Statistica Sinica Preprint No: SS-2016-0369	
Title	Nonlinear Regression Estimation Using Subset-Based Kernel Principal Components
Manuscript ID	SS-2016-0369
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0369
Complete List of Authors	Yuan Ke Degui Li and Qiwei Yao
Corresponding Author	Degui Li
E-mail	degui.li@york.ac.uk

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

NONLINEAR REGRESSION ESTIMATION USING SUBSET-BASED KERNEL PRINCIPAL COMPONENTS

Yuan Ke[1], Degui Li[2], Qiwei Yao[3]

[1]*Penn State University,* [2]*The University of York,* [3]*London School of Economics*

Abstract: We study the estimation of conditional mean regression functions through the so-called subset-based kernel principal component analysis (KPCA). Instead of using one global kernel feature space, we project a target function into different localized kernel feature spaces at different parts of the sample space. Each localized kernel feature space reflects the relationship on a subset between the response and covariates more parsimoniously. When the observations are collected from a strictly stationary and weakly dependent process, the orthonormal eigenfunctions which span the kernel feature space are consistently estimated by implementing an eigenanalysis on the subset-based kernel Gram matrix, and the estimated eigenfunctions are then used to

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

construct the estimation of the mean regression function. Under some regularity conditions, the developed estimator is shown to be uniformly consistent over the subset with a convergence rate faster than those of some well-known nonparametric estimation methods. In addition, we discuss some generalizations of the KPCA approach, and consider using the same subset-based KPCA approach to estimate the conditional distribution function. The numerical studies including three simulated examples and two data sets illustrate the reliable performance of the proposed method. In particular, the improvement over the global KPCA method is evident.

Key words and phrases: Conditional distribution function, eigen-analysis, kernel Gram matrix, KPCA, mean regression function, nonparametric regression.

1 Introduction

Let Y be a scalar response variable and $\mathbf{X}$ be a p-dimensional random vector. We are interested in estimating the conditional mean regression function

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

defined by

$$h(\mathbf{x}) = \mathsf{E}(Y|\mathbf{X} = \mathbf{x}), \quad \mathbf{x} \in \mathcal{G}, \tag{1.1}$$

where $\mathcal{G} \subset \mathcal{R}^p$ is a measurable subset of the sample space of $\mathbf{X}$, and $\mathsf{P}(\mathbf{X} \in \mathcal{G}) > 0$. We allow that the mean regression function $h(\cdot)$ is not specified, except for certain smoothness conditions, which makes (1.1) more flexible than the traditional parametric linear and nonlinear regression. Nonparametric estimation of $h(\cdot)$ has been extensively studied in the existing literature such as Green and Silverman (1994), Wand and Jones (1995), Fan and Gijbels (1996), Fan and Yao (2003) and Teräsvirta, Tjøstheim and Granger (2010). When the dimension of random covariates p is large, a direct use of nonparametric regression estimation methods, such as the spline and kernel-based smoothing, typically perform poorly due to the so-called "curse of dimensionality". Hence, some dimension-reduction techniques/assumptions (such as the additive models, single-index models and varying-coefficient models) have to be imposed when estimating the mean regression function. However, it is well known that some dimension reduction techniques may result in systematic biases in estimation. For instance, the estimation based on an additive model may perform poorly when the additive assumption does not hold.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

In this paper we propose a data-driven dimension reduction approach through the use of a *Kernel Principal Components Analysis (KPCA)* for the random covariate $\mathbf{X}$. The KPCA is a nonlinear version of the standard linear *Principal Component Analysis (PCA)* and overcomes the limitations of the linear PCA by conducting the eigendecomposition of the kernel Gram matrix, see, for example, Schölkopf, Smola and Müller (1999), Braun (2005) and Blanchard, Bousquet and Zwald (2007). See also Section 2.2 for a detailed description on the KPCA and its relation to the standard PCA. The KPCA has been applied in, among others, feature extraction and de-noising in high-dimensional regression (Rosipal *et al.* (2001)), density estimation (Girolami (2002)), robust regression (Wibowo and Desa (2011)), conditional density estimation (Fu, Shih and Wang (2011); Izbicki and Lee (2013)), and regression estimation (Lee and Izbicki (2013)).

Unlike the existing literature on KPCA, we approximate the mean regression $h(\mathbf{x})$ on different subsets of the sample space of $\mathbf{X}$ by linear combinations of different subset-based kernel principal components. The subset-based KPCA identifies nonlinear eigenfunctions in a subset, and thus reflects the relationship between Y and $\mathbf{X}$ on that set more parsimoniously than, for example, a global KPCA (see Proposition 1 in Section 2.2). The subsets may be defined according to some characteristics of $\mathbf{X}$ and/or those

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

on the relationship between Y and $\mathbf{X}$ (e.g., MACD for financial prices, different seasons/weekdays for electricity consumption, or adaptively by some change-point detection methods) and they are not necessarily connected sets. This is a marked difference from such conventional nonparametric regression techniques as the kernel smoothing and nearest neighbour methods. Meanwhile, we assume here that the observations are collected from a strictly stationary and weakly dependent process, which relaxes the independence and identical distribution assumption in the KPCA literature and makes the proposed methodology applicable to the time series data. Under some regularity conditions, we show that the estimated eigenvalues and eigenfunctions constructed through an eigenanalysis on the subset-based kernel Gram matrix are consistent. The conditional mean regression function $h(\cdot)$ is then estimated through the projection to the kernel spectral space that is spanned by a few estimated eigenfunctions whose number is determined by a simple ratio method. The developed conditional mean estimation is shown to be uniformly consistent over the subset with a convergence rate faster than those of some well-known nonparametric estimation methods. We further extend the subset-based KPCA method to estimation of the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

conditional distribution function:

$$F_{Y|\mathbf{X}}(y|\mathbf{x}) = \mathsf{P}(Y \leqslant y|\mathbf{X} = \mathbf{x}),\ \mathbf{x} \in \mathcal{G}, \tag{1.2}$$

and establish the corresponding asymptotic property.

The rest of the paper is organized as follows. Section 2 introduces the subset-based KPCA and the estimation methodology for the mean regression function. Section 3 derives the main asymptotic theorems of the proposed estimation method. Section 4 extends the proposed subset-based KPCA for estimation of conditional distribution functions. Section 5 illustrates the finite sample performance of the proposed methods by simulation. Section 6 reports on two data applications. Section 7 concludes the paper. Proofs of the theoretical results are available in an online supplementary material.

2 Methodology

Let $\{(Y_i, \mathbf{X}_i),\ 1 \leqslant i \leqslant n\}$ be observations from a strictly stationary process with the same marginal distribution as that of $(Y, \mathbf{X})$. Our aim is to estimate the mean regression function $h(\mathbf{x})$ for $\mathbf{x} \in \mathcal{G}$, as specified in (1.1). We first introduce the kernel spectral decomposition in Section 2.1, followed by the illustration on the kernel feature space and the relationship between the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

KPCA and the standard PCA in Section 2.2, and we propose an estimation method for the conditional mean regression function in Section 2.3.

2.1 Kernel spectral decomposition

Let $\mathcal{L}_2(\mathcal{G})$ be the Hilbert space consisting of all the functions defined on $\mathcal{G}$ which satisfy that, for any $f \in \mathcal{L}_2(\mathcal{G})$,

$$\int_{\mathcal{G}} f(\mathbf{x})\mathsf{P}_{\mathbf{X}}(d\mathbf{x}) = \mathsf{E}\big[f(\mathbf{X})I(\mathbf{X} \in \mathcal{G})\big] = 0,$$

$$\int_{\mathcal{G}} f^2(\mathbf{x})\mathsf{P}_{\mathbf{X}}(d\mathbf{x}) = \mathsf{E}\big[f^2(\mathbf{X})I(\mathbf{X} \in \mathcal{G})\big] < \infty,$$

where $\mathsf{P}_{\mathbf{X}}(\cdot)$ denotes the probability measure of $\mathbf{X}$, and $I(\cdot)$ is an indicator function. The inner product on $\mathcal{L}_2(\mathcal{G})$ is defined as

$$\langle f, g\rangle = \int_{\mathcal{G}} f(\mathbf{x})g(\mathbf{x})\mathsf{P}_{\mathbf{X}}(d\mathbf{x}) = \mathrm{Cov}\left\{f(\mathbf{X})I(\mathbf{X} \in \mathcal{G}),\, g(\mathbf{X})I(\mathbf{X} \in \mathcal{G})\right\}, \quad f, g \in \mathcal{L}_2(\mathcal{G}). \tag{2.1}$$

Let $K(\cdot,\cdot)$ be a Mercer kernel defined on $\mathcal{G} \times \mathcal{G}$, s a bounded and symmetric function, and for any $\mathbf{u}_1, \cdots, \mathbf{u}_k \in \mathcal{G}$ and $k \geqslant 1$, the $k \times k$ matrix with $K(\mathbf{u}_i, \mathbf{u}_j)$ being its (i,j)-th element is non-negative definite. For any fixed $\mathbf{u} \in \mathcal{G}$, $K(\mathbf{x}, \mathbf{u})$ can be seen as a function of $\mathbf{x}$. A Mercer kernel $K(\cdot,\cdot)$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

defines an operator on $\mathcal{L}_2(\mathcal{G})$ as

$$f(\mathbf{x}) \ \rightarrow \ \int_{\mathcal{G}} K(\mathbf{x},\mathbf{u}) f(\mathbf{u}) \mathsf{P}_{\mathbf{X}}(d\mathbf{u}).$$

It follows from Mercer's Theorem (Mercer (1909)) that a Mercer kernel admits a spectral decomposition

$$K(\mathbf{u},\mathbf{v}) = \sum_{k=1}^{d} \lambda_k \varphi_k(\mathbf{u}) \varphi_k(\mathbf{v}), \quad \mathbf{u},\mathbf{v} \in \mathcal{G}, \tag{2.2}$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d > 0$ are the positive eigenvalues of $K(\cdot,\cdot)$, and $\varphi_1, \varphi_2, \cdots$ are the orthonormal eigenfunctions in the sense that

$$\int_{\mathcal{G}} K(\mathbf{x},\mathbf{u}) \varphi_k(\mathbf{u}) \mathsf{P}_{\mathbf{X}}(d\mathbf{u}) = \lambda_k \varphi_k(\mathbf{x}), \quad \mathbf{x} \in \mathcal{G}, \tag{2.3}$$

$$\langle \varphi_i, \varphi_j \rangle = \int_{\mathcal{G}} \varphi_i(\mathbf{u}) \varphi_j(\mathbf{u}) \mathsf{P}_{\mathbf{X}}(d\mathbf{u}) = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases} \tag{2.4}$$

As we can see from the spectral decomposition (2.2), $d = \max\{k : \lambda_k > 0\}$ and is possibly infinity. We say that the Mercer kernel is of finite-dimension when d is finite, and of infinite-dimension when $d = \infty$. To simplify the discussion, in this section and Section 3 below, we assume d is

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

finite. This restriction will be relaxed in Section 4. We refer to Ferreira and Menegatto (2009) for Mercer's Theorem for metric spaces. The eigenvalues λ_k and the associated eigenfunctions φ_k are usually unknown, and they need to be estimated in practice. To this end, we construct the sample eigenvalues and eigenvectors through an eigenanalysis of the kernel Gram matrix defined in (2.6) below, and then obtain the estimate of the eigenfunction φ_k by the Nyström extension (Drineas and Mahoney (2005)).

Take

$$\big\{(Y_j^{\mathcal{G}}, \mathbf{X}_j^{\mathcal{G}}), j = 1, \cdots, m\big\} = \big\{(Y_i, \mathbf{X}_i) \,\big|\, 1 \leqslant i \leqslant n,\ \mathbf{X}_i \in \mathcal{G}\big\}, \tag{2.5}$$

where m is the number of observations satisfying $\mathbf{X}_i \in \mathcal{G}$, and the subset-based kernel Gram matrix as

$$\mathbf{K}_{\mathcal{G}} = \begin{pmatrix} K(\mathbf{X}_1^{\mathcal{G}}, \mathbf{X}_1^{\mathcal{G}}) & K(\mathbf{X}_1^{\mathcal{G}}, \mathbf{X}_2^{\mathcal{G}}) & \cdots & K(\mathbf{X}_1^{\mathcal{G}}, \mathbf{X}_m^{\mathcal{G}}) \\ K(\mathbf{X}_2^{\mathcal{G}}, \mathbf{X}_1^{\mathcal{G}}) & K(\mathbf{X}_2^{\mathcal{G}}, \mathbf{X}_2^{\mathcal{G}}) & \cdots & K(\mathbf{X}_2^{\mathcal{G}}, \mathbf{X}_m^{\mathcal{G}}) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{X}_m^{\mathcal{G}}, \mathbf{X}_1^{\mathcal{G}}) & K(\mathbf{X}_m^{\mathcal{G}}, \mathbf{X}_2^{\mathcal{G}}) & \cdots & K(\mathbf{X}_m^{\mathcal{G}}, \mathbf{X}_m^{\mathcal{G}}) \end{pmatrix}. \tag{2.6}$$

Let $\widehat{\lambda}_1 \geqslant \cdots \geqslant \widehat{\lambda}_m \geqslant 0$ be the eigenvalues of $\mathbf{K}_{\mathcal{G}}$, and $\widehat{\boldsymbol{\varphi}}_1, \cdots, \widehat{\boldsymbol{\varphi}}_m$ be the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

corresponding m orthonormal eigenvectors. Write

$$\widehat{\boldsymbol{\varphi}}_k = \left[\widehat{\varphi}_k(\mathbf{X}_1^{\mathcal{G}}), \cdots, \widehat{\varphi}_k(\mathbf{X}_m^{\mathcal{G}})\right]^{\mathrm{T}}. \tag{2.7}$$

We use the so-called Nyström extension to obtain estimates of the eigenfunctions. The method was originally introduced to get the approximate numerical solution of an integral equation by replacing the integral with a representative weighted sum. The integral in (2.3) can be approximated by $\frac{1}{m}\sum_{i=1}^{m} K(\mathbf{x}, \mathbf{X}_i^{\mathcal{G}})\varphi_k(\mathbf{X}_i^{\mathcal{G}})$. Under some mild conditions (e.g., Assumption 3 in Section 3), and using the Law of Large Numbers, such an approximation is sensible. Hence, the eigenfunction $\varphi_k(x)$ can be approximated by $\frac{1}{m\lambda_k}\sum_{i=1}^{m} K(\mathbf{x}, \mathbf{X}_i^{\mathcal{G}})\varphi_k(\mathbf{X}_i^{\mathcal{G}})$. Replacing λ_k and $\varphi_k(\mathbf{X}_i^{\mathcal{G}})$ by $\widehat{\lambda}_k/m$ and $\sqrt{m}\widehat{\varphi}_k(\mathbf{X}_i^{\mathcal{G}})$, respectively, we take the Nyström extension of the eigenvector $\widehat{\boldsymbol{\varphi}}_k$ as

$$\widetilde{\varphi}_k(\mathbf{x}) = \frac{\sqrt{m}}{\widehat{\lambda}_k} \cdot \sum_{i=1}^{m} K(\mathbf{x}, \mathbf{X}_i^{\mathcal{G}})\widehat{\varphi}_k(\mathbf{X}_i^{\mathcal{G}}), \quad \mathbf{x} \in \mathcal{G}, \quad k = 1, \cdots, d. \tag{2.8}$$

Let

$$\widetilde{\lambda}_k = \widehat{\lambda}_k/m, \quad k = 1, \cdots, d. \tag{2.9}$$

Proposition 3 in Section 3 shows that, for any $\mathbf{x} \in \mathcal{G}$, $\widetilde{\lambda}_k$ and $\widetilde{\varphi}_k(\mathbf{x})$ are

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

consistent estimators of λ_k and $\varphi_k(\mathbf{x})$, respectively.

Another critical issue in applications is to estimate the dimension of the Mercer kernel $K(\cdot,\cdot)$. When the dimension of $K(\cdot,\cdot)$ is d and $d \ll m$, we can estimate d by a ratio method (Lam and Yao (2012)):

$$\widehat{d} = \underset{1\leqslant k\leqslant \lfloor mc_0\rfloor}{\arg\min}\ \widehat{\lambda}_{k+1}/\widehat{\lambda}_k = \underset{1\leqslant k\leqslant \lfloor mc_0\rfloor}{\arg\min}\ \widetilde{\lambda}_{k+1}/\widetilde{\lambda}_k, \tag{2.10}$$

where $c_0 \in (0,1)$ is a pre-specified constant, such as $c_0 = 0.5$, and $\lfloor z\rfloor$ denotes the integer part of the number z. The numerical results in Sections 5 and 6 show that this ratio method works well in finite sample cases.

2.2 Kernel feature space and KPCA

Let $\mathcal{M}(K)$ be a d-dimensional linear space spanned by the eigenfunctions $\varphi_1, \cdots, \varphi_d$, and

$$\dim\{\mathcal{M}(K)\} = d = \max\{k : \lambda_k > 0\}.$$

By the spectral decomposition (2.2), $\mathcal{M}(K)$ can also be viewed as a linear space spanned by functions $g_{\mathbf{u}}(\cdot) \equiv K(\cdot,\mathbf{u})$ for all $\mathbf{u} \in \mathcal{G}$. Thus we call $\mathcal{M}(K)$ the kernel feature space as it consists of the feature functions extracted by

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

the kernel function $K(\cdot,\cdot)$, and call $\varphi_1, \cdots, \varphi_d$ the characteristic features determined by $K(\cdot,\cdot)$ and the distribution of $\mathbf{X}$ on set $\mathcal{G}$. In addition, we call $\varphi_1(\mathbf{X}), \varphi_2(\mathbf{X}), \cdots$ the kernel principal components of $\mathbf{X}$ on set $\mathcal{G}$, and one can see they are nonlinear functions of $\mathbf{X}$ in general. We next suggest how the KPCA is connected to the standard PCA.

Any $f \in \mathcal{M}(K)$ whose mean is zero on $\mathcal{G}$ can be written as

$$f(\mathbf{x}) = \sum_{j=1}^{d} \langle f, \varphi_j \rangle \varphi_j(\mathbf{x}) \quad \text{for } \mathbf{x} \in \mathcal{G}.$$

Furthermore,

$$||f||^2 \equiv \langle f, f \rangle = \mathrm{Var}\big\{ f(\mathbf{X}) I(\mathbf{X} \in \mathcal{G}) \big\} = \sum_{j=1}^{d} \langle f, \varphi_j \rangle^2.$$

Now we introduce a generalized variance incited by the kernel function $K(\cdot,\cdot)$:

$$\mathrm{Var}_K\{ f(\mathbf{X}) I(\mathbf{X} \in \mathcal{G}) \} = \sum_{j=1}^{d} \lambda_j \langle f, \varphi_j \rangle^2, \tag{2.11}$$

where λ_j is assigned as the weight on the "direction" of φ_j for $j = 1, \cdots, d$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

Then it follows from (2.2) and (2.3) that

$$\begin{aligned}
\varphi_1 &= \arg \max_{f \in \mathcal{M}(K), ||f||=1} \int_{\mathcal{G}\times\mathcal{G}} f(\mathbf{u}) f(\mathbf{v}) K(\mathbf{u},\mathbf{v}) \mathsf{P}_{\mathbf{X}}(d\mathbf{u}) \mathsf{P}_{\mathbf{X}}(d\mathbf{v}) \\
&= \arg \max_{f \in \mathcal{M}(K), ||f||=1} \sum_{j=1}^{d} \lambda_j \langle f, \varphi_j \rangle^2 \\
&= \arg \max_{f \in \mathcal{M}(K), ||f||=1} \mathrm{Var}_K\{f(\mathbf{X}) I(\mathbf{X} \in \mathcal{G})\},
\end{aligned}$$

which indicates that the function φ_1 is the "direction" which maximizes the generalized variance $\mathrm{Var}_K\{f(\mathbf{X})I(\mathbf{X} \in \mathcal{G})\}$. Similarly it can be shown that φ_k is the solution of the above maximization problem with additional constraints $\langle \varphi_k, \varphi_j \rangle = 0$ for $1 \leqslant j < k$. Hence, the kernel principal components are the orthonormal functions in the feature space $\mathcal{M}(K)$ with the maximal kernel induced variances defined in (2.11). In other words, the kernel principal components $\varphi_1, \varphi_2, \cdots$ can be treated as "directions" while their corresponding eigenvalues $\lambda_1, \lambda_2, \cdots$ can be considered as the importance of these "directions".

A related but different approach is to view $\mathcal{M}(K)$ as a reproducing kernel Hilbert space, for which the inner product is defined differently from (2.1) to serve as a penalty in estimating functions via regularization; see Section 5.8 of Hastie, Tibshirani and Friedman (2009) and Wahba (1990). Since the reproducing property is irrelevant in our context, we adopt the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

more natural inner product (2.1). For the detailed interpretation of KPCA in a reproducing kernel space, we refer to Section 14.5.4 of Hastie, Tibshirani and Friedman (2009).

We end this subsection by stating a proposition that shows that the smaller $\mathcal{G}$, the lower the dimension of $\mathcal{M}(K)$ is. This indicates that a more parsimonious representation can be obtained by using the subset-based KPCA instead of the global KPCA. The proof of the proposition follows immediately from (2.2) and Proposition 2 below.

PROPOSITION 1. *Let $\bar{\mathcal{G}}$ be a measurable subset of the sample space of $\mathbf{X}$ such that $\mathcal{G} \subset \bar{\mathcal{G}}$, and $K(\cdot,\cdot)$ be a Mercer kernel on $\bar{\mathcal{G}} \times \bar{\mathcal{G}}$. The kernel feature spaces defined with sets $\mathcal{G}$ and $\bar{\mathcal{G}}$ are denoted, respectively by $\mathcal{M}(K)$ and $\bar{\mathcal{M}}(K)$. Then $dim\{\mathcal{M}(K)\} \leqslant dim\{\bar{\mathcal{M}}(K)\}$.*

2.3 Estimation for conditional mean regression

For the simplicity of presentation, we assume that the mean of random variate $h(\mathbf{X}) = \mathsf{E}(Y|\mathbf{X})$ on set $\mathcal{G}$ is 0,

$$\mathsf{E}\left[h(\mathbf{X})I(\mathbf{X} \in \mathcal{G})\right] = \mathsf{E}\left[\mathsf{E}(Y|\mathbf{X})I(\mathbf{X} \in \mathcal{G})\right] = \mathsf{E}\left[Y I(\mathbf{X} \in \mathcal{G})\right] = 0.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

This amounts to replacing $Y_i^{\mathcal{G}}$ by $Y_i^{\mathcal{G}} - \bar{Y}^{\mathcal{G}}$ in (2.5) with $\bar{Y}^{\mathcal{G}} = m^{-1} \sum_{1 \leqslant j \leqslant m} Y_j^{\mathcal{G}}$.

In general $\mathcal{M}(K)$ is a genuine subspace of $\mathcal{L}_2(\mathcal{G})$. Suppose that on set $\mathcal{G}$,

$$h(\mathbf{x}) = \int y f_{Y|\mathbf{X}}(y|\mathbf{x}) dy = \sum_{k=1}^{d} \beta_k \varphi_k(\mathbf{x}), \quad \mathbf{x} \in \mathcal{G}, \tag{2.12}$$

where $f_{Y|\mathbf{X}}(\cdot|\mathbf{x})$ denotes the conditional density function of Y given $\mathbf{X} = \mathbf{x}$, and

$$\beta_k = \langle \varphi_k, h \rangle = \int_{\mathbf{x} \in \mathcal{G}} \varphi_k(\mathbf{x}) \mathsf{P}_{\mathbf{X}}(d\mathbf{x}) \int y f_{Y|\mathbf{X}}(y|\mathbf{x}) dy = \mathsf{E}\left[Y \varphi_k(\mathbf{X}) \, I(\mathbf{X} \in \mathcal{G}) \right].$$

This leads to the estimator for β_k constructed as

$$\widetilde{\beta}_k = \frac{1}{m} \sum_{i=1}^{m} Y_i^{\mathcal{G}} \, \widetilde{\varphi}_k(\mathbf{X}_i^{\mathcal{G}}), \quad k = 1, \cdots, d, \tag{2.13}$$

where the $(Y_i^{\mathcal{G}}, \mathbf{X}_i^{\mathcal{G}})$, $i = 1, \cdots, m$, are defined in (2.5), and the $\widetilde{\varphi}_k(\cdot)$ are given in (2.8). Consequently the estimator for $h(\cdot)$ is taken as

$$\widetilde{h}(\mathbf{x}) = \sum_{k=1}^{d} \widetilde{\beta}_k \widetilde{\varphi}_k(\mathbf{x}), \quad \mathbf{x} \in \mathcal{G}. \tag{2.14}$$

When the dimension of the kernel $K(\cdot, \cdot)$ is unknown, the sum on the right hand side here runs from $j = 1$ to $\widehat{d}$, with $\widehat{d}$ determined via (2.10).

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

The estimator in (2.14) is derived under the assumption that on set $\mathcal{G}$, $h(\mathbf{x}) \in \mathcal{M}(K)$. When this condition is unfulfilled, (2.14) is an estimator for the projection of $h(\cdot)$ on $\mathcal{M}(K)$. Hence the goodness of $\widetilde{h}(\cdot)$ as an estimator for $h(\cdot)$ depends critically on (i) kernel function K, (ii) set $\mathcal{G}$ and $\mathsf{P}_{\mathbf{X}}(\cdot)$ on $\mathcal{G}$. In the simulation studies in Section 5, we will illustrate an approach to specify $\mathcal{G}$. Ideally we would like to choose a $K(\cdot,\cdot)$ that induces a large enough $\mathcal{M}(K)$ such that $h \in \mathcal{M}(K)$. Some frequently used kernel functions include

- Gaussian kernel: $K(\mathbf{u},\mathbf{v}) = \exp(-||\mathbf{u}-\mathbf{v}||^2/c)$,
- Thin-plate spline kernel: $K(\mathbf{u},\mathbf{v}) = ||\mathbf{u}-\mathbf{v}||^2 \log(||\mathbf{u}-\mathbf{v}||)$,
- Polynomial kernel (Fu, Shih and Wang (2011)):

$$K(\mathbf{u},\mathbf{v}) = \begin{cases} [1-(\mathbf{u}'\mathbf{v})^{\ell+1}]/(1-\mathbf{u}'\mathbf{v}), & \text{if } \mathbf{u}'\mathbf{v} \neq 1, \\ \ell+1, & \text{otherwise,} \end{cases}$$

where $||\cdot||$ denotes the Euclidean norm, c is a positive constant, and $\ell \geqslant 1$ is an integer. Also note that, for any functions in $\psi_1, \cdots, \psi_d \in \mathcal{L}_2(\mathcal{G})$,

$$K(\mathbf{u},\mathbf{v}) = \sum_{k=1}^{d} \psi_k(\mathbf{u})\psi_k(\mathbf{v}) \tag{2.15}$$

is a well-defined Mercer kernel. A possible choice of the kernel function is to let $\{\psi_1(\mathbf{u}), \cdots, \psi_d(\mathbf{u})\}$ be a set of basis functions of $\mathbf{u}$, e.g., Fourier series,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

polynomial series, wavelets, B-spline, etc. The numerical studies in Sections 5 and 6 use (2.15) with appropriately chosen functions ψ_k in the estimation and dimension reduction procedure, which performs reasonably well. We have that the dimension of $\mathcal{M}(K)$ with $K(\cdot,\cdot)$ defined above is controlled by d.

PROPOSITION 2. *For the kernel function $K(\cdot,\cdot)$ defined in (2.15), $\dim\{\mathcal{M}(K)\} \leqslant d$.*

3 Large sample theory

In this section, we study the asymptotic properties for the estimators of the eigenvalues and eigenfunctions of the Mercer kernel as well as the mean regression estimation. We start with some regularity conditions which are sufficient to derive our asymptotic theory.

ASSUMPTION 1. *The process $\{(Y_i, \mathbf{X}_i)\}$ is strictly stationary and α-mixing (or strongly mixing) dependent with*

$$\alpha_t = O(t^{-\kappa}), \quad \kappa > 2\delta_* + p + \frac{3}{2}, \tag{3.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

where p is the dimension of the random covariate, and $0 \leqslant \delta_* < \infty$ such that the volume of the set $\mathcal{G}$ has the order m^{δ_*}.

ASSUMPTION 2. *The positive eigenvalues of the Mercer kernel $K(\cdot, \cdot)$ are distinct and satisfy $0 < \lambda_d < \cdots < \lambda_2 < \lambda_1 < \infty$.*

ASSUMPTION 3. *The eigenfunctions φ_j, $j = 1, \cdots, d$, are Lipschitz continuous and bounded on the set $\mathcal{G}$. The kernel $K(\cdot, \mathbf{x})$ is Lipschitz continuous and bounded on the set $\mathcal{G}$ for any $\mathbf{x} \in \mathcal{G}$.*

REMARK 1. Assumption 1 is mild and can be satisfied by some commonly-used time series models; see e.g., Section 2.6 of Fan and Yao (2003) and the references within. For example the causal ARMA processes with continuous innovations are α-mixing with exponentially decaying mixing coefficients. For the processes with exponentially decaying mixing coefficients, (3.1) is fulfilled automatically, and the technical arguments in the proofs can be simplified. We allow set $\mathcal{G}$ to expand with the size of the sub-sample in $\mathcal{G}$ in the order of m^{δ_*}, and δ_* would be 0 if $\mathcal{G}$ is bounded. Assumptions 2 and 3 impose mild restrictions on the eigenvalues and eigenfunctions of the Mercer kernel, respectively. The boundedness condition on φ_j and $K(\cdot, \mathbf{x})$ in Assumption 3 can be replaced by $2(2+\delta)$-order moment conditions for some $\delta > 0$, and Proposition 3 still holds at the cost of more lengthy arguments.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

Furthermore, by the smoothness condition on the kernel function, and using (3.2) in Proposition 3 below, we can easily show that the $\widetilde{\varphi}_j(\cdot)$ defined in (2.8), $j = 1, \cdots, d$, are Lipschitz continuous and bounded with probability tending to one.

PROPOSITION 3. *If Assumptions 1–3 hold, then*

$$\max_{1 \leqslant k \leqslant d} \left| \widetilde{\lambda}_k - \lambda_k \right| = \max_{1 \leqslant k \leqslant d} \left| \frac{1}{m} \widehat{\lambda}_k - \lambda_k \right| = O_P\left(m^{-1/2}\right) \tag{3.2}$$

$$\max_{1 \leqslant k \leqslant d} \sup_{\mathbf{x} \in \mathcal{G}} |\widetilde{\varphi}_k(\mathbf{x}) - \varphi_k(\mathbf{x})| = O_P\left(\xi_m\right), \tag{3.3}$$

where $\xi_m = m^{-1/2} \log^{1/2} m$.

REMARK 2. Proposition 3 is of independent interest. It complements some statistical properties of the KPCA in the literature such as Braun (2005) and Blanchard, Bousquet and Zwald (2007). Note that $\mathsf{P}(\mathbf{X} \in \mathcal{G})$ can be consistently estimated by m/n. If it is assumed that $\mathsf{P}(\mathbf{X} \in \mathcal{G}) = c_0 > 0$, m would be of the same order as the full sample size n (with probability tending to one). As a consequence, the convergence rates in (3.2) and (3.3) would be equivalent to $O_P\left(n^{-1/2}\right)$ and $O_P\left(n^{-1/2} \log^{1/2} n\right)$, respectively, which are not uncommon in the context of functional principal component analysis (Bosq (2000); Horváth and Kokoszka (2012)).

THEOREM 1. *If Assumptions 1–3 hold,* $\mathsf{E}[|Y|^{2+\delta}] < \infty$ *for some* $\delta > 0$ *and*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

$h(\cdot) \in \mathcal{M}(K)$, then for the ξ_m of Proposition 3,

$$\sup_{\mathbf{x}\in\mathcal{G}} \left|\widetilde{h}(\mathbf{x}) - h(\mathbf{x})\right| = O_P\left(\xi_m\right). \tag{3.4}$$

REMARK 3. The uniform convergence rate in (3.4) is thus equivalent to $O_P\left(n^{-1/2}\log^{1/2} n\right)$, which is faster than the well-known uniform convergence rate $O_P\left((nb)^{-1/2}\log^{1/2} n\right)$ in the kernel smoothing method (Fan and Yao (2003)), where b is a bandwidth that converges to zero as n tends to ∞. The intrinsic reason of the faster rate in (3.4) is that we assume the dimension of the subset-based kernel feature space is finite, and thus the number of the unknown elements in (2.12) is also finite. Section 4 below shows that the increasing dimension of the kernel feature space slows down the convergence rates.

4 Extensions of the estimation methodology

In this section, we consider two extensions of the methodology proposed in Section 2: the estimation of the conditional distribution function, and when the dimension of a kernel feature space diverges together with the sample size.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

4.1 Estimation for conditional distribution functions

Estimation of the conditional distribution function defined in (1.2) is a key aspect in such statistical topics as quantile regression, as the conditional mean regression may be not informative enough in many situations. Nonparametric estimation of the conditional distribution has been extensively studied in the literature, including Hall, Wolff and Yao (1999), Hansen (2004) and Hall and Yao (2005). In this section, we use the subset-based KPCA approach to estimate a conditional distribution function in low-dimensional kernel feature space when the random covariates are multi-dimensional.

Let $F_*(y|\mathbf{x}) = F_{Y|\mathbf{X}}(y|\mathbf{x}) - c_*$, where $c_* = \mathsf{P}(Y \leqslant y, \mathbf{X} \in \mathcal{G})$. Then $\mathsf{E}\left[F_*(y|\mathbf{X})\right] = 0$. In practice c_* can be easily estimated by the relative frequency. Suppose that

$$F_*(y|\mathbf{x}) = F_{Y|\mathbf{X}}(y|\mathbf{x}) - c_* = \int_{-\infty}^{y} f_{Y|\mathbf{X}}(z|\mathbf{x})dz - c_* = \sum_{k=1}^{d} \beta_k^* \varphi_k(\mathbf{x}), \quad \mathbf{x} \in \mathcal{G}. \tag{4.1}$$

The coefficients β_k^* in this decomposition depend on y. The orthonormality

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

of φ_i implies that

$$\begin{aligned}\beta_k^* &= \langle F_*(y|\cdot), \varphi_k\rangle = \int_{\mathcal{G}} \varphi_k(\mathbf{x})\mathsf{P}_{\mathbf{X}}(d\mathbf{x}) \left[\int_{-\infty}^{y} f_{Y|\mathbf{X}}(z|\mathbf{x})dz - c_*\right] \\ &= \int I(z \leqslant y,\, \mathbf{x} \in \mathcal{G})\varphi_k(\mathbf{x}) f_{Y|\mathbf{X}}(z|\mathbf{x})dz\mathsf{P}_{\mathbf{X}}(d\mathbf{x}) - c_* \int_{\mathcal{G}} \varphi_k(\mathbf{x})\mathsf{P}_{\mathbf{X}}(d\mathbf{x}) \\ &= \mathsf{E}\left[I(Y \leqslant y,\, \mathbf{X} \in \mathcal{G})\varphi_k(\mathbf{X})\right] - c_*\mathsf{E}\left[I(\mathbf{X} \in \mathcal{G})\,\varphi_k(\mathbf{X})\right].\end{aligned}$$

This leads to the estimator for β_k^* as

$$\widetilde{\beta}_k^* = \frac{1}{m}\sum_{i=1}^{m} I(Y_i^{\mathcal{G}} \leqslant y)\widetilde{\varphi}_k(\mathbf{X}_i^{\mathcal{G}}) - \frac{\widetilde{c}_*}{m}\sum_{i=1}^{m}\widetilde{\varphi}_k(\mathbf{X}_i^{\mathcal{G}}), \tag{4.2}$$

where $(Y_i^{\mathcal{G}}, \mathbf{X}_i^{\mathcal{G}})$ are defined in (2.5), $\widetilde{\varphi}_k(\cdot)$ are defined in (2.8), and

$$\widetilde{c}_* = \frac{1}{n}\sum_{i=1}^{n} I\left(Y_i \leqslant y,\ \mathbf{X}_i \in \mathcal{G}\right), \tag{4.3}$$

n the full sample size. Consequently, we obtain an estimator for the conditional distribution

$$\widetilde{F}_{Y|\mathbf{X}}(y|\mathbf{x}) = \sum_{k=1}^{d} \widetilde{\beta}_k^*\widetilde{\varphi}_k(\mathbf{x}) + \widetilde{c}_*. \tag{4.4}$$

The estimator $\widetilde{F}_{Y|\mathbf{X}}(\cdot|\mathbf{x})$ is not necessarily a bona fide distribution function. Some further normalization may be required to make the estimator non-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

negative, non-decreasing and between 0 and 1 (Glad, Hjort and Ushakov (2003)).

By the classic result for α-mixing sequences, we can show that $\widetilde{c}_*$ is a consistent estimator of c_* with a root-n convergence rate. Then, by Proposition 3 and the proof of Theorem 1, we have a convergence result for $\widetilde{F}_{Y|\mathbf{X}}(y|\mathbf{x})$.

THEOREM 2. *If Assumptions 1–3 hold and* $F_*(y|\cdot) \in \mathcal{M}(K)$, *then*

$$\sup_{\mathbf{x}\in\mathcal{G}} \left|\widetilde{F}_{Y|\mathbf{X}}(y|\mathbf{x}) - F_{Y|\mathbf{X}}(y|\mathbf{x})\right| = O_P\left(\xi_m\right) \tag{4.5}$$

for any given y, *where* ξ_m *is defined in Proposition 3.*

4.2 Kernel feature spaces with diverging dimensions

We next consider the case when the dimension of the kernel feature space $d_m \equiv \max\{k : \lambda_k > 0\}$ depends on m, and may diverge to infinity as m tends to infinity. Let

$$\rho_m = \min\left\{\lambda_k - \lambda_{k+1}, \quad k = 1, \cdots, d_m\right\}.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

For a more general asymptotic theory, we need to slightly modify Assumption 2.

ASSUMPTION 2*. *The positive eigenvalues of the Mercer kernel $K(\cdot,\cdot)$ are distinct and satisfy $0 < \lambda_{d_m} < \cdots < \lambda_2 < \lambda_1 < \infty$, $\sum_{k=1}^{d_m} \lambda_k < \infty$.*

PROPOSITION 4. *If Assumptions 1, 2* and 3 hold, $d_m = o\left(m\rho_m^2\lambda_{d_m}^2/\log m\right)$, and the α-mixing coefficient decays to zero at an exponential rate, then*

$$\max_{1\leqslant k\leqslant d_m}\left|\widetilde{\lambda}_k - \lambda_k\right| = \max_{1\leqslant k\leqslant d_m}\left|\frac{1}{m}\widehat{\lambda}_k - \lambda_k\right| = O_P\left(d_m^{1/2}\xi_m\right), \tag{4.6}$$

$$\max_{1\leqslant k\leqslant d_m}\sup_{\mathbf{x}\in\mathcal{G}}|\widetilde{\varphi}_k(\mathbf{x}) - \varphi_k(\mathbf{x})| = O_P\left(d_m^{1/2}\xi_m/(\rho_m\lambda_{d_m})\right). \tag{4.7}$$

REMARK 4. When d_m is fixed, if we take both ρ_m and λ_{d_m} bounded away from zero, the convergence rates in Proposition 4 are simplified. When $d_m \to \infty$ as $m \to \infty$, we usually have $\rho_m \to 0$ and $\lambda_{d_m} \to 0$. This implies that the convergence rates in (4.6) and (4.7) would be generally slower than those in (3.2) and (3.3). Let c_i, $i = 1, \cdots, 5$, be five positive constants. For any two sequences a_m and b_m, $a_m \propto b_m$ means that $0 < c_4 \leqslant a_m/b_m \leqslant c_5 < \infty$ when m is sufficiently large. If $d_m = c_1 \log m$, $\rho_m = c_2 \log^{-1} m$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

$\lambda_{d_m} = c_3 \log^{-1} m$, we have

$$d_m^{1/2}\xi_m \propto m^{-1/2}\log m, \quad d_m^{1/2}\xi_m/(\rho_m\lambda_{d_m}) \propto m^{-1/2}\log^3 m.$$

Using Proposition 4 and the proof of Theorem 1, we can obtain the uniform convergence rate for the conditional mean regression estimation when d_m is diverging.

In practice, we may encounter the more challenging case when the dimension of the Mercer kernel is infinite (e.g., $\lambda_k \propto k^{-\iota_1}$ with $\iota_1 > 0$ or $\lambda_k \propto \iota_2^k$ with $0 < \iota_2 < 1$). Then, the convergence result in Proposition 4 is not directly applicable as the rates in (4.6) and (4.7) diverge when the dimension is infinite. However, the proposed subset-based KPCA approach can still be used to estimate the conditional mean regression function. Assuming that the mean regression function $h(\mathbf{x}) \in \mathcal{M}(K)$ and noting that the dimension of $\mathcal{M}(K)$ is infinite, we have

$$h(\mathbf{x}) = \sum_{k=1}^{\infty} \beta_k \varphi_k(\mathbf{x}) = \sum_{k=1}^{d_m} \beta_k \varphi_k(\mathbf{x}) + \sum_{k=d_m+1}^{\infty} \beta_k \varphi_k(\mathbf{x}) \equiv h_1(\mathbf{x}) + h_2(\mathbf{x}), \quad (4.8)$$

where β_k and $\varphi_k(\mathbf{x})$ are defined as in Section 2, and d_m is a diverge under the condition in Proposition 4. Let $\mathcal{M}_1(K)$ be a d_m-dimensional kernel feature space spanned by $\varphi_1, \cdots, \varphi_{d_m}$. From (4.8), the mean regression

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

function $h(\mathbf{x})$ can be well approximated by its projection onto the d_m-dimensional kernel feature space $\mathcal{M}_1(K)$ as long as the approximation error $h_2(\mathbf{x})$ uniformly converges to zero at a certain rate. The latter usually holds if we impose some smoothness condition on $h(\mathbf{x})$ and let d_m diverge at an appropriate rate; this is similar to the conditions on sieve approximation accuracy (Chen (2007)).

Let $b_m^\star = \sup_{\mathbf{x}\in\mathcal{G}} |h_2(\mathbf{x})|$. We can estimate β_k and φ_k, $k = 1, \cdots, d_m$, in the same manner as in Sections 2.2 and 2.3. Denote the estimates by $\widetilde{\beta}_k$ and $\widetilde{\varphi}_k$, and let $\widetilde{h}_m(\mathbf{x}) = \sum_{k=1}^{d_m} \widetilde{\beta}_k \widetilde{\varphi}_k(\mathbf{x})$. By Proposition 4 and the proof of Theorem 1, we can establish the uniform convergence rate

$$\sup_{\mathbf{x}\in\mathcal{G}} \left|\widetilde{h}_m(\mathbf{x}) - h_1(\mathbf{x})\right| = O_P(\nu_m^\star),$$

where $\nu_m^\star = d_m^{3/2}\xi_m/(\rho_m \lambda_{d_m})$. Furthermore, we can prove, via the decomposition in (4.8), that

$$\sup_{\mathbf{x}\in\mathcal{G}} \left|\widetilde{h}_m(\mathbf{x}) - h(\mathbf{x})\right| = \sup_{\mathbf{x}\in\mathcal{G}} \left|\widetilde{h}_m(\mathbf{x}) - h_1(\mathbf{x})\right| + \sup_{\mathbf{x}\in\mathcal{G}} |h_2(\mathbf{x})| = O_P(\nu_m^\star + b_m^\star).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

5 Simulation Studies

In this section, we report on three simulations that illustrate the finite sample performance of the proposed subset-based KPCA method and compare it with the global KPCA and other existing nonparametric estimation methods: cubic spline, local linear regression and kernel ridge regression. The first simulation assesses the out-of-sample estimation performance of conditional mean function based on a multivariate nonlinear regression model. In the second simulation, we examine the one-step ahead out-of-sample forecast performance based on a multivariate nonlinear time series model.In the third simulation, we examine the finite sample performance of the estimation of conditional distribution function.

Throughout this section, the kernel function is either the Gaussian kernel or as formulated in (2.15) with $\{\psi_1(\mathbf{u}), \cdots, \psi_d(\mathbf{u})\}$ a set of normalized polynomial basis functions (with the unit norm) of $\mathbf{u} = (u_1, \cdots, u_p)^{\mathrm{T}}$ of order 2 and 3: $\{1, u_k, \cdots u_k^r,\ k = 1, \cdots, p\}$, where $r = 2, 3$ and $d = rp + 1$. Here we call the kernel *the quadratic kernel* when $r = 2$ and *the cubic kernel* when $r = 3$. In practice, d is estimated by the ratio method as in (2.10). The simulation results show that (2.10) can correctly estimate $\widehat{d} = d$ with frequency close to 1. The subset is chosen to be the $\lfloor \kappa n \rfloor$ nearest neighbors, where n is the sample size and $\kappa \in (0, 1)$ is a constant bandwidth. The

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

bandwidth κ and the tuning parameter c in the Gaussian kernel are selected by a 5-fold cross validation.

EXAMPLE 5.1. Consider the model

$$y_i = g(x_{2i}) + \sin\{\pi(x_{3i} + x_{4i})\} + x_{5i} + \log(1 + x_{6i}^2) + \varepsilon_i,$$

where $x_{1i}, \cdots, x_{6i}$ and ε_i are i.i.d. $\mathsf{N}(0, 1)$, $g(x) = \mathrm{e}^{-2x^2}$ for $x \geqslant 0$, and $g(x) = \mathrm{e}^{-x^2}$ for $x < 0$. In the model, the covariate x_{1i} is irrelevant to y_i.

We drew a training sample of size $n = 500$ or 1000 and a testing sample of size 200. We estimated the conditional mean regression function using the training sample, and then calculated the mean squared errors (MSE) and out-of-sample R^2s over the testing sample as follows:

$$\mathrm{MSE} = \frac{1}{200}\sum_{i=1}^{200}\left[y_i - \widetilde{h}(\mathbf{x}_i)\right]^2, \quad R^2 = 1 - \frac{\sum_{i=1}^{200}[y_i - \widetilde{h}(\mathbf{x}_i)]^2}{\sum_{i=1}^{200}(y_i - \bar{y})^2},$$

where $\widetilde{h}(\cdot)$ is defined as in (2.14), $\mathbf{x}_i = (x_{1i}, \cdots, x_{6i})^{\mathrm{T}}$, and $\bar{y}$ is the sample mean of y_i over the training sample.

By repeating this procedure over 200 replications, we obtained a sample of MSE and R^2 with size 200. The estimation performance was assessed by the sample mean, median and variance of MSE and R^2. The simulation

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

results are reported in Table 1. In this simulation, for the quadratic kernel and cubic kernel, the ratio method in (2.10) can always correctly estimate $\widehat{d} = rp + 1$. According to the results in Table 1, the subset-based KPCA with the quadratic kernel outperforms the global KPCA methods and other nonparametric methods as it has the smallest sample mean, median and variance of MSE and the highest R^2. In addition, both the quadratic kernel and cubic kernel perform better than the Gaussian kernel due to the fact that they can better capture different degree of smoothness in different directions.

To assess the the bandwidth choice for subset-based KPCA, we set $n = 500$, let κ vary from 0.05 to 0.8 and calculated the sample mean of MSE over 100 replications. The results are plotted in Figure 1. According to Figure 1, the subset-based KPCA method is not sensitive to the choice of κ. The smallest MSE is achieved at $\kappa = 0.27$, and any κ between 0.15 and 0.45 yields similar result.

Furthermore, we compared the computational costs between the subset-based KPCA and global KPCA approaches with quadratic kernel function. The computational cost of subset-based KPCA includes the selection of bandwidth κ. The bandwidth κ was selected by 5-fold cross validation over 10 grid points, equally spaced between 0.1 and 0.5. We let the sample size

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

increase from 400 to 1000 with a step size of 20 and recorded the computational time over 100 replications for both approaches. The comparison results are presented in Figure 2. The major computational cost of the global KPCA methods is the eigen-decomposition of the $n \times n$ gram matrix which is of computational complexity $O(n^\omega)$ for some $\omega > 2$. To see this, we calculated the empirical order of growth for both approaches:

$$\widehat{\omega} = \frac{1}{30} \sum_{l=1}^{30} \frac{\log(T_{l+1}/T_l)}{\log(n_{l+1}/n_l)},$$

where n_l is the l-th component in the set $\{400, 420, \cdots, 980, 1000\}$ and T_l is the corresponding computational cost over 100 replications. The empirical order of growth for the global KPCA is 2.69, whereas that for the subset-based KPCA is 2.17. Hence, both Figure 2 and the calculation of empirical order of growth show the subset-based KPCA method scales with sample size much better than does its global counterpart.

EXAMPLE 5.2. Consider the time series model

$$y_t = \sin(0.02\pi y_{t-1}) + \exp(-y_{t-2}^2) + \ln(1 + |y_{t-3}|) - 0.3|y_{t-4}| + 0.2\epsilon_t,$$

where $y_0 = 0$ and $\{\epsilon_t\}$ is a sequence of independent $\mathsf{N}(0, 1)$ random variables.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

Table 1: Out-of-sample estimation performance in Example 5.1

	MSE (the smaller the better)			R^2 (the larger the better)		
$n = 500$	Mean	Median	Variance	Mean	Median	Variance
sKPCA+Quadratic	1.300	1.294	0.017	0.548	0.550	0.0020
sKPCA+Cubic	1.337	1.335	0.018	0.536	0.536	0.0021
sKPCA+Gaussian	1.573	1.586	0.026	0.454	0.448	0.0031
gKPCA+Quadratic	2.389	1.871	0.059	0.170	0.350	0.0071
gKPCA+Cubic	2.586	1.937	0.064	0.102	0.327	0.0077
gKPCA+Gaussian	3.023	2.021	0.093	-0.049	0.298	0.0112
Cubic Spline	1.386	1.383	0.018	0.518	0.519	0.0022
Local Linear	1.429	1.431	0.020	0.504	0.503	0.0024
Kernel Ridge	1.897	1.866	0.048	0.340	0.351	0.0056
$n = 1000$	Mean	Median	Variance	Mean	Median	Variance
sKPCA+Quadratic	1.243	1.236	0.013	0.575	0.578	0.0015
sKPCA+Cubic	1.278	1.271	0.016	0.564	0.566	0.0018
sKPCA+Gaussian	1.531	1.528	0.025	0.477	0.478	0.0029
gKPCA+Quadratic	2.380	2.371	0.051	0.187	0.191	0.0059
gKPCA+Cubic	2.541	2.508	0.059	0.133	0.144	0.0069
gKPCA+Gaussian	3.015	2.790	0.086	-0.029	0.047	0.0100
Cubic Spline	1.371	1.372	0.017	0.532	0.531	0.0020
Local Linear	1.404	1.418	0.018	0.520	0.516	0.0021
Kernel Ridge	1.858	1.831	0.042	0.324	0.346	0.0055

"sKPCA" and "gKPCA" stand for the subset-based KPCA and global KPCA; "Quadratic", "Cubic" and "Gaussian" stand for the quadratic kernel, cubic kernel and Gaussian kernel, respectively; "Cubic Spline", "Local Linear" and "Kernel Ridge" stand for non-parametric estimation methods based on cubic spline, local linear regression and kernel ridge regression.

We estimate the conditional mean $\mathsf{E}(y_t|y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})$ and denote the estimator as $\widehat{y}_t$ which is to be used as the one-step-ahead predictor of y_t.

We generated a time series sample from this model with length $T + 100$. For $k = 1, \cdots, 100$, we used the T observations just before time $T + k$ as

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

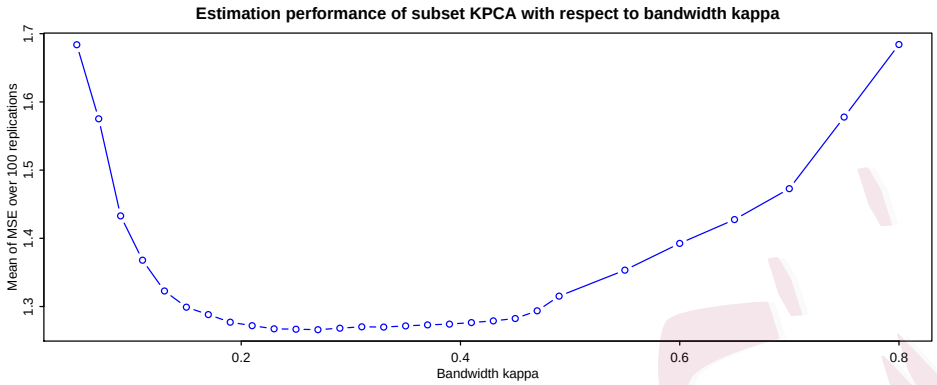

Figure 1: The out-of-sample estimation performance of the subset-based KPCA approach with the quadratic kernel with respect to the bandwidth κ when $n = 500$.

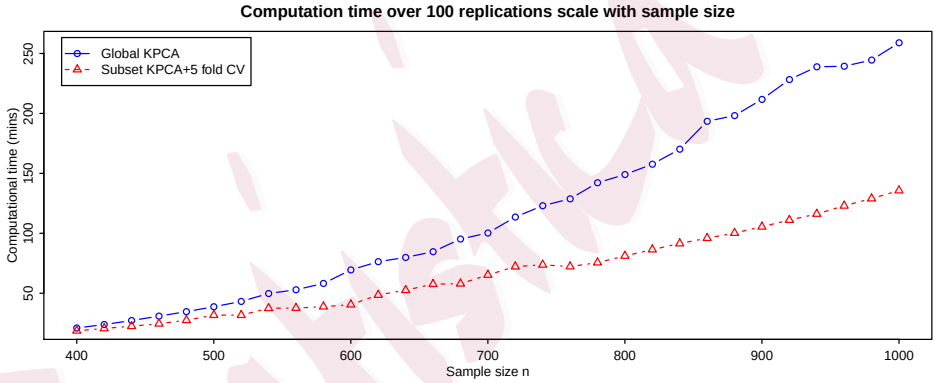

Figure 2: The computation costs for global KPCA method and subset-based KPCA method (with κ selected by 5-fold cross validation) with respect to the sample size.

the training set to predict y_{T+k}. The performance was measured by MSE

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

and out-of-sample R^2:

$$\text{MSPE} = \frac{1}{100}\sum_{k=1}^{100}(y_{T+k} - \widehat{y}_{T+k})^2, \quad R^2 = 1 - \frac{\sum_{k=1}^{100}(y_{T+k} - \widehat{y}_{T+k})^2}{\sum_{k=1}^{100}(y_{T+k} - \bar{y})^2},$$

where $\bar{y}$ is the sample mean of y_t over the training sample.

We set $T = 500$, and repeated the experiment 200 times for each method. The sample means, medians and variances of MSE and R^2 are presented in Table 2. As in Example 5.1, the subset-based KPCA method with the quadratic kernel provides the most accurate prediction. The subset-based KPCA method with the cubic kernel is a close second best in terms of both MSE and R^2. Figure 3 plots a typical path together with their one-step-ahead forecasts for each method. The typical path is the replication with median R^2. Figure 3 shows that the forecasted path from the subset-based KPCA method with the quadratic kernel follows the true path closely. A similar pattern can also be found for other subset-based KPCA method with the cubic and Gaussian kernel and the three nonparametric methods (cubic spline, local linear and kernel ridge). However the global KPCA methods fail to capture the variation of the series and tend to forecast the future values by the overall mean value, which is not satisfactory.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

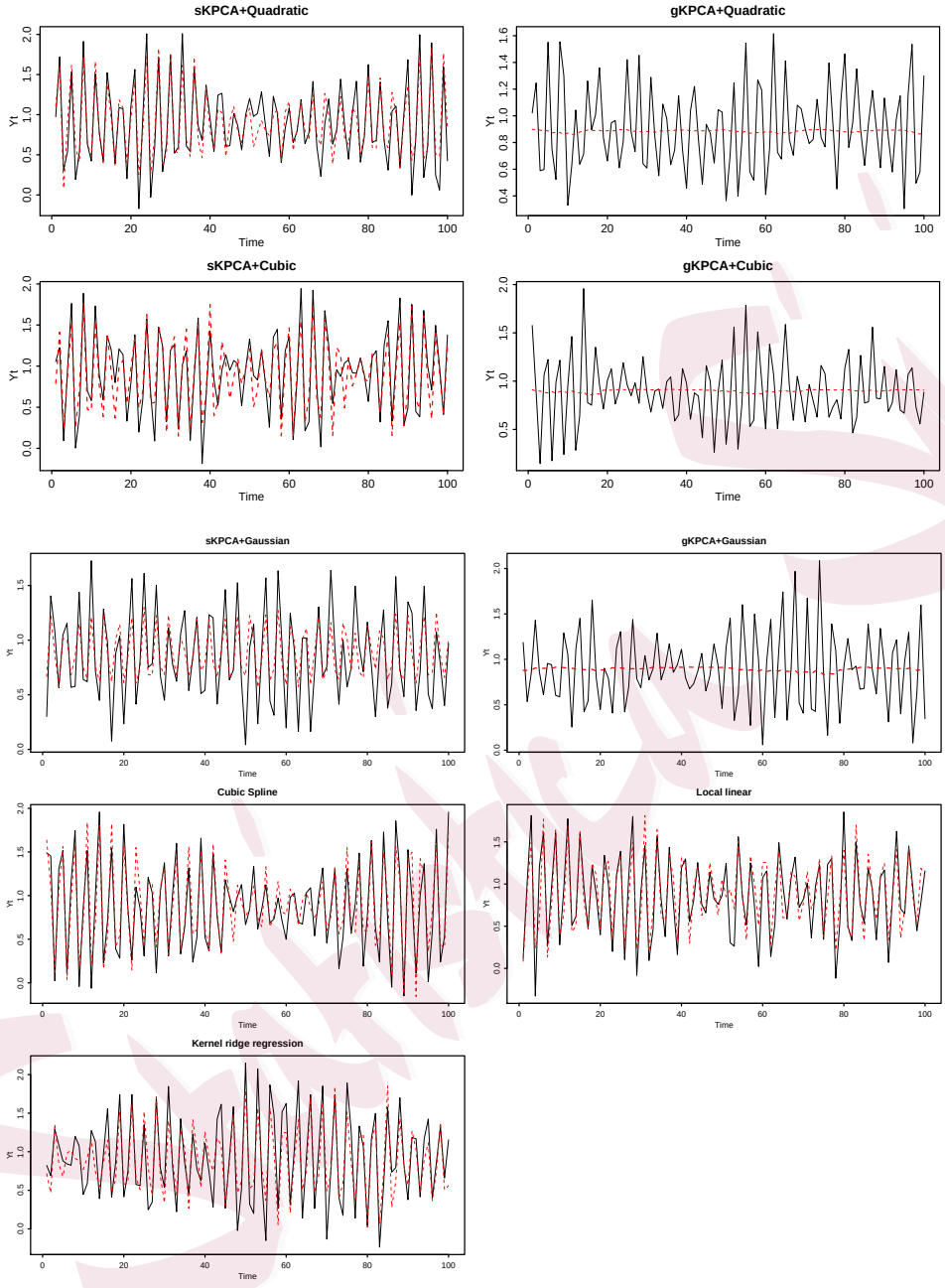

Figure 3: One-step ahead out-of-sample forecasting performance based on the replication with median R^2 for each method. The black solid line is the true value and the red dashed line is the predicted value.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

Table 2: One-step ahead forecasting performance in Example 5.2

	MSE (the smaller the better)			R^2 (the larger the better)		
	Mean	Median	Variance	Mean	Median	Variance
sKPCA+Quadratic	0.0435	0.0428	3.9×10^{-5}	0.804	0.807	7.9×10^{-4}
sKPCA+Cubic	0.0445	0.0437	4.6×10^{-5}	0.799	0.803	9.3×10^{-4}
sKPCA+Gaussian	0.0756	0.0751	4.1×10^{-4}	0.659	0.661	8.3×10^{-3}
gKPCA+Quadratic	0.1900	0.1908	6.9×10^{-4}	0.144	0.141	0.014
gKPCA+Cubic	0.2042	0.2058	9.5×10^{-4}	0.080	0.073	0.019
gKPCA+Gaussian	0.2172	0.2211	0.0038	0.022	0.041	0.077
Cubic Spline	0.0516	0.0512	4.7×10^{-5}	0.767	0.769	9.5×10^{-4}
Local Linear	0.0522	0.0515	4.9×10^{-5}	0.764	0.768	9.9×10^{-4}
Kernel Ridge	0.0721	0.0717	6.8×10^{-5}	0.675	0.677	1.4×10^{-3}

EXAMPLE 5.3. Consider the model

$$X_1 \sim \mathsf{N}(0,\ 1), \quad X_2 \sim \mathsf{N}(0,\ 1),$$

$$Y|(X_1, X_2) \sim \mathsf{N}(X_1,\ 1 + X_2^2).$$

The conditional distribution of Y given $\mathbf{X} \equiv (X_1, X_2)^{\mathrm{T}}$ is a normal distribution with mean X_1 and variance $1 + X_2^2$. According to the method proposed in Section 4.1, we estimate the conditional distribution function $F_{Y|\mathbf{X}}(y|\mathbf{x})$ using the subset-based KPCA with the quadratic kernel.

We drew a training sample of size $n = 300$ or 500 and a testing sample of size 100. The estimated conditional distribution $\widetilde{F}_{Y|\mathbf{X}}(y_i|\mathbf{x}_i)$ was obtained using the training data. We repeated the simulation 200 times and measured

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

the performance by MSE as well as largest absolute error (LAE) over the testing sample:

$$\text{MSE} = \frac{1}{100}\sum_{i=1}^{100}\left[\widetilde{F}_{Y|\mathbf{X}}(y_i|\mathbf{x}_i) - F_{Y|\mathbf{X}}(y_i|\mathbf{x}_i)\right]^2,$$

$$\text{LAE} = \sup_{(y,\mathbf{x})\in\Omega^*}\left|\widetilde{F}_{Y|\mathbf{X}}(y|\mathbf{x}) - F_{Y|\mathbf{X}}(y|\mathbf{x})\right|,$$

where Ω^* is the union of all validation sets. The results are reported in Table 3. As the values of MSE and LAE in Table 3 are small, the proposed method provides accurate estimation for the conditional distribution function.

Table 3: Estimation of the conditional distribution function

	MSE			LAE
	Mean	Median	Variance	
$n = 300$	6.0×10^{-4}	4.1×10^{-4}	3.6×10^{-7}	0.098
$n = 500$	3.7×10^{-4}	2.8×10^{-4}	8.6×10^{-8}	0.080

6 Data analysis

In this section, we apply the proposed subset-based KPCA method to two data sets. Throughout this section, the kernel function is either the Gaussian or the Quadratic kernel. The subset is chosen to be the $\lfloor \kappa n \rfloor$ nearest neighbors, where n is the sample size and $\kappa \in (0, 1)$. The bandwidth

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

κ is selected by 5-fold cross validation.

6.1 Circulatory and respiratory problem in Hong Kong

We studied the circulatory and respiratory problem in Hong Kong via an environmental data set. That contains 730 observations that were collected between January 1, 1994 and December 31, 1995. The response variable is the number of daily total hospital admissions for circulatory and respiratory problems in Hong Kong, and the covariates are daily measurements of seven pollutants and environmental factors: SO_2, NO_2, dust, temperature, change of temperature, humidity and ozone. We standardized the data so that all the covariates had zero sample mean and unit sample variance. To check the stationarity, we applied the augmented Dickey-Fuller test (Dickey and Fuller (1981)) to each variable in the data set. The tests were applied using the "urca" package in R and the lags included were selected by AIC. For each variable, the test result suggests rejection of the unit root null hypothesis. Therefore, we considered the variables in the dataset to be stationary.

The objective of this study was to estimate the number of daily total hospital admissions for circulatory and respiratory problem using the collected environmental data, to estimate the conditional mean regression function. The estimation performance was measured by the mean and variance of the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

out-of-sample R^2, that was calculated by a bootstrap method, as follows. We first randomly divided the data set into a training set of 700 observations and a testing set of 30 observations. For each observation in the testing set, we used the training set to estimate its conditional mean regression function. Then we calculated out-of-sample R^2 for the testing set as in Example 5.1. By repeating this re-sampling and estimation procedure 1000 times, we obtained a bootstrap sample of R^2s, and calculated its sample mean and variance.

Table 4: Estimation performance for the Hong Kong environmental data

	R^2 (the larger the better)	
Method	Mean	Variance
sKPCA + Quadratic	0.1544	0.0025
sKPCA + Gaussian	0.1262	0.0027
gKPCA + Quadratic	-0.3613	0.2232
gKPCA + Gaussian	-3.7058	1.6653
Cubic spline	0.0687	0.0042

We compared the performance among the five methods: the subset-based KPCA with the quadratic kernel, the subset-based KPCA with the Gaussian kernel, the global KPCA with the quadratic kernel, the global KPCA with the Gaussian kernel, and the cubic spline. The cubic spline was fitted with 10 knots using the "splines" package in R. The results are

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

presented in Table 4. According to the results in Table 4, the subset-based KPCA with the quadratic kernel has the best estimation performance and the subset-based KPCA method outperforms the global KPCA method and cubic spline.

6.2 Forecasting the log return of CPI

The CPI is a statistical estimate that measures the average change in the price paid to a market basket of goods and services. The CPI is often used as an important economic indicator in macroeconomic and financial studies. In economics, the CPI is considered closely related to the cost-of-living index and used to adjust the income eligibility levels for government assistance. In finance, the CPI is considered as an indicator of inflation and used as the deflater to translate other financial series to inflation-free ones. Hence, it is always of interest to forecast the CPI. We performed one-step-ahead forecasting for the monthly log return of CPI in the USA based on the proposed subset-based KPCA method with the quadratic kernel. The data span from January 1970 to December 2014 with 540 observations. The augmented Dickey-Fuller test suggests the monthly log return of CPI over this time span is stationary.

Instead of using the traditional linear time series models, we considered

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

that the log return of CPI follow the nonlinear AR(3) model

$$y_t = g(y_{t-1},\ y_{t-2},\ y_{t-3}) + \epsilon_t,$$

where $g(\cdot)$ is an unknown function and ϵ_t denotes an unobservable noise at time t. The regression function $g(\cdot)$ was estimated by the subset-based KPCA method with the quadratic kernel.

For comparison, we also forecast y_t based on a linear AR(p) model with the order p determined by AIC. When the testing set starts from time T and ends at time $T + S$, the forecast performance is measured by the out-of-sample R^2 as

$$\mathrm{R}^2 = 1 - \frac{\sum_{s=1}^{S} (y_{T+s} - \widehat{y}_{T+s})^2}{\sum_{s=1}^{S} (y_{T+s} - \bar{y})^2},$$

where $\widehat{y}_{T+s}$ is the estimator of y_{T+s}, and $\bar{y}$ is the sample mean of y_t over the training set.

We set the data from January 2005 to December 2014 as the testing set, which contains 120 observations. We forecast each observation in the testing set with the data up to its previous month. The out-of-sample R^2 was calculated over the testing set. The out-of-sample R^2 of the nonlinear

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

AR(3) model was 0.2318 while the R^2 of the linear AR model was 0.0412. The detailed forecasting results are plotted in Figure 4, which shows clearly that the forecast based on the subset-based KPCA method is more accurate, it captures the variations much better than the linear AR modeling method.

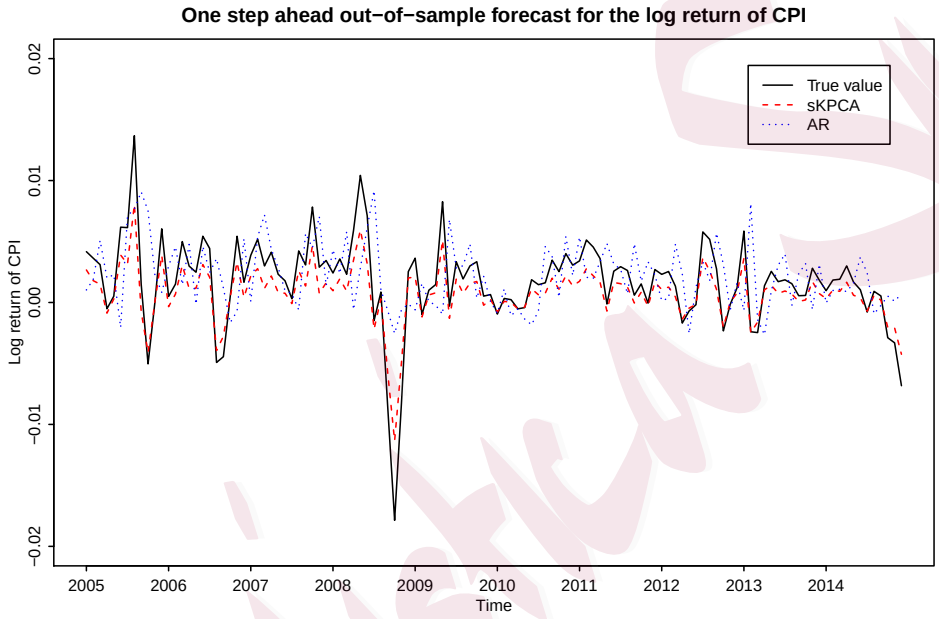

Figure 4: One step ahead out-of-sample forecast for the log return of CPI from January 2005 to December 2014. The black solid line is the true value, the red dashed line is the forecast value obtained by the subset-based KPCA, and the blue dotted line is the forecast value obtained by the linear AR model.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

7 Conclusion

In this paper, we have developed a new subset-based KPCA method for estimating nonparametric regression functions. In contrast to the conventional (global) KPCA method which builds on a global kernel feature space, we use different lower-dimensional subset-based kernel feature spaces at different locations of the sample space. Consequently the resulting localized kernel principal components provide more parsimonious representation for the target regression function, which is also reflected by the faster uniform convergence rates presented in Theorem 1, see also the discussions immediately below Theorem 1. The reported numerical results with both simulations and data sets clearly illustrate the advantages of using the subset-based KPCA method over its global counterpart. It also outperforms some popular nonparametric regression methods such as the cubic spline and kernel regression (the results on kernel regression are not reported to save the space). As well, the quadratic kernel constructed based on (2.15) using normalized univariate linear and quadratic basis functions performs better than the more conventional Gaussian kernel for the examples reported in Sections 5 and 6.

Supplementary materials

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

The online supplementary material contains the proofs of Propositions 1–4 and Theorem 1.

Acknowledgements

The authors are grateful to the editor, an associate editor and two referees for their valuable and constructive comments that substantially improved an earlier version of the paper. Yao's research is partly supported by the EPSRC research grant EP/L01226X/1.

References

Blanchard, G., Bousquet, O. and Zwald, L. (2007). Statistical properties of kernel principal component analysis. *Machine Learning*, **66**, 259–294.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Lecture Notes in Statistics, Springer.

Braun, M. L. (2005). *Spectral Properties of the Kernel Matrix and Their Relation to Kernel Methods in Machine Learning*. PhD Thesis, University of Bonn, Germany.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, **76**, North Holland, Amsterdam.

Dickey, A.D. and Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica*, **49**, 1057–1072.

Drineas, P. and Mahoney, M. (2005). On the Nyström method for approximating a Gram matrix for improved kernel-based learning. *Journal of Machine Learning Research*, **6**, 2153–2175.

Ferreira, J. C. and Menegatto, V. A. (2009). Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations and Operator Theory*, **64**, 61–81.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York.

Fu, G., Shih, F.Y. and Wang, H. (2011). A kernel-based parametric method for conditional density estimation. *Pattern Recognition*, **44**, 284-294.

Girolami, M. (2002). Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, **14**, 669-688.

Glad, I.K., Hjort, N.L. and Ushakov, N.G. (2003). Correction of density

estimators that are not densities. *Scandinavian Journal of Statistics*, **30**, 415-427.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall/CRC.

Hall, P., Wolff, R.C.L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, **94**, 154-163.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution functions using dimension reduction. *The Annals of Statistics*, **33**, 1404-1421.

Hansen, B. (2004). Nonparametric estimation of smooth conditional distributions. Working paper available at *http://www.ssc.wisc.edu/∼bhansen/papers/cdf.pdf*.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning* (2nd Edition). Springer, New York.

Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

Izbicki, R. and Lee, A.B. (2013). Nonparametric conditional density estimation in high-dimensional regression setting. *Manuscript*.

Lam, C. and Yao, Q. (2012). Factor modelling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, **40**, 694-726.

Lee, A.B. and Izbicki, R. (2013). A spectral series approach to high-dimensional nonparametric regression. *Manuscript*.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London*, **A**, **209**, 415-446.

Rosipal, R., Girolami, M., Trejo, L.J. and Cichocki, A. (2001). Kernel PCA for feature extraction and de-noising in nonlinear regression. *Neural Computing & Applications*, **10**, 231-243.

Schölkopf, B., Smola, A. J. and Müller, K. R. (1999). Kernel principal component analysis. *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, 327–352.

Teräsvirta, T., Tjøstheim, D. and Granger, C. (2010). *Modelling Nonlinear Economic Time Series*. Oxford University Press.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0369

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. Chapman and Hall/CRC.

Wibowo, A. and Desa, I.M. (2011). Nonlinear robust regression using kernel principal component analysis and R-estimators. *International Journal of Computer Science Issues*, **8**, 75-82.

Department of Statistics, Pennsylvania State University, PA 16802, U.S.A.

E-mail: yzk62@psu.edu

Department of Mathematics, The University of York, YO10 5DD, U.K.

E-mail: degui.li@york.ac.uk

Department of Statistics, London School of Economics, WC2A 2AE, U.K.

E-mail: q.yao@lse.ac.uk