Nonlinear Regression Estimation Using Subset-based Kernel Principal Components

Yuan Ke
Department of ORFE
Princeton University
Princeton, 08544, U.S.A.

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

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

24 October 2015

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 a response and its 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 construct the estimation of the mean regression function. Under some regularity conditions, the developed estimation 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 also 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 real data sets illustrate the reliable performance of the proposed method. Especially the improvement over the global KPCA method is evident.

Keywords: Conditional distribution function, eigenfunctions, eigenvalues, 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 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 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 *et al.* (2010). When the dimension of the random covariates p is large, a direct use of the nonparametric regression estimation methods such as spline and the 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 a additive model may perform poorly when the data generation process deviates from the additive assumption.

In this paper we propose a data-driven dimension reduction approach through using 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 *et al.* (1999), Braun (2005) and Blanchard *et al.* (2007). See also Section 2.2 below 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 *et al.* 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 the 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 below). The subsets may be defined according to some characteristics of $\mathbf{X}$ and/or those 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 some conventional nonparametric regression techniques such as the kernel smoothing and nearest neighbour methods. Meanwhile, we assume that the observations in the present paper 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 which are 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 which 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 for estimating the 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 associated 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 two real data applications. Section 7 concludes the paper. All the proofs of the theoretical results are provided in an appendix.

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). To make the presentation clear, this section is organized as follows: 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 KPCA and the standard PCA in Section 2.2, and then 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 the following conditions: for any $f \in \mathcal{L}_2(\mathcal{G})$,

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

and

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

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 $\mathcal{G} \times \mathcal{G}$, i.e. $K(\cdot,\cdot)$ is 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 $\mathbf{u} \in \mathcal{G}$, $K(\mathbf{x}, \mathbf{u}) \in \mathcal{L}_2(\mathcal{G})$ can be seen as a function of $\mathbf{x}$. A Mercer kernel $K(\cdot,\cdot)$ defines an operator on $\mathcal{L}_2(\mathcal{G})$ as follows:

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

It follows from Mercer's Theorem (Mercer, 1909) that a Mercer kernel admits the following 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

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

and

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

As we can see from the spectral decomposition (2.2), $d = \max\{k : \lambda_k > 0\}$ and is possible to be 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 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 which is defined in (2.6) below, and then obtain the estimate of the eigenfunction φ_k by the Nyström extension (Drineas and Mahoney, 2005).

Define

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

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

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

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 corresponding m orthonormal eigenvectors. Write

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

By (2.3), (2.6) and the Nyström extension of the eigenvector $\widehat{\boldsymbol{\varphi}}_k$, we may define

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

Proposition 3 in Section 3 below shows that, for any $\mathbf{x} \in \mathcal{G}$, $\widetilde{\lambda}_k$ and $\widetilde{\varphi}_k(\mathbf{x})$ are consistent estimators of λ_k and $\varphi_k(\mathbf{x})$, respectively.

Another critical issue in practical application is to estimate the dimension of the Mercer kernel $K(\cdot,\cdot)$. When the dimension of the kernel $K(\cdot,\cdot)$ is d and $d \ll m$, we may estimate d by the following ratio method (c.f., Lam and Yao, 2012):

$$\widehat{d} = \underset{1\leqslant k\leqslant \lfloor mc_0\rfloor}{\arg\min}\ \widehat{\lambda}_{k+1}/\widehat{\lambda}_k = \underset{1\leqslant k\leqslant \lfloor mc_0\rfloor}{\arg\min}\ \widetilde{\lambda}_{k+1}/\widetilde{\lambda}_k, \tag{2.9}$$

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 indicate that this ratio method works well in finite sample cases.

2.2 Kernel feature space and KPCA

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

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

Then we have $\mathcal{M}(K) \subset \mathcal{L}_2(\mathcal{G})$. By 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 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 give an interpretation to see how the KPCA is connected to the standard PCA.

Any $f \in \mathcal{M}(K)$ whose mean is zero on set $\mathcal{G}$ admits the following expression,

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

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 $K(\cdot,\cdot)$,

$$\mathrm{Var}_K\{f(\mathbf{X})I(\mathbf{X}\in\mathcal{G})\} = \sum_{j=1}^{d} \lambda_j \langle f, \varphi_j\rangle^2, \tag{2.10}$$

where λ_j is assigned as the weight on the "direction" of φ_j for $j = 1, \cdots, d$. 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 (2.10). 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 different from (2.1) to serve as a penalty in estimating functions via regularization; see section 5.8 of Hastie *et al.* (2009) and Wahba (1990). Since the reproducing property is irrelevant in our context, we adopt the 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 *et al.* (2009).

We end this subsection by stating a proposition which shows that the smaller $\mathcal{G}$ is, 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.

PROPOSITION 1. *Let $\mathcal{G}_*$ be a measurable subset of the sample space of $\mathbf{X}$ such that $\mathcal{G} \subset \mathcal{G}_*$, and $K(\cdot, \cdot)$ be a Mercer kernel on $\mathcal{G}_* \times \mathcal{G}_*$. The kernel feature spaces defined with sets $\mathcal{G}$ and $\mathcal{G}_*$ are denoted, respectively by $\mathcal{M}(K)$ and $\mathcal{M}_*(K)$. Furthermore, for any eigenfunctions $\phi_k^*(\cdot)$ on $\mathcal{M}_*(K)$, assume that there exists $\mathbf{x} \in \mathcal{G}$ such that $\phi_k^*(\mathbf{x}) \neq 0$. Then $dim\{\mathcal{M}(K)\} \leqslant dim\,\{\mathcal{M}_*(K)\}$.*

2.3 Estimation for conditional mean regression

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

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

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 $\mathcal{G}$, $h(\mathbf{x}) = \mathsf{E}(Y|\mathbf{X} = \mathbf{x}) \in \mathcal{M}(K)$, i.e. $h(\mathbf{x})$ may be expressed as

$$h(\mathbf{x}) = \int y f_{Y|\mathbf{X}}(y|\mathbf{x})dy = \sum_{k=1}^{d} \beta_k \varphi_k(\mathbf{x}), \quad \mathbf{x} \in \mathcal{G}, \tag{2.11}$$

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 which is constructed as

$$\widetilde{\beta}_k = \frac{1}{m}\sum_{i=1}^{m} Y_i^{\mathcal{G}}\, \widetilde{\varphi}_k(\mathbf{X}_i^{\mathcal{G}}), \quad k = 1, \cdots, d, \tag{2.12}$$

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

$$\widetilde{h}(\mathbf{x}) = \sum_{k=1}^{d} \widetilde{\beta}_k \widetilde{\varphi}_k(\mathbf{x}), \quad \mathbf{x} \in \mathcal{G}. \tag{2.13}$$

When the dimension of the kernel $K(\cdot,\cdot)$ is unknown, the sum on the right hand side of the above expression runs from $j = 1$ to $\widehat{d}$ with $\widehat{d}$ determined via (2.9).

The estimator in (2.13) is derived under the assumption that on set $\mathcal{G}$, $h(\mathbf{x}) \in \mathcal{M}(K)$. When this condition is unfulfilled, (2.13) 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 below, 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 *et al.* 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 $\psi_1, \cdots, \psi_d \in \mathcal{L}_2(\mathcal{G})$,

$$K(\mathbf{u},\mathbf{v}) = \sum_{k=1}^{d} \psi_k(\mathbf{u})\psi_k(\mathbf{v}) \tag{2.14}$$

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, polynomial series, wavelets, B-spline, etc. The numerical studies in Sections 5 and 6 use (2.14) with appropriately chosen functions ψ_k in the estimation and dimension reduction procedure, which performs reasonably well. The following proposition shows 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 (2.14), 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 the mixing coefficient satisfying

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

where p is the dimension of the random covariate, $0 \leqslant \delta_* < \infty$ 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 $\mathcal{G}$. Furthermore, the kernel $K(\cdot,\mathbf{x})$ is Lipschitz continuous on the set $\mathcal{G}$ for any $\mathbf{x} \in \mathcal{G}$.

In Assumption 1, we allow the process to be stationary and α-mixing dependent, which 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. Note that 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_* is 0 if $\mathcal{G}$ is bounded. Assumptions 2 and 3 impose mild restrictions on the eigenvalues and eigenfunctions of the Mercer kernel, respectively. They are crucial to ensure the consistency of the sample eigenvalues and eigenvectors constructed in Section 2.1. The bounded condition on φ_j and $K(\cdot, \mathbf{x})$ in Assumption 3 can be replaced by the $2(2+\delta)$-order moment conditions for some $\delta > 0$. Proposition 3 below still holds at the cost of more lengthy arguments.

PROPOSITION 3. *Suppose that Assumptions 1–3 are satisfied. Then we have*

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

and

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

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

Proposition 3 presents the convergence rates of the estimated eigenvalues and eigenfunctions of Mercer kernel $K(\cdot, \cdot)$. The result is of independent interest. It complements some statistical properties of the KPCA in the literature such as Braun (2005) and Blanchard *et al.* (2007). Note that it is implied by $\mathsf{P}(\mathbf{X} \in \mathcal{G}) > 0$ that m is of the same order as the full sample size n. Hence, the convergence rates in (3.2) and (3.3) are equivalent to $O_P\left(n^{-1/2}\right)$ and $O_P\left(n^{-1/2} \log^{1/2} n\right)$, which are not uncommon in the context of functional principal component analysis (e.g., Bosq, 2000; Horváth and Kokoszka, 2012). Based on Proposition 3, we can easily derive the following uniform consistency result for $\widehat{h}(\cdot)$.

THEOREM 1. *Suppose that Assumptions 1–3 are satisfied,* $\mathsf{E}[|Y|^{2+\delta}] < \infty$ *for some* $\delta > 0$ *and* $h(\cdot) \in \mathcal{M}(K)$. *Then it holds that*

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

where ξ_m *is defined in Proposition 3.*

As stated above, the uniform convergence rate in (3.4) is 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 (c.f., Fan and Yao, 2003), where b is a bandwidth which 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 (2.11) 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 for the conditional distribution function, and the case when the dimension of a kernel feature space diverges together with the sample size.

4.1 Estimation for conditional distribution functions

Estimation of the conditional distribution function defined in (1.2) is a key aspect in various statistical topics (such as the 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 *et al.* (1999), Hansen (2004) and Hall and Yao (2005). In this section, we use the subset-based KPCA approach discussed above 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 $F_*(y|\cdot) \in \mathcal{M}(K)$, i.e.

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

Note that the coefficients β_k^* in the above decomposition depend on y. The orthonormality 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 following estimator for β_k^*:

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

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 is the full sample size. Consequently, we obtain the 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-negative, non-decreasing and between 0 and 1 (e.g., Glad *et al.* 2003).

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

THEOREM 2. *Suppose that Assumptions 1–3 are satisfied and $F_*(y|\cdot) \in \mathcal{M}(K)$. Then it holds that*

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

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

4.2 Kernel feature spaces with diverging dimensions

We next study the case when the dimension of the kernel feature space $d_m \equiv \max\{k : \lambda_k > 0\}$ depends on m. It may diverge to infinity as m tends to infinity. In order to derive a more general asymptotic theory, we need to modify Assumption 2 in Section 3.

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

$$\frac{d_m^2 \log m}{m\rho_m^2\lambda_{d_m}^2} = o(1), \quad \rho_m = \min\left\{\lambda_k - \lambda_{k+1},\ \ k = 1, \cdots, d_m - 1\right\}.$$

The following proposition shows that the diverging d_m would slow down the convergence rates in Proposition 3.

PROPOSITION 4. *Suppose that Assumptions 1, 2* and 3 are satisfied, and the α-mixing coefficient decays to zero at an exponential rate. Then it holds that*

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

and

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

When $m\to\infty$, $d_m\to\infty$ and usually $\rho_m\to 0$ and $\lambda_{d_m}\to 0$. This implies that the convergence rates in (4.6) and (4.7) would be slower than those in (3.2) and (3.3). Let c_i, $i=0,1,\cdots,4$, be positive constants. For any two sequences a_m and b_m, $a_m\propto b_m$ means that $0<c_3\leqslant a_m/b_m\leqslant c_4<\infty$ when m is sufficiently large. If $d_m=c_0\log m$, $\rho_m=c_1\log^{-1}m$ and $\lambda_{d_m}=c_2\log^{-1}m$, we have

$$d_m\xi_m\propto m^{-1/2}\log^{3/2}m,\quad d_m\xi_m/(\rho_m\lambda_{d_m})\propto m^{-1/2}\log^{7/2}m.$$

Using the above proposition and following the proof of Theorem 1 in the appendix, we can easily show the uniform convergence rate for the conditional mean regression estimation and the conditional distribution estimation is of the order $O_P\left(d_m^2\xi_m/(\rho_m\lambda_{d_m})\right)$, which is slower than those in Theorems 1 and 2.

5 Simulation Studies

In this section, we use three simulated examples to illustrate the finite sample performance of the proposed subset-based KPCA method and compare it with some existing estimation methods, i.e. the global KPCA and cubic spline. Throughout this section, the kernel function is either the Gaussian kernel or formulated as in (2.14) with $\{\psi_1(\mathbf{u}),\cdots,\psi_d(\mathbf{u})\}$ being a set of normalized (i.e. with the unit norm) polynomial basis functions of $\mathbf{u}=(u_1,\cdots,u_p)^{\mathrm{T}}$ up to order 2, i.e., $\left\{1,u_k,u_k^2,\ k=1,\cdots,p\right\}$, where p is the dimension of $\mathbf{u}$. For the latter case, we have $d=2p+1$ and call the kernel as *the quadratic kernel* for the simplicity of presentation. In practice, d is estimated by the ratio method as in (2.9). The simulation results shows (2.9) 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 bandwidth κ and the tuning parameter c in the Gaussian kernel are selected by a 10-fold cross validation.

We start with an example to assess the out-of-sample estimation performance of conditional mean regression function based on a multivariate nonlinear regression model. Then, in the second example, we examine the one-step ahead out-of-sample forecast performance based on a multivariate nonlinear time series. Finally, in the third example, we examine the finite sample performance of the estimation of conditional distribution function.

EXAMPLE 5.1. Consider the following model:

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

where $x_{k1}, \cdots, x_{k6}$ and ε_i are independently and $\mathsf{N}(0,1)$, and $g(x) = \mathrm{e}^{-2x^2}$ 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 draw a training sample of size n and a test sample of size 200. We estimate the regression function (1.1) using the training sample, and then calculate the mean squared errors over the testing sample as follows:

$$\text{MSE} = \frac{1}{200}\sum_{i=1}^{200}\left[\widehat{h}(\mathbf{x}_i) - y_i\right]^2. \tag{5.1}$$

By repeating this procedure over 200 replications, we obtain a sample of MSE with size 200. The estimation performance is assessed by the sample mean, median and variance of MSE. The size of the training sample n is set to be 500 or 1000. The simulation results are reported in Table 1. In this simulation, for the quadratic kernel, the ratio method in (2.9) can always correctly estimate $\widehat{d} = 13$. According to the results in Table 1, the subset-based KPCA with the quadratic kernel outperforms the other methods as it has smaller sample mean, median and variance of MSE. In addition, the quadratic kernel performs better than the Gaussian kernel due to the fact that the quadratic kernel captures different degree of smoothness on different directions.

EXAMPLE 5.2. Consider the following 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 $\{\epsilon_t\}$ is a sequence of independent $\mathsf{N}(0,1)$ random variables. We want to estimate the conditional mean $\mathsf{E}(y_t|y_{t-1}, \cdots y_{t-4})$ and denote the estimator as $\widehat{y}_t$. Note $\widehat{y}_t$ is a one-step-ahead predictor for y_t.

We generate a time series from the above model with the length $n + 100$. For each $k = 1, \cdots, 100$, we predict the value of y_{n+k} by $\widehat{y}_{n+k}$ which is estimated based on the n observations

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

	$n = 500$			$n = 1000$		
MSE	Mean	Median	Variance	Mean	Median	Variance
sKPCA+Quadratic	1.3002	1.2941	0.0169	1.2438	1.2365	0.0137
sKPCA+Gaussian	1.5729	1.5855	0.0259	1.5310	1.5282	0.0253
gKPCA+Gaussian	3.0228	2.0214	0.0933	3.0151	2.9900	0.0859
Cubic spline	1.3864	1.3828	0.0181	1.3707	1.3720	0.0052

"sKPCA+Quadratic" stands for the subset-based KPCA with the quadratic kernel; "sKPCA+Gaussian" stands for the subset-based KPCA with the Gaussian kernel; "gKPCA+Gaussian" stands for the global KPCA with the Gaussian kernel.

$y_k, \cdots, y_{n+k-1}$. The performance is measured by the mean squared prediction error (MSPE) and mean relative prediction error (MRPE) defined as:

$$\text{MSPE} = \frac{1}{100}\sum_{k=1}^{100}(\widehat{y}_{n+k} - y_{n+k})^2, \qquad \text{MRPE} = \frac{1}{100}\sum_{k=1}^{100}\left|\frac{\widehat{y}_{n+k} - y_{n+k}}{y_{n+k}}\right|.$$

We set $n = 500$, and repeat the experiment 200 times, leading to a sample of MSPE and a sample of MRPE with size 200 for each of the estimation methods. The sample means, medians and variances of MSPE and MRPE are presented in Table 2. Similar to Example 5.1, the subset-based KPCA method with the quadratic kernel provides the most accurate forecasting with the cubic spline as a close second best in terms of MSPE. Judging by MRPE, the subset-based KPCA method with the quadratic kernel still outperforms the other three methods. But the cubic spline method is no longer attractive as its mean MRPE is greater than that of both the subset-based KPCA method with the Gaussian kernel and the global KPCA method. Figure 1 plots a typical path together with their one-step-ahead forecasts for each of the four methods. The typical path is the one with its MSPE equal to the sample median. Figure 1 indicates that the forecasted path from the subset-based KPCA method with the quadratic kernel follows the true path closely. This is also true, to some extent, for the subset-based KPCA method with the Gaussian kernel and the cubic spline method. However the global KPCA method fails to capture the dynamic variation of the series and tends to forecast the future values by the overall mean value, which is not satisfactory.

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

	MSPE			MRPE		
	Mean	Median	Variance	Mean	Median	Variance
sKPCA+Quadratic	0.0435	0.0428	3.9×10^{-5}	0.2192	0.2162	3.8×10^{-4}
sKPCA+Gaussian	0.0756	0.0751	4.1×10^{-4}	0.2529	0.2527	0.0011
gKPCA+Gaussian	0.2172	0.2211	0.0038	0.3889	0.3901	0.0028
Cubic spline	0.0455	0.0443	0.0049	0.6712	0.4797	0.0325

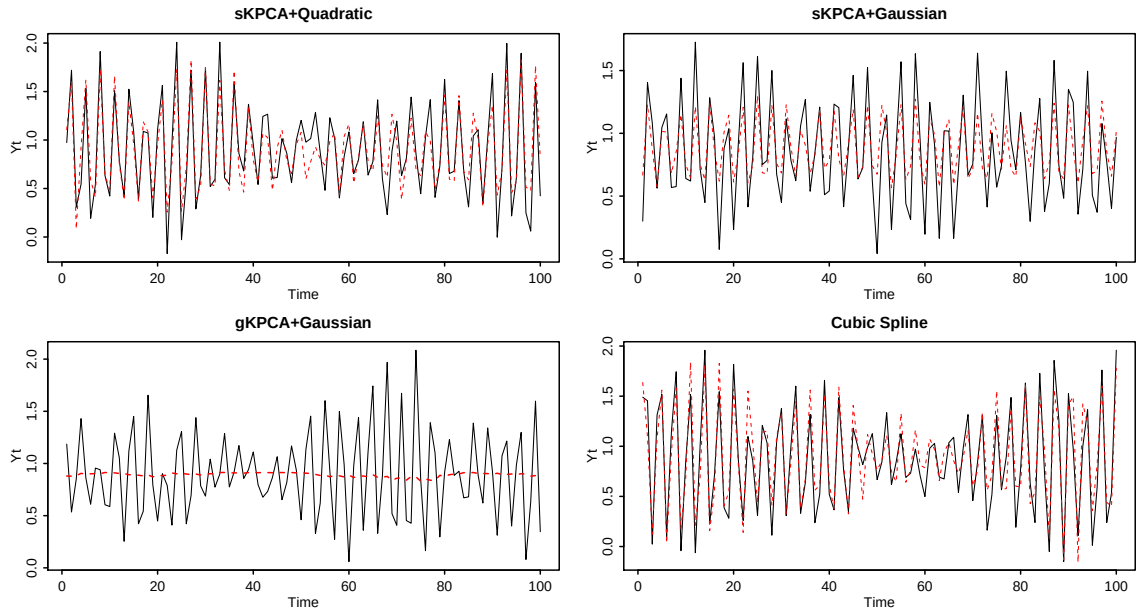

Figure 1: One-step ahead out-of-sample forecasting performance based on the replication with median MSPE for each method. The black solid line is the true value and the red dashed line is the predicted value.

Example 5.3. Consider now the model:

$$X_1 \sim \mathsf{N}(0,\ 1), \quad X_2 \sim \mathsf{N}(0,\ 1),$$
$$Y|(X_1, X_2) \sim \mathsf{N}(X_1,\ 1 + X_2^2),$$

Now 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 $1 + X_2^2$. The aim is to estimate the conditional distribution function $F_{Y|\mathbf{X}}(y|\mathbf{x})$ based on the method proposed in Section 4.1.

We draw a training sample of size n and a test sample of size 100. The estimated condi-

tional distribution $\widetilde{F}_{Y|\mathbf{X}}(y_i|\mathbf{x}_i)$ is obtained using the training data. We check the performance by calculating the mean squared errors over the test sample as follows:

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

By repeating this experiment 200 times, we obtain a sample of MSE of size 200. Table 3 lists the sample means, medians and variances of the MSE for $n = 300$ and $n = 500$. Also reported in Table 3 is the largest absolute error (LAE):

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

where Ω^* is the union of all validation sets. As those values of LAE are very small, the proposed method provides very accurate estimation for the conditional distribution functions.

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 Real data analysis

In this section, we apply the proposed subset-based KPCA method to two real data examples. The kernel functions and the choices for the subsets and the tuning parameter are specified in the same manner as in Section 5.

6.1 Circulatory and respiratory problem in Hong Kong

We study the circulatory and respiratory problem in Hong Kong via an environmental data set. This data set contains 730 observations and was 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 standardize the data so that all covariates have zero sample mean and unit sample variance.

The objective of this study is to estimate the number of daily total hospital admissions for circulatory and respiratory problem using the collected environmental data, i.e. estimate the conditional mean regression function. For a given observation $(y, \mathbf{x})$, we define the relative estimation error (REE) as

$$\text{REE} = \left| \frac{\widehat{\xi} - y}{y} \right|,$$

where $\widehat{\xi}$ is the estimator of the conditional expectation of y given $\mathbf{x}$. In this study, the estimation performance is measured by the mean and variance of the REE, which are calculated by a bootstrap method described as follows. We first randomly divide the data set into a training set of 700 observations and a test set of 30 observations. Then for each observation in the test set, we use the training set to estimate the conditional mean regression function and calculate the REE. By repeating this re-sampling and estimation procedure 1,000 times, we obtain a bootstrap sample of REEs with size 30,000. The sample mean and variance are used as the mean and variance of the REE.

We compare the performances of the three methods: the subset-based KPCA with the quadratic kernel, the subset-based KPCA with the Gaussian kernel and the global KPCA with the Gaussian kernel. The results are presented in Table 4. According to the results in Table 4, the sKPCA with the quadratic kernel has the best estimation performance. The subset-based KPCA method outperforms the global KPCA method as the latter has the largest mean and variance of REE.

Table 4: Estimation performance for the Hong Kong environmental data

	REE	
Method	Mean	Variance
sKPCA + Quadratic	0.1601	5.9×10^{-4}
sKPCA + Gaussian	0.1856	7.7×10^{-4}
gKPCA + Gaussian	0.3503	1.9×10^{-3}

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 by the urban customers. The CPI is often used as an important economic indicator in macroeconomic and financial studies. For example, in economics, CPI is considered as closely related to the cost-of-living index and used to adjust the income eligibility levels for government assistance. In finance, 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 perform one-step-ahead forecasting for the monthly log return of CPI in USA based on the proposed subset-based KPCA method with the quadratic kernel. The data concerned are collected for the period from January 1970 to December 2014 with the total 540 observations. Instead of using the traditional linear time series models, we consider the log return of CPI follows a 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. For a comparison purpose, we also forecast y_t based on a linear AR(p) model with the order p determined by AIC. Suppose the forecast period starts form time t and ends at time $t+S$, the forecast error is measured by the mean squared error (MSE) defined as

$$\text{MSE} = \frac{1}{S}\sum_{s=1}^{S}(\widehat{y}_{t+s} - y_{t+s})^2,$$

where $\widehat{y}_{t+s}$ is the estimator of y at time $t+s$.

For each of the 120 months in the period of January 2005 – December 2014, we forecast its log return based on the models fitted using the data up to its previous month. The MSPE are calculated over the 120 months. The MSE of the subset-based KPCA method based on the nonlinear AR(3) model is 2.9×10^{-6} while the MSPE of the linear AR model is 1.5×10^{-5}. The detailed forecast results are plotted in Figure 2, which shows clearly that the forecast based on the subset-based KPCA method is more accurate as it captures the local variations much better than the linear AR modelling method.

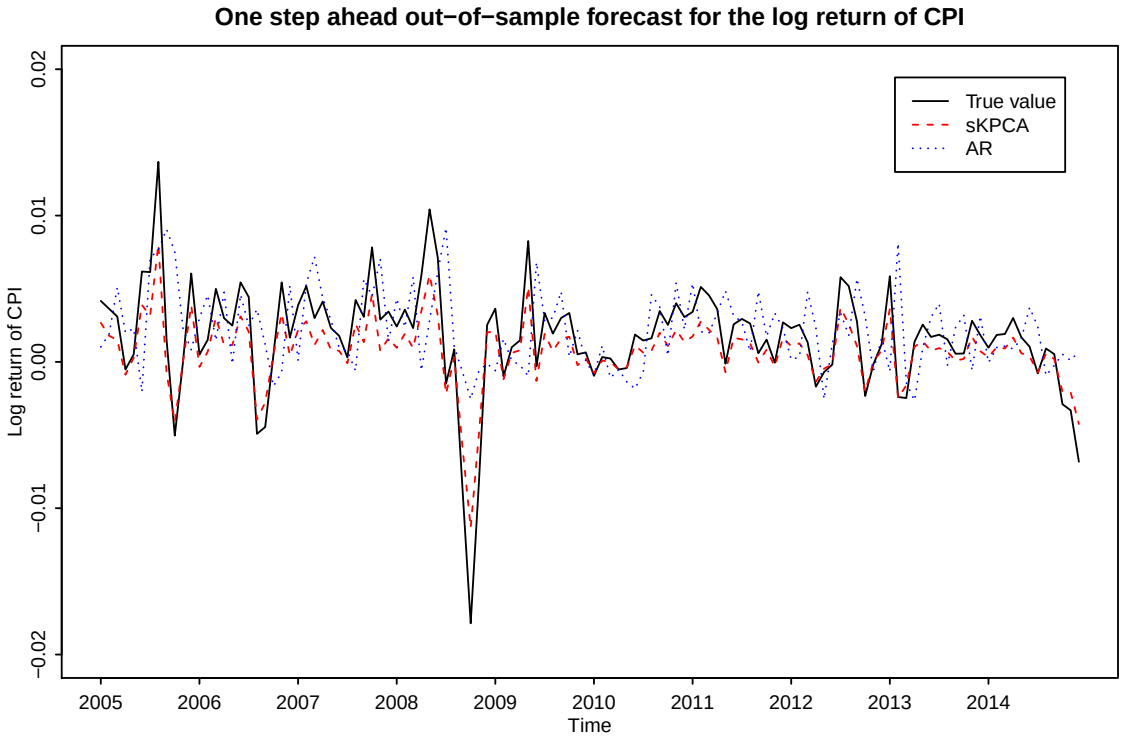

Figure 2: One step ahead out-of-sample forecast for the log return of CPI from January 2005 to December 2014. The black solid line is the true value, the red dashed line is the forecast value obtained by the subset-based KPCA, and the blue dotted line is the forecast value obtained by the linear AR model.

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 simulated and real data sets illustrate clearly the advantages of using the subset-based KPCA method over its global counterpart. It also outperforms some popular nonparametric regression methods such as cubic spline and kernel regression. (The results on kernel regression are not reported to save

the space.) It is also worth mentioning that the quadratic kernel constructed based on (2.14) using normalized univariate linear and quadratic basis functions performs better than the more conventional Gaussian kernel for all the examples reported in Sections 5 and 6.

Appendix: Proofs of the theoretical results

This appendix provides the detailed proofs of the theoretical results given in Sections 2 and 3. We start with the proofs of Propositions 1 and 2.

Proof of Proposition 1. By Mercer's Theorem, for $\mathbf{u}, \mathbf{v} \in \mathcal{G}_*$, the kernel function has the following spectral decomposition:

$$K(\mathbf{u}, \mathbf{v}) = \sum_{k=1}^{d_*} \lambda_k^* \varphi_k^*(\mathbf{u}) \varphi_k^*(\mathbf{v}), \tag{A.1}$$

where $\lambda_1^* \geqslant \lambda_2^* \geqslant \cdots \geqslant 0$ are the eigenvalues of $K(\cdot, \cdot)$ on set $\mathcal{G}_*$, $\varphi_1^*, \varphi_2^*, \cdots$ are the corresponding orthonormal eigenfunctions, and $d_* = \max\{k : \lambda_k^* > 0\} = \dim\{\mathcal{M}_*(K)\}$. Recall that $\{\lambda_k, \varphi_k\}$, $k = 1, \cdots, d$, are pairs of eigenvalues and eigenfunctions of $K(\cdot, \cdot)$ on set $\mathcal{G}$ with $d = \dim\{\mathcal{M}(K)\}$. Hence, we next only need to show that $d \leqslant d_*$. Note that for any $k = 1, \cdots, d$,

$$\varphi_k^*(\mathbf{x}) = \sum_{j=1}^{d} a_{kj} \varphi_j(\mathbf{x}), \quad \mathbf{x} \in \mathcal{G}, \tag{A.2}$$

where $a_{kj} = \langle \varphi_k^*, \varphi_j \rangle = \int_{\mathcal{G}} \varphi_k^*(\mathbf{x}) \varphi_j(\mathbf{x}) \mathsf{P}_{\mathbf{X}}(d\mathbf{x})$. By the assumption in the proposition, we may show that at least one of a_{kj} is non-zero for $j = 1, \cdots, d$; otherwise $\varphi_k^*(\mathbf{x}) = 0$ for any $\mathbf{x} \in \mathcal{G}$. In view of (2.2)–(2.4), (A.1) and (A.2), we may show that, for any $k = 1, \cdots, d$,

$$\begin{aligned}
\lambda_k^* &= \int_{\mathcal{G}_* \times \mathcal{G}_*} \varphi_k^*(\mathbf{u}) K(\mathbf{u}, \mathbf{v}) \varphi_k^*(\mathbf{v}) \mathsf{P}_{\mathbf{X}}(d\mathbf{v}) \mathsf{P}_{\mathbf{X}}(d\mathbf{u}) \\
&= \int_{\mathcal{G} \times \mathcal{G}} \varphi_k^*(\mathbf{u}) K(\mathbf{u}, \mathbf{v}) \varphi_k^*(\mathbf{v}) \mathsf{P}_{\mathbf{X}}(d\mathbf{v}) \mathsf{P}_{\mathbf{X}}(d\mathbf{u}) + \\
&\quad \int_{\mathcal{G}_* \times \mathcal{G}_* - \mathcal{G} \times \mathcal{G}} \varphi_k^*(\mathbf{u}) K(\mathbf{u}, \mathbf{v}) \varphi_k^*(\mathbf{v}) \mathsf{P}_{\mathbf{X}}(d\mathbf{v}) \mathsf{P}_{\mathbf{X}}(d\mathbf{u}) \\
&\geqslant \int_{\mathcal{G} \times \mathcal{G}} \varphi_k^*(\mathbf{u}) K(\mathbf{u}, \mathbf{v}) \varphi_k^*(\mathbf{v}) \mathsf{P}_{\mathbf{X}}(d\mathbf{v}) \mathsf{P}_{\mathbf{X}}(d\mathbf{u}) \\
&= \sum_{j=1}^{d} a_{kj}^2 \lambda_j > 0,
\end{aligned}$$

where the first inequality holds as the kernel function is non-negative definite and the last strict inequality holds as $\lambda_j > 0$, $j = 1, \cdots, d$ and at least one of a_{kj}, $j = 1, \cdots, d$, is non-zero. Hence, we can prove that $d_* = \max\{k : \lambda_k^* > 0\} \geqslant d = \max\{k : \lambda_k > 0\}$, which indicates that $\dim\{\mathcal{M}_*(K)\} \geqslant \dim\{\mathcal{M}(K)\}$ and completes the proof of Proposition 1. ■

PROOF OF PROPOSITION 2. Let $\lambda_1^\diamond, \lambda_2^\diamond, \cdots \lambda_{d^\diamond}^\diamond$ be the eigenvalues of the Mercer kernel $K(\cdot,\cdot)$ defined in (2.14) and let $\varphi_1^\diamond, \varphi_2^\diamond, \cdots, \varphi_{d^\diamond}^\diamond$ be the corresponding orthonormal eigenfunctions. Then, by Mercer's Theorem, we have

$$K(\mathbf{u},\mathbf{v}) = \sum_{j=1}^{d^\diamond} \lambda_j^\diamond \varphi_j^\diamond(\mathbf{u})\varphi_j^\diamond(\mathbf{v}), \quad \mathbf{u},\mathbf{v} \in \mathcal{G}, \tag{A.3}$$

in which $\lambda_1^\diamond \geqslant \lambda_2^\diamond \geqslant \cdots \geqslant \lambda_{d^\diamond}^\diamond > 0$. We next show that the conclusion of $d^\diamond > d$ would lead to a contradiction. For $\psi_j(\cdot)$, $j = 1, \cdots, d$, we may show that $\psi_j(\mathbf{x}) = \sum_{k=1}^{d^\diamond} a_{jk}^\diamond \varphi_k^\diamond(\mathbf{x})$, $\mathbf{x} \in \mathcal{G}$, where $a_{jk}^\diamond = \langle \varphi_k^\diamond, \psi_j \rangle$. Let $\mathbf{A}$ be a $d \times d^\diamond$ matrix with the (j,k)-entry being $a_{jk}^\diamond$,

$$\boldsymbol{\psi}(\cdot) = [\psi_1(\cdot), \cdots, \psi_d(\cdot)]^{\mathrm{T}}, \quad \boldsymbol{\varphi}^\diamond(\cdot) = [\varphi_1^\diamond(\cdot), \cdots, \varphi_{d^\diamond}^\diamond(\cdot)]^{\mathrm{T}}.$$

Then, we have

$$\boldsymbol{\psi}(\mathbf{x}) = \mathbf{A}\boldsymbol{\varphi}^\diamond(\mathbf{x}), \quad \mathbf{x} \in \mathcal{G}, \tag{A.4}$$

and the rank of $\mathbf{A}$ is strictly smaller than $d^\diamond$ when $d < d^\diamond$. However, by (A.4) and the definition of $K(\cdot,\cdot)$ in (2.14),

$$K(\mathbf{u},\mathbf{v}) = \boldsymbol{\psi}(\mathbf{u})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{v}) = \boldsymbol{\varphi}^\diamond(\mathbf{u})^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{\varphi}^\diamond(\mathbf{v}),$$

which, together with (A.3), indicates that $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ has $d^\diamond$ positive eigenvalues. Hence, it is not possible to conclude that $d^\diamond > d$, which completes the proof of Proposition 2. ■

To prove Proposition 3 in Section 3, we need to make use of the following technical lemma on uniform consistency.

LEMMA 1. *Suppose that Assumptions 1–3 are satisfied. Then we have*

$$\max_{1 \leqslant k \leqslant d} \sup_{\mathbf{x} \in \mathcal{G}} \left| \frac{1}{m} \sum_{i=1}^{m} K(\mathbf{x}, \mathbf{X}_i^{\mathcal{G}}) \varphi_k(\mathbf{X}_i^{\mathcal{G}}) - \lambda_k \varphi_k(\mathbf{x}) \right| = O_P(\xi_m). \tag{A.5}$$

where $\xi_m = \sqrt{(\log m)/m}$

PROOF. To simplify the presentation, we let $Z_{ik}(\mathbf{x}) = K(\mathbf{x}, \mathbf{X}_i^{\mathcal{G}})\varphi_k(\mathbf{X}_i^{\mathcal{G}})$. By (2.3), it is easy to verify that $\mathsf{E}[Z_{ik}(\mathbf{x})] = \lambda_k \varphi_k(\mathbf{x})$ for any $1 \leqslant k \leqslant d$ and $\mathbf{x} \in \mathcal{G}$. The proof of (A.5) is standard by using the finite covering techniques. We consider covering the set $\mathcal{G}$ by a finite number of subsets $\mathcal{G}_j$ which are centered at $\mathbf{c}_j$ with radius ξ_m. Letting N_m be the total number of these subsets, $N_m = O(m^{\delta *}/\xi_m^p)$ which is diverging with m.

Observe that

$$\begin{aligned}
& \max_{1\leqslant k\leqslant d}\sup_{\mathbf{x}\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^{m}K(\mathbf{x},\mathbf{X}_i^{\mathcal{G}})\varphi_k(\mathbf{X}_i^{\mathcal{G}})-\lambda_k\varphi_k(\mathbf{x})\right| \\
= & \max_{1\leqslant k\leqslant d}\sup_{\mathbf{x}\in\mathcal{G}}\left|\frac{1}{m}\sum_{i=1}^{m}\{Z_{ik}(\mathbf{x})-\mathsf{E}\left[Z_{ik}(\mathbf{x})\right]\}\right| \\
\leqslant & \max_{1\leqslant k\leqslant d}\max_{1\leqslant j\leqslant N_m}\left|\frac{1}{m}\sum_{i=1}^{m}\{Z_{ik}(\mathbf{c}_j)-\mathsf{E}\left[Z_{ik}(\mathbf{c}_j)\right]\}\right| + \\
& \max_{1\leqslant k\leqslant d}\max_{1\leqslant j\leqslant N_m}\sup_{\mathbf{x}\in\mathcal{G}_j}\left|\frac{1}{m}\sum_{i=1}^{m}\left[Z_{ik}(\mathbf{x})-Z_{ik}(\mathbf{c}_j)\right]\right| + \\
& \max_{1\leqslant k\leqslant d}\max_{1\leqslant j\leqslant N_m}\sup_{\mathbf{x}\in\mathcal{G}_j}\left|\lambda_k\left[\varphi_k(\mathbf{x})-\varphi_k(\mathbf{c}_j)\right]\right| \\
\equiv & \Pi_m(1)+\Pi_m(2)+\Pi_m(3). \qquad \text{(A.6)}
\end{aligned}$$

By the Lipschitz continuity in Assumption 2, we readily have

$$\Pi_m(2)+\Pi_m(3)=O_P(\xi_m). \qquad \text{(A.7)}$$

Therefore, to complete the proof of (A.5), we only need to show

$$\Pi_m(1)=O_P(\xi_m). \qquad \text{(A.8)}$$

Using the exponential inequality for the α-mixing sequence (e.g., Theorem 2.18 (ii) in Fan and Yao, 2003), we may show that

$$\begin{aligned}
\mathsf{P}\left\{\Pi_m(1)>C_1\xi_m\right\} &= \mathsf{P}\left\{\max_{1\leqslant k\leqslant d}\max_{1\leqslant j\leqslant N_m}\left|\frac{1}{m}\sum_{i=1}^{m}\{Z_{ik}(\mathbf{c}_j)-\mathsf{E}\left[Z_{ik}(\mathbf{c}_j)\right]\}\right|>C_1\xi_m\right\} \\
&\leqslant \sum_{k=1}^{d}\sum_{j=1}^{N_m}\mathsf{P}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\{Z_{ik}(\mathbf{c}_j)-\mathsf{E}\left[Z_{ik}(\mathbf{c}_j)\right]\}\right|>C_1\xi_m\right\} \\
&\leqslant O_P\left(N_m\exp\{-C_1\log m\}+N_m q^{\kappa+3/2}m^{-\kappa}\right),
\end{aligned}$$

where C_1 is a positive constant which can be sufficiently large and $q=\lfloor m^{1/2}\log^{1/2}m\rfloor$. Then, by (3.1), we may show that

$$\mathsf{P}\left\{\Pi_m(1)>C_1\xi_m\right\}=o(1), \qquad \text{(A.9)}$$

which completes the proof of (A.8). The proof of Lemma 1 has been completed. ■

We next prove the asymptotic theorems in Section 3.

Proof of Proposition 3. The proof of (3.2) is a generalization of the argument in the proof of Theorem 3.65 in Braun (2005) from the independence and identical distribution assumption to the stationary and α-mixing dependence assumption.

By the spectral decomposition (2.2), the (i,j)-entry of the $m \times m$ kernel Gram matrix can be written as

$$K(\mathbf{X}_i^{\mathcal{G}}, \mathbf{X}_j^{\mathcal{G}}) = \sum_{k=1}^{d} \lambda_k \varphi_k(\mathbf{X}_i^{\mathcal{G}}) \varphi_k(\mathbf{X}_j^{\mathcal{G}}). \tag{A.10}$$

Therefore, the kernel Gram matrix $\mathbf{K}_{\mathcal{G}}$ can be expressed as

$$\mathbf{K}_{\mathcal{G}} = \mathbf{\Phi}_m \mathbf{\Lambda}_m \mathbf{\Phi}_m^{\mathrm{T}} \tag{A.11}$$

with $\mathbf{\Lambda}_m = \mathsf{diag}(m\lambda_1, \cdots, m\lambda_d)$ and

$$\mathbf{\Phi}_m = \frac{1}{\sqrt{m}} \begin{bmatrix} \varphi_1(\mathbf{X}_1^{\mathcal{G}}) & \cdots & \varphi_d(\mathbf{X}_1^{\mathcal{G}}) \\ \vdots & \ddots & \vdots \\ \varphi_1(\mathbf{X}_m^{\mathcal{G}}) & \cdots & \varphi_d(\mathbf{X}_m^{\mathcal{G}}) \end{bmatrix}.$$

Then, using the Ostrowski's Theorem (e.g., Theorem 4.5.9 and Corollary 4.5.11 in Horn and Johnson, 1985; or Corollary 3.59 in Braun, 2005), we may show that

$$\max_{1 \leqslant k \leqslant d} \left| \widehat{\lambda}_k - m\lambda_k \right| \leqslant \max_{1 \leqslant k \leqslant d} |m\lambda_k| \cdot \left\| \mathbf{\Phi}_m^{\mathrm{T}} \mathbf{\Phi}_m - \mathbf{I}_d \right\|, \tag{A.12}$$

where $\mathbf{I}_d$ is a $d \times d$ identity matrix, and for a $d \times d$ matrix $\mathbf{M}$

$$\|\mathbf{M}\| = \sup_{\mathbf{x} \in \mathcal{R}^d, \|\mathbf{x}\|=1} \|\mathbf{M}\mathbf{x}\|.$$

By (A.14) and Assumption 2, we readily have

$$\max_{1 \leqslant k \leqslant d} \left| \frac{1}{m} \widehat{\lambda}_k - \lambda_k \right| \leqslant O_P\left(\left\| \mathbf{\Phi}_m^{\mathrm{T}} \mathbf{\Phi}_m - \mathbf{I}_d \right\| \right). \tag{A.13}$$

When d is fixed, by (2.4), Assumptions 1 and 3 as well as Theorem A.5 in Hall and Heyde (1980), we can prove that

$$\left\| \mathbf{\Phi}_m^{\mathrm{T}} \mathbf{\Phi}_m - \mathbf{I}_d \right\| = O_P\left(m^{-1/2} \right), \tag{A.14}$$

which together with (A.13), completes the proof of (3.2).

We next turn to the proof of (3.3), which can be seen as a modification of the proof of Lemma 4.3 in Bosq (2000). By Lemma 1 and (3.2), we may show that

$$\begin{aligned} \max_{1 \leqslant k \leqslant d} \left\| \frac{1}{m} \mathbf{K}_{\mathcal{G}} \boldsymbol{\varphi}_k - \widetilde{\lambda}_k \boldsymbol{\varphi}_k \right\| &\leqslant \max_{1 \leqslant k \leqslant d} \left\| \frac{1}{m} \mathbf{K}_{\mathcal{G}} \boldsymbol{\varphi}_k - \lambda_k \boldsymbol{\varphi}_k \right\| + \max_{1 \leqslant k \leqslant d} \left\| \lambda_k \boldsymbol{\varphi}_k - \widetilde{\lambda}_k \boldsymbol{\varphi}_k \right\| \\ &= O_P\left(\xi_m + m^{-1/2} \right) = O_P(\xi_m), \end{aligned} \tag{A.15}$$

where $\boldsymbol{\varphi}_k = \frac{1}{\sqrt{m}}\left[\varphi_k(\mathbf{X}_1^{\mathcal{G}}), \cdots, \varphi_k(\mathbf{X}_m^{\mathcal{G}})\right]^{\mathrm{T}}$ and $\xi_m = \sqrt{(\log m)/m}$. On the other hand, note that, for any $1 \leqslant k \leqslant d$,

$$\begin{aligned}\left\|\frac{1}{m}\mathbf{K}_{\mathcal{G}}\boldsymbol{\varphi}_k - \widetilde{\lambda}_k\boldsymbol{\varphi}_k\right\|^2 &= \sum_{j=1}^{m}\left\|\langle\tfrac{1}{m}\mathbf{K}_{\mathcal{G}}\boldsymbol{\varphi}_k, \widehat{\boldsymbol{\varphi}}_j\rangle - \widetilde{\lambda}_k\langle\boldsymbol{\varphi}_k, \widehat{\boldsymbol{\varphi}}_j\rangle\right\|^2 \\ &\geqslant (1+o_P(1))\cdot\min_{j\neq k}|\lambda_j-\lambda_k|^2\cdot\sum_{j=1,\neq k}^{m}\delta_{kj}^2, \end{aligned} \tag{A.16}$$

where $\delta_{kj} = \langle\boldsymbol{\varphi}_k, \widehat{\boldsymbol{\varphi}}_j\rangle$. By (A.15), (A.16) and Assumption 2, we readily have

$$\max_{1\leqslant k\leqslant d}\Delta_k^2 \equiv \max_{1\leqslant k\leqslant d}\sum_{j=1,\neq k}^{m}\delta_{kj}^2 = O_P(\xi_m^2), \tag{A.17}$$

where $\Delta_k^2 = \sum_{j=1,\neq k}^{m}\delta_{kj}$.

For any $1 \leqslant k \leqslant d$, we may write $\boldsymbol{\varphi}_k$ as

$$\boldsymbol{\varphi}_k = \sqrt{\|\boldsymbol{\varphi}_k\|^2 - \Delta_k^2}\cdot\widehat{\boldsymbol{\varphi}}_k + \sum_{j=1,\neq k}^{m}\delta_{kj}\cdot\widehat{\boldsymbol{\varphi}}_j. \tag{A.18}$$

In view of (2.8), (3.2) and (A.18), we have

$$\begin{aligned}\widetilde{\varphi}_k(\mathbf{x}) &= \frac{1}{\widehat{\lambda}_k}\sum_{i=1}^{m}K(\mathbf{x},\mathbf{X}_i^{\mathcal{G}})\varphi_k(\mathbf{X}_i^{\mathcal{G}}) + \frac{\sqrt{m}}{\widehat{\lambda}_k}\sum_{i=1}^{m}K(\mathbf{x},\mathbf{X}_i^{\mathcal{G}})\left[\widehat{\varphi}_k(\mathbf{X}_i^{\mathcal{G}}) - \frac{1}{\sqrt{m}}\boldsymbol{\varphi}_k(\mathbf{X}_i^{\mathcal{G}})\right] \\ &= \frac{1}{\lambda_k}\cdot\frac{1}{m}\sum_{i=1}^{m}K(\mathbf{x},\mathbf{X}_i^{\mathcal{G}})\varphi_k(\mathbf{X}_i^{\mathcal{G}}) + \left(1-\sqrt{\|\boldsymbol{\varphi}_k\|^2-\Delta_k^2}\right)\cdot\frac{\sqrt{m}}{\widehat{\lambda}_k}\sum_{i=1}^{m}K(\mathbf{x},\mathbf{X}_i^{\mathcal{G}})\widehat{\varphi}_k(\mathbf{X}_i^{\mathcal{G}}) + \\ &\quad \sum_{j=1,\neq k}^{m}\delta_{kj}\cdot\frac{\sqrt{m}}{\widehat{\lambda}_k}\sum_{i=1}^{m}K(\mathbf{x},\mathbf{X}_i^{\mathcal{G}})\widehat{\varphi}_j(\mathbf{X}_i^{\mathcal{G}}) + O_P\left(m^{-1/2}\right) \\ &\equiv \Xi_k(1)+\Xi_k(2)+\Xi_k(3)+O_P\left(m^{-1/2}\right). \end{aligned} \tag{A.19}$$

By Lemma 1, (3.2) and some standard arguments, we have

$$\max_{1\leqslant k\leqslant d}\sup_{\mathbf{x}\in\mathcal{G}}\left|\frac{1}{\lambda_k}\cdot\frac{1}{m}\sum_{i=1}^{m}K(\mathbf{x},\mathbf{X}_i^{\mathcal{G}})\varphi_k(\mathbf{X}_i^{\mathcal{G}}) - \varphi_k(\mathbf{x})\right| = O_P(\xi_m) \tag{A.20}$$

and

$$\begin{aligned}\max_{1\leqslant k\leqslant d}\sup_{\mathbf{x}\in\mathcal{G}}|\Xi_k(2)| &= \max_{1\leqslant k\leqslant d}\sup_{\mathbf{x}\in\mathcal{G}}\left|(1+o_P(1))\left(1-\sqrt{\|\boldsymbol{\varphi}_k\|^2-\Delta_k^2}\right)\cdot\frac{1}{\lambda_k\sqrt{m}}\sum_{i=1}^{m}K(\mathbf{x},\mathbf{X}_i^{\mathcal{G}})\widehat{\varphi}_k(\mathbf{X}_i^{\mathcal{G}})\right| \\ &= O_P\left(\max_{1\leqslant k\leqslant d}\left|1-\sqrt{\|\boldsymbol{\varphi}_k\|^2-\Delta_k^2}\right|\sup_{\mathbf{x}\in\mathcal{G}}\left[\frac{1}{m}\sum_{i=1}^{m}K^2(\mathbf{x},\mathbf{X}_i^{\mathcal{G}})\right]^{1/2}\right) \\ &= O_P\left(\xi_m\right). \end{aligned} \tag{A.21}$$

By the spectral decomposition (2.2), (3.2), (A.17) and the Cauchy-Schwarz inequality, we may show that, uniformly for $1 \leqslant k \leqslant d$ and $\mathbf{x} \in \mathcal{G}$,

$$\begin{aligned}
\Xi_k(3) &= (1+o_P(1)) \cdot \frac{1}{\lambda_k} \sum_{j=1,\neq k}^{m} \frac{\delta_{kj}}{\sqrt{m}} \sum_{i=1}^{m} K(\mathbf{x}, \mathbf{X}_i^{\mathcal{G}}) \widehat{\varphi}_j(\mathbf{X}_i^{\mathcal{G}}) \\
&= (1+o_P(1)) \cdot \frac{1}{\lambda_k} \sum_{j=1,\neq k}^{m} \frac{\delta_{kj}}{\sqrt{m}} \sum_{i=1}^{m} \left[\sum_{l=1}^{d} \lambda_l \varphi_l(\mathbf{x}) \varphi_l(\mathbf{X}_i^{\mathcal{G}}) \right] \widehat{\varphi}_j(\mathbf{X}_i^{\mathcal{G}}) \\
&= (1+o_P(1)) \cdot \left[\varphi_k(\mathbf{x}) \sum_{j=1,\neq k}^{m} \delta_{kj}^2 + \sum_{l=1,\neq k}^{d} \frac{\lambda_l}{\lambda_k} \varphi_l(\mathbf{x}) \delta_{kl} \delta_{ll} + \sum_{l=1,\neq k}^{d} \frac{\lambda_l}{\lambda_k} \varphi_l(\mathbf{x}) \sum_{j=1,\neq k,\neq l}^{m} \delta_{kj} \delta_{lj} \right] \\
&= O_P(\xi_m^2 + \xi_m) = O_P(\xi_m). \qquad \text{(A.22)}
\end{aligned}$$

Then we complete the proof of (3.3) in view of (A.19)–(A.22). The proof of Proposition 3 has been completed. ■

Proof of Theorem 1. Observe that

$$\begin{aligned}
\widetilde{h}(\mathbf{x}) - h(\mathbf{x}) &= \sum_{k=1}^{d} \widetilde{\beta}_k \widetilde{\varphi}_k(\mathbf{x}) - \sum_{k=1}^{d} \beta_k \varphi_k(\mathbf{x}) \\
&= \sum_{k=1}^{d} (\widetilde{\beta}_k - \beta_k) \widetilde{\varphi}_k(\mathbf{x}) + \sum_{k=1}^{d} \beta_k \left[\widetilde{\varphi}_k(\mathbf{x}) - \varphi_k(\mathbf{x}) \right].
\end{aligned}$$

For any $1 \leqslant k \leqslant d$, by (3.3) in Proposition 3, we have

$$\begin{aligned}
\widetilde{\beta}_k - \beta_k &= \frac{1}{m} \sum_{i=1}^{m} Y_i^{\mathcal{G}} \, \widetilde{\varphi}_k(\mathbf{X}_i^{\mathcal{G}}) - \beta_k \\
&= \frac{1}{m} \sum_{i=1}^{m} Y_i^{\mathcal{G}} \, \varphi_k(\mathbf{X}_i^{\mathcal{G}}) - \beta_k + O_P(\xi_m) \\
&= O_P\left(m^{-1/2} + \xi_m \right) = O_P(\xi_m),
\end{aligned}$$

which indicates that

$$\sup_{\mathbf{x} \in \mathcal{G}} \left| \sum_{k=1}^{d} (\widetilde{\beta}_k - \beta_k) \widetilde{\varphi}_k(\mathbf{x}) \right| = O_P(\xi_m). \qquad \text{(A.23)}$$

Noting that $\max_{1 \leqslant k \leqslant d} |\beta_k|$ is bounded, by (3.3) in Proposition 3 again, we have

$$\sup_{\mathbf{x} \in \mathcal{G}} \left| \sum_{k=1}^{d} \beta_k \left[\widetilde{\varphi}_k(\mathbf{x}) - \varphi_k(\mathbf{x}) \right] \right| = O_P(\xi_m). \qquad \text{(A.24)}$$

In view of (A.23) and (A.24), we complete the proof of (3.4). ■

Proof of Proposition 4. The proof is similar to the proof of Proposition 3 above. Thus we next only sketch the modification.

Following the proof of Lemma 1, we may show that

$$\max_{1\leqslant j,k\leqslant m}\left|\frac{1}{m}\sum_{i=1}^{m}\varphi_j(\mathbf{X}_i^{\mathcal{G}})\varphi_k(\mathbf{X}_i^{\mathcal{G}})-I(j=k)\right|=O_P(\xi_m),$$

which indicates that

$$\left\|\mathbf{\Phi}_m^{\mathrm{T}}\mathbf{\Phi}_m-\mathbf{I}_{d_m}\right\|=O_P\left(d_m\xi_m\right). \tag{A.25}$$

Using (A.25) and following the proof of (3.2), we may complete the proof of (4.6).

On the other hand, note that when d_m is diverging,

$$\max_{1\leqslant k\leqslant d_m}\left\|\frac{1}{m}\mathbf{K}_{\mathcal{G}}\boldsymbol{\varphi}_k-\widetilde{\lambda}_k\boldsymbol{\varphi}_k\right\|\leqslant\max_{1\leqslant k\leqslant d_m}\left\|\frac{1}{m}\mathbf{K}_{\mathcal{G}}\boldsymbol{\varphi}_k-\lambda_k\boldsymbol{\varphi}_k\right\|+\max_{1\leqslant k\leqslant d_m}\left\|\lambda_k\boldsymbol{\varphi}_k-\widetilde{\lambda}_k\boldsymbol{\varphi}_k\right\|=O_P(d_m\xi_m), \tag{A.26}$$

and

$$\left\|\frac{1}{m}\mathbf{K}_{\mathcal{G}}\boldsymbol{\varphi}_k-\widetilde{\lambda}_k\boldsymbol{\varphi}_k\right\|^2=\sum_{j=1}^{m}\left\|\langle\tfrac{1}{m}\mathbf{K}_{\mathcal{G}}\boldsymbol{\varphi}_k,\widehat{\boldsymbol{\varphi}}_j\rangle-\widetilde{\lambda}_k\langle\boldsymbol{\varphi}_k,\widehat{\boldsymbol{\varphi}}_j\rangle\right\|^2\geqslant(1+o_P(1))\rho_m^2\cdot\sum_{j=1,\neq k}^{m}\delta_{kj}^2. \tag{A.27}$$

By (A.26), (A.27) and Assumption 2*, we readily have

$$\max_{1\leqslant k\leqslant d_m}\Delta_k^2\equiv\max_{1\leqslant k\leqslant d_m}\sum_{j=1,\neq k}^{m}\delta_{kj}^2=O_P(d_m^2\xi_m^2/\rho_m^2). \tag{A.28}$$

Using (A.28) and (A.19)–(A.22) (with slight modification), we may complete the proof of (4.7). The proof of Proposition 4 has been completed. ■

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.

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., and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press.

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.

Horn, R. A., and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

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

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.

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.