algorithms by employing the MCM and Nyström approximation, which exploits the specific structure of $\widetilde{\mathbf{K}}$ to efficiently conduct kernel selection.

However, before designing the algorithms, we solve the theoretical problems faced by approximate kernel selection, i.e., investigate how the approximation on the kernel matrix impacts the criterion. To do so, we analyze the discrepancy between the approximate criterion $\mathcal{C}(\widetilde{\mathbf{K}})$ and the accurate one $\mathcal{C}(\mathbf{K})$. More specifically, for finite samples, we should give the upper bound of the discrepancy between the approximate and accurate criteria; for large samples, we need to show under what conditions and at what speed, the discrepancy between the approximate and accurate criteria converges to 0. We define and analyze the approximate consistency to solve these problems. We denote an approximate kernel selection algorithm $\mathcal{AKS}$ as a 2-tuple: $\mathcal{AKS} = (\mathcal{C}(\mathbf{K}), \mathcal{A})$.

Definition 1: For an approximate kernel selection algorithm $\mathcal{AKS} = (\mathcal{C}(\mathbf{K}), \mathcal{A})$, we say $\mathcal{AKS}$ is of strong approximate consistency if $|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\widetilde{\mathbf{K}})| \leq \varepsilon(l)$, where $\lim_{l \to \infty} \varepsilon(l) \to 0$. We say $\mathcal{AKS}$ is of p-order approximate consistency if $|\mathcal{C}(\mathbf{K}) - \mathcal{C}(\widetilde{\mathbf{K}})| \leq \varepsilon(l)$, where $\lim_{l \to \infty} \varepsilon(l)/l^p \to 0$.[1]

A. Approximate Consistency of MCM Approximation

To facilitate representation, we introduce the following notations. We use $\mathbb{N}$ to denote a set of positive integers. For $m \in \mathbb{N}$, $[m] = \{0, 1, \ldots, m-1\}$. For a positive integer p and $\boldsymbol{m} = (m_0, m_1, \ldots, m_{p-1}) \in \mathbb{N}^p$, we denote

$\Pi_{\boldsymbol{m}} = m_0 m_1 \ldots m_{p-1}$, (continued product)

$[\boldsymbol{m}] = [m_0] \times [m_1] \times \cdots \times [m_{p-1}]$ (Cartesian product).

A circulant matrix can be represented in the following form:

$$\mathbf{C} = \begin{bmatrix} c_0 & c_{m-1} & \ldots & c_1 \\ c_1 & c_0 & \ldots & c_2 \\ \vdots & \vdots & \ddots & \vdots \\ c_{m-1} & c_{m-2} & \ldots & c_0 \end{bmatrix}.$$

[1] This definition is a refinement of that in [42]. Taking the approximate consistency as a basic property of $\mathcal{AKS}$ instead of $\mathcal{A}$ is more appropriate since the approximate consistency is closely related to both $\mathcal{C}(\mathbf{K})$ and $\mathcal{A}$.

Each column of $\mathbf{C}$ is a cyclic shift of its left column. It is fully determined by its first column.

Multilevel circulant matrices [48] are defined recursively. For an integer $s \geq 1$, an $(s+1)$-level circulant matrix is a block matrix, where each block is an s-level circulant matrix. For $\boldsymbol{m} \in \mathbb{N}^p$, we use multidimensional indices $\boldsymbol{i} = (i_0, \ldots, i_{p-1})$, $\boldsymbol{j} = (j_0, \ldots, j_{p-1}) \in [\boldsymbol{m}]$ to locate the entries of a p-level circulant matrix $\mathbf{A}_{\boldsymbol{m}}$. According to [49], for $\boldsymbol{m} \in \mathbb{N}^p$, $\mathbf{A}_{\boldsymbol{m}} = [a_{\boldsymbol{i},\boldsymbol{j}} : \boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]]$ is a p-level circulant matrix if, for any $\boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]$

$$a_{\boldsymbol{i},\boldsymbol{j}} = a_{i_0 - j_0 (\mathrm{mod}\ m_0), \ldots, i_{p-1} - j_{p-1} (\mathrm{mod}\ m_{p-1})}.$$

We can fully determine $\mathbf{A}_{\boldsymbol{m}}$ by its first column $a_{\boldsymbol{i},\boldsymbol{0}}$ with $\boldsymbol{0} = (0, \ldots, 0) \in \mathbb{R}^p$, so we write $\mathbf{A}_{\boldsymbol{m}} = \mathrm{circ}_{\boldsymbol{m}}[a_{\boldsymbol{i}} : \boldsymbol{i} \in [\boldsymbol{m}]]$, where $a_{\boldsymbol{i}} = a_{\boldsymbol{i},\boldsymbol{0}}$, for $\boldsymbol{i} \in [\boldsymbol{m}]$. We introduce an example of the MCM to explain the above notations [45] in the Supplement Material.

In the following, we will show how to approximate the kernel matrix with an MCM. We consider the radial basis function (RBF) kernels for MCM approximation, such as the Gaussian kernel [50]. For $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, we assume that $K \in L^1(\mathbb{R})$ on $\mathcal{X}$, such that $\kappa(\boldsymbol{x}, \boldsymbol{x}') = K(\|\boldsymbol{x} - \boldsymbol{x}'\|_2)$. Since $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x}', \boldsymbol{x})$, K is always an even function. Without loss of generality, for $\boldsymbol{m} \in \mathbb{N}^p$, it is assumed that the number of data points in $\mathcal{S}$ is $\Pi_{\boldsymbol{m}}$, that is, $|\mathcal{S}| = l = \Pi_{\boldsymbol{m}}$. The data points in $\mathcal{S}$ can be relabeled as $\mathcal{S} = \{(\boldsymbol{x}_{\boldsymbol{i}}, y_{\boldsymbol{i}}) : \boldsymbol{i} \in [\boldsymbol{m}]\}$. Now, $\mathbf{K}$ is rewritten as $\mathbf{K}_{\boldsymbol{m}} = [K(\|\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{x}_{\boldsymbol{j}}\|_2) : \boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]]$.

Algorithm 1 Construction of an MCM

Input: p, $\boldsymbol{m} \in \mathbb{N}^p$, K, a sequence of positive numbers $\boldsymbol{h}_{\boldsymbol{m}} = (h_{\boldsymbol{m},0}, \ldots, h_{\boldsymbol{m},p-1}) \in \mathbb{R}^p$;
Output: The MCM $\mathbf{U}_{\boldsymbol{m}}$;
1: For any $\boldsymbol{i} \in [\boldsymbol{m}]$, calculate

$$t_{\boldsymbol{i}} = K(\|[i_s h_{\boldsymbol{m},s} : s \in [p]]\|_2);$$

2: For any $\boldsymbol{i} \in [\boldsymbol{m}]$ and $s \in [p]$, let

$$D_{\boldsymbol{i},s} = \begin{cases} \{0\}, & i_s = 0, \\ \{i_s, m_s - i_s\}, & 1 \leq i_s \leq m_s - 1, \end{cases}$$

and $D_{\boldsymbol{i}} = D_{\boldsymbol{i},0} \times D_{\boldsymbol{i},1} \times \cdots \times D_{\boldsymbol{i},p-1}$;
3: For any $\boldsymbol{i} \in [\boldsymbol{m}]$ calculate $u_{\boldsymbol{i}} = \sum_{\boldsymbol{j} \in D_{\boldsymbol{i}}} t_{\boldsymbol{j}}$;
4: **Return** $\mathbf{U}_{\boldsymbol{m}} = \mathrm{circ}_{\boldsymbol{m}}[u_{\boldsymbol{i}} : \boldsymbol{i} \in [\boldsymbol{m}]]$;

We introduce Algorithm 1 to demonstrate the construction of an MCM $\mathbf{U}_{\boldsymbol{m}}$ as the approximation of $\mathbf{K}_{\boldsymbol{m}}$ [24]. We give an example [45] to explain Algorithm 1 in the Supplement Material.

$\mathbf{U}_{\boldsymbol{m}}$, shown in Algorithm 1, is an MCM specifically designed for kernel approximation. First, $\mathbf{U}_{\boldsymbol{m}}$ is symmetric, and $\mathbf{U}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}}$ is invertible. The invertibility of $\mathbf{U}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}}$ was proven in [24]. Since $\mu l \mathbf{I}_{\boldsymbol{m}}$ is symmetric, we only need to guarantee the symmetry of $\mathbf{U}_{\boldsymbol{m}}$. The symmetry of $\mathbf{U}_{\boldsymbol{m}}$ is proven in Proposition 1. We provide the Proof of Proposition 1 in the Supplement Material [45]. It is worth noting that step 2 in Algorithm 2 guarantees the symmetry of the MCM $\mathbf{U}_{\boldsymbol{m}}$. Second, we can see that all the approximate elements $t_{\boldsymbol{i}} = K(\|[i_s h_{\boldsymbol{m},s} : s \in [p]]\|_2)$ of the original kernel matrix have been summed into the first column of the MCM $\mathbf{U}_{\boldsymbol{m}}$. This is carried out in steps 1 and 3 in Algorithm 2.

Algorithm 2 MCM Approximate Kernel Selection

Input: $\boldsymbol{y} = \{y_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{m}]\}$, $\mathcal{K} = \{\kappa_1, \ldots, \kappa_N\}$, μ;
Output: The optimal kernel κ^*;
1: **Initialize**: $\mathcal{C}^* = \infty$;
2: **for** each $\kappa \in \mathcal{K}$ **do**
3: Calculate $[u_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{m}]]$ according to Algorithm 1;
4: Calculate $\boldsymbol{v} = \boldsymbol{\Phi}[u_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{m}]]$ by mFFT;
5: Calculate $\boldsymbol{\eta} = \boldsymbol{\Phi}\boldsymbol{y}$ using mFFT;
6: Calculate $\boldsymbol{\tau} = \mathrm{diag}\left(\frac{1}{v_{\boldsymbol{j}} + \mu l} : \boldsymbol{j} \in [\boldsymbol{m}]\right)\boldsymbol{\eta}$;
7: Calculate $\boldsymbol{\zeta} = \frac{1}{\Pi_{\boldsymbol{m}}}\boldsymbol{\Phi}^{\mathrm{H}}\boldsymbol{\tau}$ using inverse mFFT;
8: $\mathcal{C}_{\mathrm{ree}}(\mathbf{U}_{\boldsymbol{m}}) = \mu \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\zeta}$;
9: **if** $\mathcal{C}_{\mathrm{ree}}(\mathbf{U}_{\boldsymbol{m}}) \leq \mathcal{C}^*$ **then**
10: $\mathcal{C}^* = \mathcal{C}_{\mathrm{ree}}(\mathbf{U}_{\boldsymbol{m}})$;
11: $\kappa^* = \kappa$;
12: **Return** κ^*;

Proposition 1: For p and $\boldsymbol{m} \in \mathbb{N}^p$, the MCM $\mathbf{U}_{\boldsymbol{m}}$ constructed by Algorithm 1 is symmetric.

Theorem 1 shows the strong approximate consistency of MCM approximation under $\mathcal{C}_{\mathrm{ree}}(\mathbf{K}_{\boldsymbol{m}})$. We prove the upper bound of the discrepancy between the approximate and accurate criteria and then analyze the convergence of the discrepancy. The proof of Theorem 1 is given in the Supplement Material.

Theorem 1: If the assumptions H1–H2 hold, and furthermore, there are positive constants c_3 and r_1 satisfying $|y_{\boldsymbol{i}}| \leq c_3 \mathrm{e}^{-r_1 v_{\boldsymbol{m}}(\boldsymbol{i})}$, for any $\boldsymbol{m} \in \mathbb{N}^p$ and $\boldsymbol{i} \in [\boldsymbol{m}]$, where $v_{\boldsymbol{m}}(\boldsymbol{i}) = \|(\boldsymbol{m}/2) - \boldsymbol{i}\|_2$, and then, we have

$$\lim_{\boldsymbol{m} \to \infty} |\mathcal{C}_{\mathrm{ree}}(\mathbf{K}_{\boldsymbol{m}}) - \mathcal{C}_{\mathrm{ree}}(\mathbf{U}_{\boldsymbol{m}})| = 0$$

where $\boldsymbol{m} \to \infty$ indicates that all components of $\boldsymbol{m}$ go to infinity.

H1: There are positive constants c_0 and β such that $|K(x) - K(x')| \leq c_0 |x - x'|^{\beta}$ for $x, x' \in \mathbb{R}$.
H2: There is a positive constant h such that $h_{\boldsymbol{m},j} \geq h$ for $\boldsymbol{m} \in \mathbb{N}^p$ and $j \in [p]$.
H3: There are positive constants λ_1 and c_1 such that $|K(x)| \leq c_1 \mathrm{e}^{-\lambda_1 |x|}$ for $x \in \mathbb{R}$.
H4: There are positive constants λ_2 and c_2 such that

$$\begin{aligned} &|\|x_{\boldsymbol{i}} - x_{\boldsymbol{j}}\|_2 - \|[(i_s - j_s) h_{\boldsymbol{m},s} : s \in [p]]\|_2| \\ &\qquad \leq c_2 \sum_{s \in [p]} (\mathrm{e}^{-\lambda_2 \delta_{m_s}(i_s)} + \mathrm{e}^{-\lambda_2 \delta_{m_s}(j_s)}) \end{aligned}$$

for $\boldsymbol{m} \in \mathbb{N}^p$ and $\boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]$, where $\delta_m(j) = (m/2) - |(m/2) - j|$ for $m \in \mathbb{N}$ and $j \in [m]$.

We further study the approximate consistency of MCM approximation under $\mathcal{C}_{\mathrm{ipe}}(\mathbf{K}_{\boldsymbol{m}})$. Theorem 2 shows the upper bound on the variance and the bias term of $\mathcal{C}_{\mathrm{ipe}}(\mathbf{K}_{\boldsymbol{m}})$. Based on Theorem 2, we can obtain Theorem 3, which shows the strong approximate consistency of MCM approximation under $\mathcal{C}_{\mathrm{ipe}}(\mathbf{K}_{\boldsymbol{m}})$. The Proof of Theorem 2 is given in the Supplement Material.

Theorem 2: If the assumptions in Theorem 1 hold, we have

$$|\text{variance}(\mathbf{K}_{\boldsymbol{m}}) - \text{variance}(\mathbf{U}_{\boldsymbol{m}})| \leq c\sigma^2(m_{\min})^{-1}$$

for a positive constant c, where $m_{\min} = \min\{m_s : s \in [p]\}$. If for any $\boldsymbol{m} \in \mathbb{N}^p$ and $\boldsymbol{i} \in [\boldsymbol{m}]$, there exist positive constants c_3 and r_1 such that $|\dot{y}_{\boldsymbol{i}}| \leq c_3 \mathrm{e}^{-r_1 \nu_{\boldsymbol{m}}(\boldsymbol{i})}$, where $\nu_{\boldsymbol{m}}(\boldsymbol{i}) = \|(\boldsymbol{m}/2) - \boldsymbol{i}\|_2$, and then, for any $\boldsymbol{m} \in \mathbb{N}^p$

$$|\text{bias}(\mathbf{K}_{\boldsymbol{m}}) - \text{bias}(\mathbf{U}_{\boldsymbol{m}})| \leq c\mu^2 \Pi_{\boldsymbol{m}}^{3/2} \mathrm{e}^{-r' m_{\min}}$$

for positive constants c and r'.

Theorem 3: If the assumptions in Theorem 1 hold, we have

$$\lim_{\boldsymbol{m}\to\infty} |\mathcal{C}_{\text{ipe}}(\mathbf{K}_{\boldsymbol{m}}) - \mathcal{C}_{\text{ipe}}(\mathbf{U}_{\boldsymbol{m}})| = 0.$$

B. Approximate Consistency of Nyström Approximation

We now briefly review the Nyström approximation [28]. We first randomly select c columns of $\mathbf{K}$. We denote $\mathbf{C}$ as an $l \times c$ matrix formed by the selected columns. We use $\mathbf{W}$ to denote the $c \times c$ matrix composed of the intersection between the selected c columns and the corresponding c rows of $\mathbf{K}$. The Nyström approximate matrix is $\widetilde{\mathbf{K}} = \mathbf{C}\mathbf{W}_k^+\mathbf{C}^{\mathrm{T}} \approx \mathbf{K}$, where $\mathbf{W}_k$ is the optimal rank k approximation to $\mathbf{W}$ and $\mathbf{W}_k^+$ is the generalized inverse of $\mathbf{W}_k$.

Modified Nyström approximation [34] presents a tighter approximation error bound than the classical Nyström approximation but has a higher computational cost. We can write the approximate matrix of modified Nyström approximation as $\widetilde{\mathbf{K}} = \mathbf{C}(\mathbf{C}^+\mathbf{K}(\mathbf{C}^+)^{\mathrm{T}})\mathbf{C}^{\mathrm{T}}$.

Although there are many different versions of Nyström approximation with different sampling strategies [28]–[32], [34], we concentrate on the approximations with $(1+\epsilon)$ relative-error bounds, where ϵ is independent of l. The bound for the Nyström approximation [32] is derived using leverage score-based column sampling [37], which states that for $\epsilon \in (0, 1]$ and a failure probability $\delta \in (0, 1]$

$$\|\mathbf{K} - \widetilde{\mathbf{K}}\|_* \leq (1+\epsilon)\|\mathbf{K} - \mathbf{K}_k\|_* \tag{4}$$

holds with a probability of at least $0.6 - \delta$. For the modified Nyström approximation [34], the bound is derived by combining the near-optimal sampling [51] and the error-driven adaptive sampling [52]

$$\mathbb{E}(\|\mathbf{K} - \widetilde{\mathbf{K}}\|_{\mathrm{F}}) \leq (1+\epsilon)\|\mathbf{K} - \mathbf{K}_k\|_{\mathrm{F}}.$$

Before analyzing the approximate consistency of the classical and modified Nyström approximations, we introduce two assumptions.

Assumption 1: For the rank $k \leq c \ll l$ and $\rho \in (0, 1/2)$, we assume that $\lambda_k(\mathbf{K}) = \Omega(l/c^\rho)$ and $\lambda_{k+1}(\mathbf{K}) = O(l/c^{1-\rho})$, where ρ is used to characterize the gap between the kth and $(k+1)$th eigenvalues.

Assumption 2: We assume that the sampling size c is a small ratio r of l, and the rank parameter k is a constant.

Assumption 1 is not a strong assumption. As suggested in [53], the eigenvalues of the kernel matrix have polynomial or exponential decay. The eigenvalues of the Gaussian kernels have exponential decay [18]. Assumption 1 is always weaker than exponential decay, even when ρ goes to 0. When ρ is close to 1/2, Assumption 1 is weaker than polynomial decay. Assumption 1 was adopted in [40] and [41] and experimentally tested in [40]. Assumption 2 is a common setting of the Nyström approximation. The constant rank was adopted in [34].

Theorem 4 shows the (1/2)-order approximate consistency of the Nyström approximation under $\mathcal{C}_{\text{ree}}(\mathbf{K})$. Theorem 5 shows the strong approximate consistency of the modified Nyström approximation under $\mathcal{C}_{\text{ree}}(\mathbf{K})$. The proofs are given in the Supplement Material.

Theorem 4: For $\mathcal{C}_{\text{ree}}(\mathbf{K})$, if Assumptions 1 and 2 hold, we have that $|\mathcal{C}_{\text{ree}}(\mathbf{K}) - \mathcal{C}_{\text{ree}}(\widetilde{\mathbf{K}})| \leq \varepsilon(l)$ for $\delta \in (0, 1]$ and $\epsilon \in (0, 1]$ holds with a probability of at least $0.6 - \delta$, where $\widetilde{\mathbf{K}}$ is produced by the Nyström approximation with leverage score sampling

$$\varepsilon(l) = \frac{\tau M^2(1+\epsilon)}{\mu r^{1-\rho} l^{1-\rho}}(l-k)$$

for constant τ and $\lim_{l\to\infty} \varepsilon(l)/l^{(1/2)} \to 0$.

Theorem 5: For $\mathcal{C}_{\text{ree}}(\mathbf{K})$, if Assumptions 1 and 2 hold, we have $\mathbb{E}(|\mathcal{C}_{\text{ree}}(\mathbf{K}) - \mathcal{C}_{\text{ree}}(\widetilde{\mathbf{K}})|) \leq \varepsilon(l)$, where $\widetilde{\mathbf{K}}$ is produced by the modified Nyström approximation

$$\varepsilon(l) = \frac{\tau M^2(1+\epsilon)}{\mu r^{1-\rho} l^{1-\rho}}\sqrt{l-k}$$

for constant τ and $\lim_{l\to\infty} \varepsilon(l) \to 0$.

VI. Approximate Kernel Selection Algorithms

Under the theoretical guarantee of approximate consistency, we design approximate kernel selection algorithms using the MCM and Nyström approximations.

A. Approximate Kernel Selection With MCM Approximation

Here, we adopt the inverse of $\mathbf{U}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}}$ to approximate the inverse of $\mathbf{K}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}}$, where $\mathbf{I}_{\boldsymbol{m}}$ is an identity matrix. The eigenvalues and eigenvectors of an MCM $\mathbf{U}_{\boldsymbol{m}}$ [49][2] can be represented as $\mathbf{U}_{\boldsymbol{m}} = (1/\Pi_{\boldsymbol{m}})\boldsymbol{\Phi}^{\mathrm{H}}\text{diag}(\boldsymbol{v})\boldsymbol{\Phi}$, where the vector of eigenvalues is $\boldsymbol{v} = \boldsymbol{\Phi}[u_{\boldsymbol{j}} : \boldsymbol{j} \in [\boldsymbol{m}]]$. $\boldsymbol{\Phi}$ is the Kronecker product of the Fourier matrices. For any vector $\boldsymbol{x} = [x_{\boldsymbol{i}} : \boldsymbol{i} \in [\boldsymbol{m}]]$, $\boldsymbol{\Phi}\boldsymbol{x}$ is the multidimensional discrete Fourier transform (mDFT) of $\boldsymbol{x}$. Therefore, we can compute $\boldsymbol{\Phi}\boldsymbol{x}$ through a multidimensional FFT (mFFT). We can compute the eigenvalues of an MCM by applying mFFT to its first column [49]. It follows that

$$(\mathbf{U}_{\boldsymbol{m}} + \mu l \mathbf{I}_{\boldsymbol{m}})^{-1} = \frac{1}{\Pi_{\boldsymbol{m}}}\boldsymbol{\Phi}^{\mathrm{H}}\text{diag}\left(\frac{1}{\upsilon_{\boldsymbol{j}} + \mu l} : \boldsymbol{j} \in [\boldsymbol{m}]\right)\boldsymbol{\Phi}.$$

Now, we design an MCM approximate kernel selection algorithm (see Algorithm 2) for $\mathcal{C}_{\text{ree}}$. The time complexity of Algorithm 2 is shown in Theorem 6.

Theorem 6: The time complexity of Algorithm 2 is $O(N(l \max\{\log(l), p\}))$.

Proof: We assume that $l = \Pi_{\boldsymbol{m}}$. The time complexity of step 3 is $O(lp)$ for RBF kernels, where p is the number of

[2]For complete lemmas, please see Section V in the Supplement Material.

levels of the MCM. We know that the time complexity of steps 4, 5, and 7 is $O(l\log(l))$, since mFFT can be applied with the $O(l\log(l))$ complexity. The time complexity of step 6 is $O(l)$. Because the number of candidate kernels is N, the total time complexity is $O(N(l\max\{\log(l), p\}))$. ■

The time complexity of computing $\mathcal{C}_{\text{ree}}(\mathbf{U}_{\boldsymbol{m}})$ for each candidate kernel κ is $O(l\max\{\log(l), p\})$, which is quasi-linear in the number of samples l and much lower than $O(l^3)$. The space complexity of $\mathcal{C}_{\text{ree}}(\mathbf{U}_{\boldsymbol{m}})$ is $O(l)$ since we only need to store the first column of $\mathbf{U}_{\boldsymbol{m}}$, which is lower than the $O(l^2)$ space complexity required to store the kernel matrix.

We further discuss the approximate computation of the kernel selection criterion $\mathcal{C}_{\text{ipe}}$, defined in (2). We can approximately compute the bias term $\mu^2 l\dot{\boldsymbol{y}}^{\mathrm{T}}(\mathbf{K}_{\boldsymbol{m}}+\mu l\mathbf{I}_{\boldsymbol{m}})^{-2}\dot{\boldsymbol{y}}$ of $\mathcal{C}_{\text{ipe}}$, using steps 5–7 in Algorithm 2, twice. For the variance term $(\sigma^2/l)\text{trace}(\mathbf{K}_{\boldsymbol{m}}^2(\mathbf{K}_{\boldsymbol{m}}+\mu l\mathbf{I}_{\boldsymbol{m}})^{-2})$ of $\mathcal{C}_{\text{ipe}}$, we compute the eigenvalues of $\mathbf{U}_{\boldsymbol{m}}$ as the approximations of the eigenvalues of the kernel matrix $\mathbf{K}_{\boldsymbol{m}}$. Therefore

$$\text{trace}\left(\mathbf{K}_{\boldsymbol{m}}^2(\mathbf{K}_{\boldsymbol{m}}+\mu l\mathbf{I}_{\boldsymbol{m}})^{-2}\right)\approx\sum_{\boldsymbol{i}\in[\boldsymbol{m}]}\left(\frac{v_{\boldsymbol{i}}}{v_{\boldsymbol{i}}+\mu l}\right)^2.$$

B. Approximate Kernel Selection With the Nyström Approximation

The Nyström approximate matrix can be represented as $\widetilde{\mathbf{K}}=\mathbf{C}\mathbf{W}_k^{+}\mathbf{C}^{\mathrm{T}}\approx\mathbf{K}$. The SVD of $\mathbf{W}$ is $\mathbf{W}=\mathbf{U}_W\boldsymbol{\Sigma}_W\mathbf{U}_W^{\mathrm{T}}$, so $\mathbf{W}_k^{+}=\mathbf{U}_{\mathbf{W},k}\boldsymbol{\Sigma}_{\mathbf{W},k}^{+}\mathbf{U}_{\mathbf{W},k}^{\mathrm{T}}$. Usually, we consider the case where $k<\text{rank}(\mathbf{W})$ since when $k\geq\text{rank}(\mathbf{W})$, the Nyström approximation is exact [54]. Therefore, all elements of $\boldsymbol{\Sigma}_{\mathbf{W},k}$ are positive. Now, we have

$$\widetilde{\mathbf{K}}=\underbrace{\mathbf{C}\mathbf{U}_{\mathbf{W},k}\sqrt{\boldsymbol{\Sigma}_{\mathbf{W},k}^{+}}}_{\mathbf{V}}\underbrace{\left(\mathbf{C}\mathbf{U}_{\mathbf{W},k}\sqrt{\boldsymbol{\Sigma}_{\mathbf{W},k}^{+}}\right)^{\mathrm{T}}}_{\mathbf{V}^{\mathrm{T}}}.$$

We adopt $\widetilde{\mathbf{K}}+\mu l\mathbf{I}$ as an approximation of $\mathbf{K}+\mu l\mathbf{I}$. Since $\widetilde{\mathbf{K}}$ is positive semidefinite, the invertibility of $\widetilde{\mathbf{K}}+\mu l\mathbf{I}$ is guaranteed. Using the Woodbury formula, we can obtain

$$(\mu l\mathbf{I}+\widetilde{\mathbf{K}})^{-1}=\frac{1}{\mu l}(\mathbf{I}-\mathbf{V}(\mu l\mathbf{I}_k+\mathbf{V}^{\mathrm{T}}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}) \tag{5}$$

where $\mathbf{I}_k$ is a $k\times k$ identity matrix. We let $\boldsymbol{u}=(\mu l\mathbf{I}+\widetilde{\mathbf{K}})^{-1}\boldsymbol{y}$. According to (5), we have $\boldsymbol{u}=(1/\mu l)(\boldsymbol{y}-\mathbf{V}(\mu l\mathbf{I}_k+\mathbf{V}^{\mathrm{T}}\mathbf{V})^{-1}\mathbf{V}^{\mathrm{T}}\boldsymbol{y})$. We introduce a temporary variable $\boldsymbol{t}:\ (\mu l\mathbf{I}_k+\mathbf{V}^{\mathrm{T}}\mathbf{V})\boldsymbol{t}=\mathbf{V}^{\mathrm{T}}\boldsymbol{y}$, $\boldsymbol{u}=(1/\mu l)(\boldsymbol{y}-\mathbf{V}\boldsymbol{t})$. We present a Nyström approximate kernel selection algorithm (see Algorithm 3) for $\mathcal{C}_{\text{ree}}$.

Theorem 7: The computational complexity of Algorithm 3 is $O(N(c^3+lc\max\{d,k\}))$.

Proof: The complexity of step 4 is $O(lcd)$, where d is the dimension of the input data. The complexity of step 5 is $O(c^3)$ since this step conducts SVD. Step 6 has a complexity of $O(lck)$. In step 7, we solve the inverse of $(\mu l\mathbf{I}_k+\mathbf{V}^{\mathrm{T}}\mathbf{V})$ with the complexity $O(k^3)$. Computing the matrix of the linear system takes $O(lk^2)$ multiplications. Therefore, the total complexity of step 7 is $O(lk^2)$. The complexity of step 8 is $O(lk)$. Since $k<c\ll l$ and the number of candidate kernels is N, the total time complexity of Algorithm 3 is $O(N(c^3+lc\max\{d,k\}))$. ■

Algorithm 3 Nyström Approximate Kernel Selection

Input: $\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{l}$, $\mathcal{K}=\{\kappa_1,\ldots,\kappa_N\}$, c, k, μ;
Output: The optimal kernel κ^*;
1: **Initialize**: $\mathcal{C}^*=\infty$;
2: **for** each $\kappa\in\mathcal{K}$ **do**
3: Sample c indices from $\{1,\ldots,l\}$ to form the index set $\mathcal{I}$;
4: Generate $\mathbf{C}$ and $\mathbf{W}$ using $\mathcal{S}$ and κ according to $\mathcal{I}$;
5: Calculate the SVD of $\mathbf{W}$ as $\mathbf{W}=\mathbf{U}_{\mathbf{W}}\boldsymbol{\Sigma}_{\mathbf{W}}\mathbf{U}_{\mathbf{W}}^{\mathrm{T}}$;
6: Let $\mathbf{V}=\mathbf{C}\mathbf{U}_{\mathbf{W},k}\sqrt{\boldsymbol{\Sigma}_{\mathbf{W},k}^{+}}$;
7: Solve $\left(\mu l\mathbf{I}_k+\mathbf{V}^{\mathrm{T}}\mathbf{V}\right)\boldsymbol{t}=\mathbf{V}^{\mathrm{T}}\boldsymbol{y}$ to obtain $\boldsymbol{t}$;
8: $\boldsymbol{u}=\dfrac{1}{\mu l}(\boldsymbol{y}-\mathbf{V}\boldsymbol{t})$;
9: $\mathcal{C}_{\text{ree}}(\widetilde{\mathbf{K}})=\mu\boldsymbol{y}^{\mathrm{T}}\boldsymbol{u}$;
10: **if** $\mathcal{C}_{\text{ree}}(\widetilde{\mathbf{K}})\leq\mathcal{C}^*$ **then**
11: $\mathcal{C}^*=\mathcal{C}_{\text{ree}}(\widetilde{\mathbf{K}})$;
12: $\kappa^*=\kappa$;
13: **Return** κ^*;

We further discuss the approximate computation of the kernel selection criterion $\mathcal{C}_{\text{ipe}}$. The computation of the bias term of $\mathcal{C}_{\text{ipe}}$ is similar to that of $\mathcal{C}_{\text{ree}}$. For the variance term of $\mathcal{C}_{\text{ipe}}$, we need the sum of the eigenvalues of $\mathbf{K}^2\mathbf{K}_{\mu}^{-2}$. Actually, the Nyström approximate matrix $\widetilde{\mathbf{K}}$ corresponds to an approximate eigendecomposition of $\mathbf{K}$ [28]

$$\widetilde{\mathbf{K}}=\underbrace{\sqrt{\frac{c}{l}}\mathbf{C}\mathbf{U}_{\mathbf{W},k}\boldsymbol{\Sigma}_{\mathbf{W},k}^{+}}_{\widetilde{\mathbf{U}}}\underbrace{\frac{l}{c}\boldsymbol{\Sigma}_{\mathbf{W},k}}_{\widetilde{\boldsymbol{\Sigma}}}\underbrace{\left(\sqrt{\frac{c}{l}}\mathbf{C}\mathbf{U}_{\mathbf{W},k}\boldsymbol{\Sigma}_{\mathbf{W},k}^{+}\right)^{\mathrm{T}}}_{\widetilde{\mathbf{U}}^{\mathrm{T}}}.$$

We use $\widetilde{\boldsymbol{\Sigma}}=(l/c)\boldsymbol{\Sigma}_{\mathbf{W},k}=(l/c)\text{diag}(\tilde{\lambda}_1,\tilde{\lambda}_2,\ldots,\tilde{\lambda}_k)$ as the eigenvalues of $\widetilde{\mathbf{K}}$ to approximate the eigenvalues of $\mathbf{K}$, where $\tilde{\lambda}_1,\tilde{\lambda}_2,\ldots,\tilde{\lambda}_k$ are the top-k eigenvalues of $\mathbf{W}$. Therefore

$$\text{trace}(\mathbf{K}^2\mathbf{K}_{\mu}^{-2})\approx\sum_{i=1}^{k}\left(\frac{(l/c)\tilde{\lambda}_i}{(l/c)\tilde{\lambda}_i+\mu l}\right)^2=\sum_{i=1}^{k}\left(\frac{\tilde{\lambda}_i}{\tilde{\lambda}_i+\mu c}\right)^2.$$

According to the above-mentioned analysis, we can obtain the approximate kernel selection algorithm for the criterion $\mathcal{C}_{\text{ipe}}$, which is similar to Algorithm 3 and omitted here. Its computational complexity is $O(N(c^3+lc\max\{d,k\}))$.

VII. Experiments

In this section, we empirically verified the theoretical findings of the approximate consistency and evaluated the effectiveness of the proposed approximate algorithms.

We conducted experiments on benchmark data sets that are publicly available from the UCI Repository, StatLib data sets, and Weka data sets for both regression and classification problems. We randomly split each data set into training and test sets (50% of all samples for training and the other 50% for testing). We adopted $\kappa(\boldsymbol{x},\boldsymbol{x}')=\exp(-\gamma\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2)$ with a variable width γ as the kernel set $\mathcal{K}$. Since the focus of

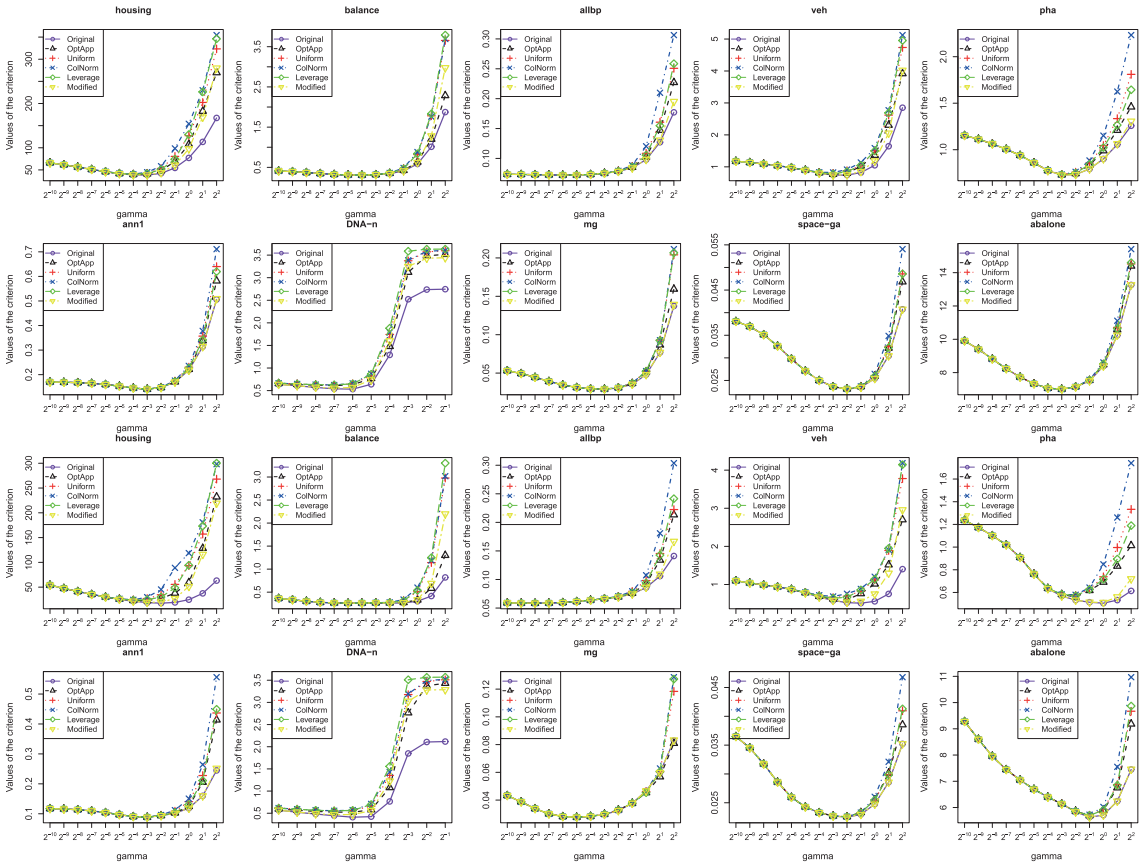


Fig. 1. Approximate consistency of different kernel matrix approximation algorithms under $\mathcal{C}_{\text{ree}}(\mathbf{K})$ and $\mathcal{C}_{\text{ipe}}(\mathbf{K})$ for regression data.

this work was not on tuning the regularization parameter μ, we just set it as a fixed value of 0.005. We set the parameter σ in $\mathcal{C}_{\text{ipe}}$ to be 1% of the standard deviation of the response vector $\boldsymbol{y}$. All the implementations were in the R language.

In the first experiment, we evaluated the theoretical findings on the approximate consistency. We compared six kernel matrix approximation algorithms under $\mathcal{C}_{\text{ree}}$ and $\mathcal{C}_{\text{ipe}}$, including the optimal-rank k approximation derived with SVD (OptApp), the Nyström approximation with uniform sampling (Uniform) [30], the Nyström approximation with column-norm sampling (ColNorm) [29], the Nyström approximation with leverage-score sampling (Leverage) [32],[3] the modified Nyström approximation (Modified)[4] [34], and MCM approximation (MCM) [26]. We set $k = 20$ and $c = 0.2l$. To avoid randomness, all experiments for the compared Nyström methods were repeated 20 times. For MCM approximation, a three-level circulant matrix was adopted. Based on H4 in Theorem 1, we tuned h to minimize the Frobenius norm of the difference between $\mathbf{X}_{\boldsymbol{m}} = [\|\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{x}_{\boldsymbol{j}}\| : \boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]]$ and $\mathbf{H}_{\boldsymbol{m}} = [h\|\boldsymbol{i} - \boldsymbol{j}\| : \boldsymbol{i}, \boldsymbol{j} \in [\boldsymbol{m}]]$.

We generated synthetic data following the settings in [25].[5] The target function [25] was $f(\boldsymbol{x}) = \mathrm{e}^{-8(3-\|\boldsymbol{x}\|_2)^2} - \mathrm{e}^{-8(1.5-\|\boldsymbol{x}\|_2)^2} - \mathrm{e}^{-8(2-\|\boldsymbol{x}\|_2)^2}$. We used $\{(\boldsymbol{x}_{\boldsymbol{j}}, y_{\boldsymbol{j}}), \boldsymbol{j} \in [\boldsymbol{m}]\} \in \mathbb{R}^2 \times \mathbb{R}$ for $\boldsymbol{m} = (10, 10), (20, 20), (30, 30), (40, 40)$ as data points. Each dimension of the sampled inputs $\boldsymbol{x}_{\boldsymbol{j}}$ was centered at 0. For two successive points, there was a fixed difference of 0.1 between them, and $y_{\boldsymbol{j}} = f(\boldsymbol{x}_{\boldsymbol{j}}) + \xi$, where ξ was a Gaussian random variable with mean 0 and standard deviation 0.01.

For each kernel parameter γ, we observed the values of the original criterion $\mathcal{C}(\mathbf{K})$ and the approximate criterion $\mathcal{C}(\widetilde{\mathbf{K}})$. The results of $\mathcal{C}_{\text{ree}}$ and $\mathcal{C}_{\text{ipe}}$ for regression were shown in Fig. 1. The results for the classification were shown in Fig. 2. We found that the curves of the original criterion and the approximate criteria were close for most data sets. The curves of OptApp and Modified were closer to the curve of the original criterion than the Nyström approximations. The results on the synthetic data were also provided (see Fig. 3). We observed that the more samples, the closer the curves of the original criteria and approximate ones were.

In the second experiment, we evaluated the performance of the optimal kernels selected by the proposed approximate algorithms in Section VI. We determined the effectiveness to assess the performance. Effectiveness includes efficiency and

[3]Although in the theoretical analysis, we only considered the Nyström approximation with $(1 + \epsilon)$ relative-error bound, in experiments, for the purpose of comparison, we also considered the Nyström approximation with other sampling distributions.

[4]For the modified Nyström approximation, we used the uniform+adpative^2 sampling [35].

[5]To strictly satisfy H4 in Theorem 1, experiments for MCM approximation were only conducted on synthetic data in the first experiment.

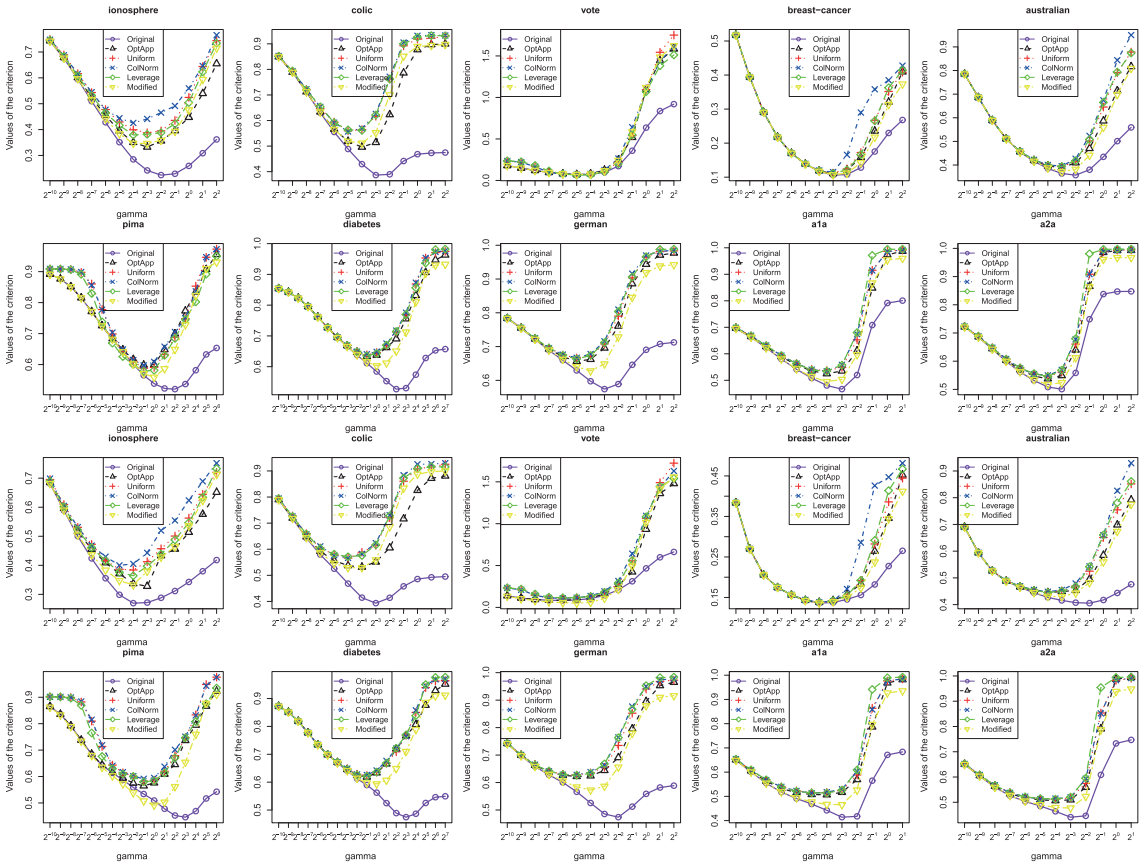


Fig. 2. Approximate consistency of different kernel matrix approximation algorithms under $\mathcal{C}_{\text{ree}}(\mathbf{K})$ and $\mathcal{C}_{\text{ipe}}(\mathbf{K})$ for classification data.

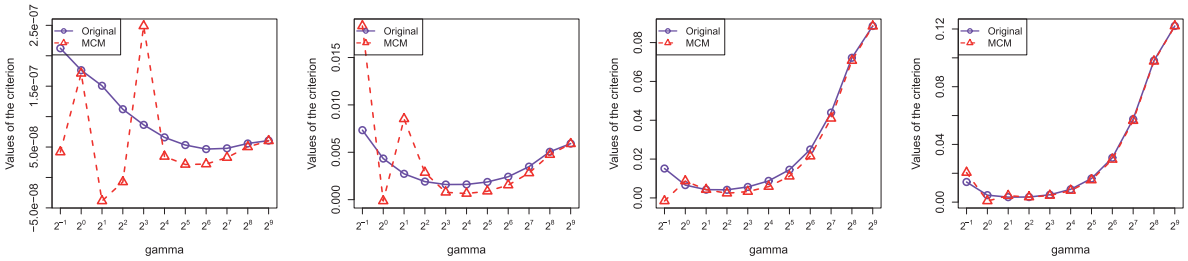


Fig. 3. Evolution of the approximate consistency of the MCM approximation as the number of samples $\Pi_{\boldsymbol{m}}$ increases. $\Pi_{\boldsymbol{m}}$ increases from 100 to 1600. We can find that the more the samples, the closer the curves of "MCM" and "Original." (a) 100. (b) 400. (c) 900. (d) 1600.

generalization, where the former is measured by averaging the computational time for kernel selection and the latter is measured by the mean testing error of the trained model with the kernel selected by the kernel selection algorithm. For the regression and classification problems, we used KRR and least squares support vector machine (LSSVM) as the base models, respectively.

In the first step of the experiment, we selected the optimal kernel from the candidate kernel set $\mathcal{K}$ by minimizing the accurate or approximate kernel selection criterion on the training set. Then, we trained the model using the selected optimal kernel, again on the training set. Finally, we evaluated the test performance of the trained model on the test set. The test performance was evaluated in terms of the mean square error (mse) for the regression problems and the MCA for the classification tasks. The experimental results for the criteria $\mathcal{C}_{\text{ree}}$ and $\mathcal{C}_{\text{ipe}}$ were shown in Tables I and II, respectively. "Accurate" refers to the optimal kernel selected by minimizing the accurate kernel selection criterion using the original kernel matrix. "MCM" and "Nyström" denote the MCM kernel selection algorithm and the Nyström kernel selection algorithm, respectively.

According to the Wilcoxon signed rank test [55], "Accurate" was statistically superior to neither "MCM" nor "Nyström" on $\mathcal{C}_{\text{ree}}$ and $\mathcal{C}_{\text{ipe}}$ at the 95% level of significance. Meanwhile, Tables I and II also showed that approximate algorithms were faster than the accurate ones for larger data sets, especially the "MCM" approximate kernel selection. The efficiency gain

TABLE I

COMPARISON OF THE MSES OR THE MEAN CLASSIFICATION ACCURACIES (MCAS) AND THE TIME (SECONDS) BETWEEN THE ACCURATE AND THE PROPOSED APPROXIMATE KERNEL SELECTION ALGORITHMS FOR $\mathcal{C}_{\text{ree}}$

regression (♯ samples)	Accurate		MCM		Nyström	
	MSE	Time (sec)	MSE	Time (sec)	MSE	Time (sec)
housing (506)	2.79±0.48(1e1)	4.48(1e-3)	3.06±0.50(1e1)	7.44(1e-3)	2.80±0.47(1e1)	2.27(1e-3)
balance (625)	2.49±0.14(1e-1)	4.20(1e-3)	2.49±0.13(1e-1)	7.16(1e-3)	2.48±0.14(1e-1)	2.22(1e-3)
allbp (840)	5.45±0.83(1e-2)	1.28(1e-2)	5.44±0.82(1e-2)	1.36(1e-2)	5.50±0.79(1e-2)	0.51(1e-2)
veh (846)	5.67±0.35(1e-1)	1.44(1e-2)	5.99±0.45(1e-1)	1.15(1e-2)	5.73±0.32(1e-1)	0.57(1e-2)
pha (1070)	6.13±0.38(1e-1)	4.08(1e-2)	6.15±0.67(1e-1)	1.56(1e-2)	6.04±0.35(1e-1)	1.28(1e-2)
ann1 (1131)	9.39±1.05(1e-2)	4.10(1e-2)	1.20±0.19(1e-1)	1.60(1e-2)	9.39±0.41(1e-2)	1.32(1e-2)
DNA-n (1275)	4.95±0.13(1e-1)	4.79(1e-2)	5.19±0.12(1e-1)	1.53(1e-2)	4.96±0.12(1e-1)	1.38(1e-2)
mg (1385)	2.11±0.08(1e-2)	3.61(1e-2)	2.06±0.07(1e-2)	2.26(1e-2)	1.98±0.12(1e-2)	1.21(1e-2)
space-ga (3107)	2.05±0.15(1e-2)	4.85(1e-1)	2.00±0.17(1e-2)	4.79(1e-2)	1.98±0.15(1e-2)	1.08(1e-1)
abalone (4177)	6.06±0.30(1e0)	9.70(1e-1)	5.67±0.31(1e0)	6.15(1e-2)	5.75±0.38(1e0)	2.05(1e-1)
classification (♯ samples)	MCA	Time (sec)	MCA	Time (sec)	MCA	Time (sec)
ionosphere (351)	93.65%	1.46(1e-3)	93.40%	4.40(1e-3)	93.38%	8.68(1e-4)
colic (368)	75.55%	1.39(1e-3)	76.70%	4.17(1e-3)	78.74%	8.82(1e-4)
vote (435)	94.79%	1.47(1e-3)	93.76%	4.09(1e-3)	94.72%	8.86(1e-4)
breast-cancer (683)	96.86%	4.31(1e-3)	96.62%	7.60(1e-3)	96.86%	2.22(1e-3)
australian (690)	86.34%	1.37(1e-2)	85.15%	1.23(1e-2)	86.38%	5.62(1e-3)
pima (768)	73.11%	1.35(1e-2)	76.75%	1.16(1e-2)	76.22%	5.62(1e-3)
diabetes (768)	73.05%	1.44(1e-2)	73.28%	1.20(1e-2)	76.29%	5.67(1e-3)
german (1000)	73.94%	1.48(1e-2)	75.23%	1.14(1e-2)	75.19%	5.65(1e-3)
a1a (1605)	81.86%	1.25(1e-1)	80.03%	2.42(1e-2)	82.23%	2.86(1e-2)
a2a (2265)	81.54%	2.73(1e-1)	80.75%	3.29(1e-2)	81.72%	5.78(1e-2)

TABLE II

COMPARISON OF THE MSES OR THE MCAS AND THE TIME (SECONDS) BETWEEN THE ACCURATE AND THE PROPOSED APPROXIMATE KERNEL SELECTION ALGORITHMS FOR $\mathcal{C}_{\text{ipe}}$

regression (♯ samples)	Accurate		MCM		Nyström	
	MSE	Time (sec)	MSE	Time (sec)	MSE	Time (sec)
housing (506)	3.10±0.52(1e1)	2.80(1e-2)	3.21±0.46(1e1)	7.89(1e-3)	2.87±0.42(1e1)	2.29(1e-3)
balance (625)	2.79±0.45(1e-1)	2.78(1e-2)	2.46±0.12(1e-1)	7.37(1e-3)	2.46±0.12(1e-1)	2.25(1e-3)
allbp (840)	6.01±0.82(1e-2)	1.05(1e-1)	5.59±0.90(1e-2)	1.39(1e-2)	5.48±0.89(1e-2)	5.27(1e-3)
veh (846)	6.22±0.52(1e-1)	1.06(1e-1)	6.02±0.67(1e-1)	1.38(1e-2)	5.84±0.43(1e-1)	5.39(1e-3)
pha (1070)	6.28±0.60(1e-1)	3.85(1e-1)	6.40±0.43(1e-1)	1.70(1e-2)	6.31±0.43(1e-1)	1.70(1e-2)
ann1 (1131)	9.09±0.73(1e-2)	3.85(1e-1)	9.35±1.06(1e-2)	1.69(1e-2)	8.79±0.87(1e-2)	1.25(1e-2)
DNA-n (1275)	5.67±0.25(1e-1)	3.81(1e-1)	4.96±0.15(1e-1)	1.62(1e-2)	4.96±0.14(1e-1)	1.33(1e-2)
mg (1385)	1.91±0.08(1e-2)	3.85(1e-1)	1.95±0.08(1e-2)	1.65(1e-2)	1.91±0.08(1e-2)	1.28(1e-2)
space-ga (3107)	1.92±0.22(1e-2)	6.15(1e0)	1.93±0.21(1e-2)	4.81(1e-2)	1.93±0.22(1e-2)	1.09(1e-1)
abalone (4177)	5.57±0.19(1e0)	1.48(1e1)	5.96±0.18(1e0)	6.32(1e-2)	5.57±0.19(1e0)	2.06(1e-1)
classification (♯ samples)	MCA	Time (sec)	MCA	Time (sec)	MCA	Time (sec)
ionosphere (351)	93.42%	6.30(1e-3)	93.31%	4.44(1e-3)	93.65%	9.00(1e-4)
colic (368)	74.65%	6.22(1e-3)	77.63%	4.18(1e-3)	79.87%	9.06(1e-4)
vote (435)	93.69%	2.81(1e-2)	92.67%	7.30(1e-3)	94.70%	2.27(1e-3)
breast-cancer (683)	96.54%	2.77(1e-2)	96.89%	7.70(1e-3)	96.79%	2.24(1e-3)
australian (690)	84.49%	1.06(1e-1)	83.55%	1.19(1e-2)	86.48%	5.70(1e-3)
pima (768)	68.01%	1.06(1e-1)	75.27%	1.13(1e-2)	76.04%	5.75(1e-3)
diabetes (768)	68.34%	1.10(1e-1)	76.71%	1.13(1e-2)	76.29%	5.62(1e-3)
german (1000)	71.28%	1.10(1e-1)	73.79%	1.12(1e-2)	74.31%	5.60(1e-3)
a1a (1605)	79.94%	1.06(1e0)	82.32%	2.42(1e-2)	82.40%	2.85(1e-2)
a2a (2265)	80.13%	2.65(1e0)	81.30%	3.31(1e-2)	81.26%	5.85(1e-2)

for $\mathcal{C}_{\text{ipe}}$ is more obvious due to its higher computational complexity. In summary, the proposed approximate algorithms can reduce the computational cost of kernel selection, while, at the same time, exhibiting a competitive performance.

VIII. CONCLUSION

In this article, we proposed a novel approach for kernel selection based on kernel matrix approximation. We theoretically justified the introduction of kernel matrix approximation into kernel selection by defining and analyzing the approximate consistency, which provides the foundations for approximate kernel selection. Under the theoretical guarantee of the approximate consistency, we designed approximate algorithms for kernel selection by exploiting the computational virtues of the MCM and Nyström approximation, whose computational complexities are quasi-linear or linear in the number of samples and significantly lower than the accurate approaches. The approximate consistency of different kernel matrix approximation algorithms under two error-minimization criteria was empirically verified, and the results showed that even for a not-so-large number of samples, the phenomenon of the consistency between the approximate and accurate criteria was present. Furthermore, the effectiveness of the experiments demonstrated that the approximate algorithms can improve

the kernel selection efficiency compared to the accurate algorithms, without sacrificing predictive performance.

REFERENCES

[1] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc.*, vol. 39, no. 1, pp. 1–49, 2002.

[2] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA, USA: MIT Press, 2004.

[3] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[4] X. Tian, Y. Li, T. Liu, X. Wang, and D. Tao, "Eigenfunction-based multitask learning in a reproducing kernel Hilbert space," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1818–1830, Jun. 2019.

[5] B. Nguyen and B. De Baets, "Kernel-based distance metric learning for supervised k-means clustering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3084–3095, Oct. 2019.

[6] F. Liu, X. Huang, C. Gong, J. Yang, and J. A. K. Suykens, "Indefinite kernel logistic regression with concave-inexact-convex procedure," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 765–776, Mar. 2019.

[7] L. Ding *et al.*, "Linear kernel tests via empirical likelihood for high-dimensional data," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 3454–3461.

[8] L. Ding *et al.*, "Two generator game: Learning to sample via linear goodness-of-fit test," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019, pp. 11257–11268.

[9] B. Schölkopf and A. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.

[10] S. Zhou, "Sparse LSSVM in primal using Cholesky factorization for large-scale problems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 783–795, Apr. 2016.

[11] Y. Xu, Z. Yang, and X. Pan, "A novel twin support-vector machine with pinball loss," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 2, pp. 359–370, Feb. 2017.

[12] Y. Liu, S. Jiang, and S. Liao, "Efficient approximation of cross-validation for kernel methods using Bouligand influence function," in *Proc. 31st Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 324–332.

[13] C. A. Micchelli and M. Pontil, "Learning the kernel function via regularization," *J. Mach. Learn. Res.*, vol. 6, pp. 1099–1125, Jul. 2005.

[14] Y. Liu, H. Lin, L. Ding, W. Wang, and S. Liao, "Fast cross-validation," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 2497–2503.

[15] Y. Liu, S. Liao, S. Jiang, L. Ding, H. Lin, and W. Wang, "Fast cross-validation for kernel-based algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[16] P. L. Bartlett, S. Boucheron, and G. Lugosi, "Model selection and error estimation," *Mach. Learn.*, vol. 48, nos. 1–3, pp. 85–113, 2002.

[17] D. Anguita, A. Ghio, L. Oneto, and S. Ridella, "In-sample and out-of-sample model selection and error estimation for support vector machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 9, pp. 1390–1406, Sep. 2012.

[18] C. Cortes, M. Kloft, and M. Mohri, "Learning kernels using local Rademacher complexity," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2760–2768.

[19] J. Li, Y. Liu, R. Yin, H. Zhang, L. Ding, and W. Wang, "Multi-class learning: From theory to algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1591–1600.

[20] L. Ding and S. Liao, "Model selection with the covering number of the ball of RKHS," in *Proc. 23rd ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2014, pp. 1159–1168.

[21] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, nos. 1–3, pp. 131–159, 2002.

[22] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *J. Mach. Learn. Res.*, vol. 11, pp. 2079–2107, Jul. 2010.

[23] L. Ding, S. Liao, Y. Liu, P. Yang, and X. Gao, "Randomized kernel selection with spectra of multilevel circulant matrices," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 2910–2917.

[24] G. H. Song and Y. S. Xu, "Approximation of high-dimensional kernel matrices by multilevel circulant matrices," *J. Complex.*, vol. 26, no. 4, pp. 375–405, 2010.

[25] G. H. Song, "Approximation of kernel matrices in machine learning," Ph.D. dissertation, Dept. Math., Syracuse Univ., Syracuse, NY, USA, 2010.

[26] L. Ding and S. Liao, "Approximate model selection for large scale LSSVM," *J. Mach. Learn. Res. Proc. Track*, vol. 20, pp. 165–180, Nov. 2011.

[27] R. Yin, Y. Liu, W. Wang, and D. Meng, "Sketch kernel ridge regression using circulant matrix: Algorithm and theory," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2019.2944959.

[28] C. K. I. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2001, pp. 682–688.

[29] P. Drineas and M. W. Mahoney, "On the Nyström method for approximating a Gram matrix for improved kernel-based learning," *J. Mach. Learn. Res.*, vol. 6, pp. 2153–2175, Dec. 2005.

[30] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling methods for the Nyström method," *J. Mach. Learn. Res.*, vol. 13, pp. 981–1006, Apr. 2012.

[31] K. Zhang and J. T. Kwok, "Clustered Nyström method for large scale manifold learning and dimension reduction," *IEEE Trans. Neural Netw.*, vol. 21, no. 10, pp. 1576–1587, Oct. 2010.

[32] A. Gittens and M. W. Mahoney, "Revisiting the Nyström method for improved large-scale machine learning," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 567–575.

[33] L. Lan, Z. Wang, S. Zhe, W. Cheng, J. Wang, and K. Zhang, "Scaling up kernel SVM on limited resources: A low-rank linearization approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 369–378, Feb. 2019.

[34] S. Wang and Z. Zhang, "Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling," *J. Mach. Learn. Res.*, vol. 14, pp. 2729–2769, Sep. 2013.

[35] S. Wang and Z. Zhang, "Efficient algorithms and error analysis for the modified Nyström method," in *Proc. 17th Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2014, pp. 996–1004.

[36] X. Chang, Y. Zhong, Y. Wang, and S. Lin, "Unified low-rank matrix estimate via penalized matrix least squares approximation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 474–485, Feb. 2019.

[37] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error CUR matrix decompositions," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 844–881, 2008.

[38] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 3, pp. 697–702, 2009.

[39] C. Cortes, M. Mohri, and A. Talwalkar, "On the impact of kernel approximation on learning accuracy," in *Proc. 13th Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2010, pp. 113–120.

[40] T. B. Yang, Y. F. Li, M. Mahdavi, R. Jin, and Z. H. Zhou, "Nyström method vs random Fourier features: A theoretical and empirical comparison," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2012, pp. 1060–1068.

[41] R. Jin, T. Yang, M. Mahdavi, Y.-F. Li, and Z.-H. Zhou, "Improved bounds for the Nyström method with application to kernel classification," *IEEE Trans. Inf. Theory*, vol. 5, no. 10, pp. 6939–6949, Oct. 2013.

[42] L. Ding and S. Liao, "Approximate consistency: Towards foundations of approximate kernel selection," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discovery Database (ECML PKDD)*, 2014, pp. 354–369.

[43] L. Ding *et al.*, "Approximate kernel selection with strong approximate consistency," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 3462–3469.

[44] L. Ding and S. Liao, "Nyström approximate model selection for LSSVM," in *Proc. 16th Pacific–Asia Conf. Adv. Knowl. Discovery Data Mining (PAKDD)*, 2012, pp. 282–293.

[45] L. Ding and S. Liao, "An approximate approach to automatic kernel selection," *IEEE Trans. Cybern.*, vol. 47, no. 3, pp. 554–565, Mar. 2017.

[46] Y. Xu and H. Zhang, "Refinement of reproducing kernels," *J. Mach. Learn. Res.*, vol. 10, pp. 107–140, Jan. 2009.

[47] G. S. Kimeldorf and G. Wahba, "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines," *Ann. Math. Statist.*, vol. 41, no. 2, pp. 495–502, 1970.

[48] P. J. Davis, *Circulant Matrices*. New York, NY, USA: Wiley, 1979.

[49] E. E. Tyrtyshnikov, "A unifying approach to some old and new theorems on distribution and clustering," *Linear Algebra Appl.*, vol. 232, pp. 1–43, Jan. 1996.

[50] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 1–50, 2000.

[51] C. Boutsidis, P. Drineas, and M. Magdon-Ismail, "Near optimal column-based matrix reconstruction," in *Proc. IEEE 52nd Annu. Symp. Found. Comput. Sci. (FOCS)*, 2011, pp. 305–314.

[52] A. Deshpande, L. Rademacher, S. Vempala, and G. Wang, "Matrix approximation and projective clustering via volume sampling," in *Proc. 17th Annu. ACM-SIAM Symp. Discrete Algorithm (SODA)*, 2006, pp. 1117–1126.
[53] F. Bach, "Sharp analysis of low-rank kernel matrix approximations," in *Proc. 26th Annu. Conf. Learn. Theory (COLT)*, 2013, pp. 185–209.
[54] S. Kumar, M. Mohri, and A. Talwalkar, "On sampling-based approximate spectral decomposition," in *Proc. 26th Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 553–560.
[55] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, Jan. 2006.

Lizhong Ding received the B.E. degree in computer science and technology and the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2009 and 2015, respectively.

He is currently a Research Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. His research interests include model selection, kernel methods, and deep generative models.

Shizhong Liao received the B.S. degree in computer software from the Dalian University of Technology, Dalian, China, in 1985, the M.S. degree in computer science from Jilin University, Changchun, China, in 1988, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 1997.

He is currently a Professor with the School of Computer Science and Technology, Tianjin University, Tianjin, China. His research interests include artificial intelligence and theoretical computer science.

Yong Liu received the Ph.D. degree in computer science from Tianjin University, Tianjin, China, in 2016.

He is currently an Assistant Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. His research interests include large-scale kernel methods, large-scale model selection, and machine learning.

Li Liu received the B.Eng. degree in electronic information engineering from Xi'an Jiaotong University, Xi'an, China, in 2011, and the Ph.D. degree from the Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield, U.K., in 2014.

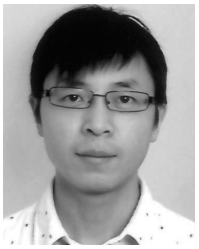
He is currently with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. His current research interests include computer vision, machine learning, and data mining.

Fan Zhu received the M.Sc. degree (Hons.) in electrical engineering and the Ph.D. degree in computer vision from The University of Sheffield, Sheffield, U.K., in 2011 and 2015, respectively.

He was a Post-Doctoral Research Fellow with the Electrical and Computer Engineering Department, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates, and a Data Scientist with Pegasus LLC, Palos Verdes Estates, CA, USA. He is currently with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi. His research interests include deep feature learning for 2-D images and 3-D shapes, scene understanding, video analytics, and adversarial learning.

Yazhou Yao received the B.Sc. and M.Sc. degrees from Nanjing Normal University, Nanjing, China, in 2010 and 2013, respectively, and the Ph.D. degree in computer science from the University of Technology Sydney, Ultimo, NSW, Australia, in 2018, with the support from the China Scholarship Council.

He is a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing. His research interests include social multimedia processing and machine learning.



Ling Shao is currently the CEO and the Chief Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. He is also a Professor with the School of Computing Sciences, University of East Anglia, Norwich, U.K.

Dr. Shao is also an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and several other journals.

Xin Gao received the B.S. degree in computer science from Tsinghua University, Beijing, China, in 2004, and the Ph.D. degree in computer science from University of Waterloo, Waterloo, ON, Canada, in 2009.

He is currently an Associate Professor of computer science with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. He is also a PI with the Computational Bioscience Research Center, KAUST, and an Adjunct Faculty Member with the David R. Cheriton School of Computer Science, University of Waterloo. His group focuses on building computational models, developing machine learning methods, and designing efficient and effective algorithms, with particular a focus on applications to key open problems in biology. He has coauthored more than 100 research articles in the fields of machine learning and bioinformatics.