Noname manuscript No.
(will be inserted by the editor)

Nonlinear Dimension Reduction for Conditional Quantiles

Eliana Christou[1] · **Annabel Settle**[1] · **Andreas Artemiou**[2]

Received: date / Accepted: date

[1] Department of Mathematics and Statistics, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC
[2] School of Mathematics, Cardiff University, Cardiff CF10 3AT, United Kingdom
Corresponding author: Eliana Christou, echris15@uncc.edu

Abstract In practice, data often display heteroscedasticity, making quantile regression (QR) a more appropriate methodology. Modeling the data, while maintaining a flexible nonparametric fitting, requires smoothing over a high-dimensional space which might not be feasible when the number of the predictor variables is large. This problem makes necessary the use of dimension reduction techniques for conditional quantiles, which focus on extracting *linear combinations* of the predictor variables without losing any information about the conditional quantile. However, nonlinear features can achieve greater dimension reduction. We, therefore, present the *first nonlinear extension* of the linear algorithm for estimating the central quantile subspace (CQS) using kernel data. First, we describe the feature CQS within the framework of reproducing kernel Hilbert space, and second, we illustrate its performance through simulation examples and real data applications. Specifically, we emphasize on visualizing various aspects of the data structure using the first two feature extractors, and we highlight the ability to combine the proposed algorithm with classification and regression linear algorithms. The results show that the feature CQS is an effective kernel tool for performing nonlinear dimension reduction for conditional quantiles.

Keywords Classification · Dimension reduction · Quantile Regression · Reproducing kernel Hilbert space · Visualization

1 Introduction

In many situations, data exhibit heteroscedasticity, a characteristic of great scientific importance which is often overlooked. Koenker and Bassett (1978) introduced quantile regression (QR), an alternative to ordinary least squares regression, and considered the linear model $Q_\tau(Y|\mathbf{x}) = \alpha_\tau + \boldsymbol{\beta}_\tau^\top \mathbf{x}$, where Y denotes a univariate response, $\mathbf{X}$ a p-dimensional set of predictors, $Q_\tau(Y|\mathbf{x})$ a τ-th conditional quantile of Y given $\mathbf{X} = \mathbf{x}$, $0 < \tau < 1$, and $\alpha_\tau \in \mathbb{R}$, $\boldsymbol{\beta}_\tau \in \mathbb{R}^p$. Since then, QR has received growing interest and several authors considered the completely flexible nonparametric estimation of the conditional quantiles; see, e.g., Truong (1989), Chaudhuri (1991), Yu and Jones (1998), Takeuchi et al. (2006), Kong et al. (2010), and Guerre and Sabbah (2012).

A fully nonparametric approach for estimating the conditional quantiles can be very challenging when the set of the predictors is large, thus requiring *dimension reduction* techniques. *Linear* dimension reduction techniques for QR focus on extracting the fewest linear combinations of $\mathbf{X}$ that contain all the information about the conditional quantile and have been extensively researched. Wu et al. (2010), Kong and Xia (2012), and Christou and Akritas (2016) considered the single-index quantile regression (SIQR) model, while Kong and Xia (2014) extended to a multi-index quantile regression (MIQR) model. In addition, Luo et al. (2014) introduced a sufficient dimension reduction with respect to any conditional statistical functional, e.g., conditional quantile, while Christou (2020) proposed an alternative algorithm that achieves substantial performance gain.

These *linear* dimension reduction techniques fail to find important *nonlinear* features. Therefore, the overall goal of this paper is to find a nonlinear feature extractor in order to explore conditional quantiles of complex, high-dimensional data, with nonlinear structures.

To construct a nonlinear extension of a linear algorithm, we use the so-called 'kernel trick'. This concept was first introduced by Aizerman et al. (1964), although the name seems to originate in the influential paper of Schölkopf et al. (2004). The main idea is to transform the data into a *very high-dimensional feature reproducing kernel Hilbert space* (RKHS; Aronszajn 1950), and then seek for low-dimensional projections by applying a linear algorithm. In another words, linear directions in the feature space correspond to nonlinear directions in the original data space. In this work, we extend the linear algorithm of Christou (2020) by considering a nonlinear embedding of the data into an RKHS. We demonstrate the performance of the proposed algorithm through simulation examples and real data applications. Specifically, we emphasize on visualizing various aspects of the data structure using the first two feature extractors, and we highlight the ability to combine the proposed algorithm with classification and regression linear algorithms.

The paper is organized as follows. Section 2 gives a brief review of the τth central quantile subspace (CQS) and its estimation. Section 3 introduces the kernel extension of the algorithm by mapping the data into an RKHS using a kernel function. Section 4 presents results from several simulation examples, while Section 5 illustrates the performance of the methodology through real data applications. A brief discussion is given in Section 6.

2 The τth central quantile subspace

A dimension reduction subspace is the column space of any matrix $\mathbf{A}$ such that Y and $\mathbf{X}$ are conditionally independent given $\mathbf{A}^\top\mathbf{X}$, and the central subspace (CS), denoted by $\mathcal{S}_{Y|\mathbf{X}}$, is the dimension reduction subspace with the smallest dimension. However, when the error term is heteroscedastic and the conditional quantile of the response given the predictors is of interest, the CS cannot be used as it can be larger and provide more directions than necessary.

Christou (2020) introduced the concept of the τth central quantile subspace (τ-CQS), a special case of Definition 1 of Luo et al. (2014). Specifically, for a matrix $\mathbf{B}_\tau$, if

$$Y \perp\!\!\!\perp Q_\tau(Y|\mathbf{X})|\mathbf{B}_\tau^\top\mathbf{X}, \tag{1}$$

then the space spanned by the columns of $\mathbf{B}_\tau$ is a τth quantile dimension reduction subspace for the regression of Y on $\mathbf{X}$. The τ-CQS, denoted by $\mathcal{S}_{Q_\tau(Y|\mathbf{X})}$, is the smallest τth quantile dimension reduction subspace. For the rest of the paper, we assume that the CS is spanned by the matrix $\mathbf{A}$, and the τ-CQS is spanned by the matrix $\mathbf{B}_\tau$.

For a fixed $\tau \in (0,1)$, Christou (2020) showed that the slope vector from regressing $Q_\tau(Y|\mathbf{X})$ on $\mathbf{X}$ is contained in the linear subspace spanned by $\mathbf{B}_\tau$.

If the dimension of the τ-CQS, denoted by $d_{Q_\tau(Y|\mathbf{X})}$, is one, then the slope vector will be exhaustive. However, if $d_{Q_\tau(Y|\mathbf{X})} > 1$, then the slope vector will be inconsistent and a different method is necessary to produce more vectors in the linear subspace spanned by $\mathbf{B}_\tau$; see part (b) of Theorem 1.

Theorem 1 ***(Christou 2020).*** *For a given $\tau \in (0,1)$, assume that $Y \perp\!\!\!\perp Q_\tau(Y|\mathbf{X})|\mathbf{B}_\tau^\top\mathbf{X}$. If the conditional expectation $E(\mathbf{b}_\tau^\top\mathbf{X}|\mathbf{B}_\tau^\top\mathbf{X})$ is linear in $\mathbf{B}_\tau^\top\mathbf{X}$ for every $\mathbf{b}_\tau \in \mathbb{R}^p$ (linearity condition), then*

(a) $\boldsymbol{\beta}_\tau^* \in \mathcal{S}_{Q_\tau(Y|\mathbf{X})}$, where

$$(\alpha_\tau^*, \boldsymbol{\beta}_\tau^*) = \arg\min_{(a_\tau, \mathbf{b}_\tau)} E\left\{Q_\tau(Y|\mathbf{A}^\top\mathbf{X}) - a_\tau - \mathbf{b}_\tau^\top\mathbf{X}\right\}^2,$$

and $\mathcal{S}(\mathbf{A}) = \mathcal{S}_{Y|\mathbf{X}}$.

(b) $E\{Q_\tau(Y|U_\tau)\mathbf{X}\} \in \mathcal{S}_{Q_\tau(Y|\mathbf{X})}$, where U_τ is a measurable function of $\mathbf{B}_\tau^\top\mathbf{X}$, provided that $Q_\tau(Y|U_\tau)\mathbf{X}$ is integrable.

Theorem 1 (b) suggests that if $d_{Q_\tau(Y|\mathbf{X})} > 1$, then we can create more vectors by setting $\boldsymbol{\beta}_{\tau,0} = \boldsymbol{\beta}_\tau^*$ and $\boldsymbol{\beta}_{\tau,j} = E[Q_\tau\{Y|u_\tau(\boldsymbol{\beta}_{\tau,j-1}^\top\mathbf{X})\}\mathbf{X}]$, for $j = 1, 2, \ldots, p-1$; the author used $u_\tau(t) = t$.

The above procedure suggests the following estimation method. First, use a standard dimension reduction technique to estimate $\mathbf{A}$ by $\widehat{\mathbf{A}}$ and form the new predictor vector $\widehat{\mathbf{A}}^\top\mathbf{X}$. Next, use the independent and identically distributed (iid) observations $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ to estimate $\boldsymbol{\beta}_\tau^*$ by

$$(\widehat{\alpha}_\tau, \widehat{\boldsymbol{\beta}}_\tau) = \arg\min_{(a_\tau, \mathbf{b}_\tau)} \sum_{i=1}^n \{\widehat{Q}_\tau(Y|\widehat{\mathbf{A}}^\top\mathbf{X}_i) - a_\tau - \mathbf{b}_\tau^\top\mathbf{X}_i\}^2,$$

where $\widehat{Q}_\tau(Y|\widehat{\mathbf{A}}^\top\mathbf{X}_i)$ is a nonparametric estimate of $Q_\tau(Y|\widehat{\mathbf{A}}^\top\mathbf{X}_i)$. Specifically, take $\widehat{Q}_\tau(Y|\widehat{\mathbf{A}}^\top\mathbf{X}_i) = \widehat{q}_\tau(\mathbf{X}_i)$, where

$$(\widehat{q}_\tau(\mathbf{X}_i), \widehat{\mathbf{s}}_\tau(\mathbf{X}_i)) = \arg\min_{(q_\tau, \mathbf{s}_\tau)} \sum_{k=1}^n \rho_\tau\{Y_k - q_\tau - \mathbf{s}_\tau^\top\widehat{\mathbf{A}}^\top(\mathbf{X}_k - \mathbf{X}_i)\} \times K\left\{\frac{\widehat{\mathbf{A}}^\top(\mathbf{X}_k - \mathbf{X}_i)}{h}\right\}, \quad (2)$$

for $\rho_\tau(u) = \{\tau - I(u < 0)\}u$ a loss function, $K(\cdot)$ a kernel function, and $h > 0$ a bandwidth.

Following, if $d_{Q_\tau(Y|\mathbf{X})} = 1$, then stop and report $\widehat{\boldsymbol{\beta}}_\tau$ as the estimated basis vector for $\mathcal{S}_{Q_\tau(Y|\mathbf{X})}$. Otherwise, set $\widehat{\boldsymbol{\beta}}_{\tau,0} = \widehat{\boldsymbol{\beta}}_\tau$ and form the vectors $\widehat{\boldsymbol{\beta}}_{\tau,j} = n^{-1}\sum_{i=1}^n \widehat{Q}_\tau(Y|\widehat{\boldsymbol{\beta}}_{\tau,j-1}^\top\mathbf{X}_i)\mathbf{X}_i$, for $j = 1, \ldots, p-1$, where $\widehat{Q}_\tau(Y|\widehat{\boldsymbol{\beta}}_{\tau,j-1}^\top\mathbf{X}_i)$ is the local linear conditional quantile estimate of $Q_\tau(Y|\widehat{\boldsymbol{\beta}}_{\tau,j-1}^\top\mathbf{X}_i)$, i.e., $\widehat{Q}_\tau(Y|\widehat{\boldsymbol{\beta}}_{\tau,j-1}^\top\mathbf{X}_i) = \widehat{q}_\tau(\mathbf{X}_i)$ from (2) but $\widehat{\mathbf{A}}$ is replaced with $\widehat{\boldsymbol{\beta}}_{\tau,j-1}$. Finally, form the $p \times p$ matrix $\widehat{\mathbf{V}}_\tau = (\widehat{\boldsymbol{\beta}}_{\tau,0}, \ldots, \widehat{\boldsymbol{\beta}}_{\tau,p-1})$ and choose the eigenvectors $\widehat{\mathbf{v}}_{\tau,k}$, $k =$

$1, \ldots, d_{Q_\tau(Y|\mathbf{X})}$, corresponding to the $d_{Q_\tau(Y|\mathbf{X})}$ largest eigenvalues of $\widehat{\mathbf{V}}_\tau \widehat{\mathbf{V}}_\tau^\top$. Then,

$$\widehat{\mathbf{B}}_\tau = (\widehat{\mathbf{v}}_{\tau,1}, \ldots, \widehat{\mathbf{v}}_{\tau,d_{Q_\tau(Y|\mathbf{X})}}) \tag{3}$$

is an estimated basis matrix for $\mathcal{S}_{Q_\tau(Y|\mathbf{X})}$.

3 The feature τth central quantile subspace

3.1 Population level

As the data cloud of independent variables cannot always be characterized by projections into a low-dimensional linear subspace, nonlinear components are necessary. A nonlinear generalization of (1) is to replace the linear function $\mathbf{B}_\tau^\top \mathbf{X}$ with the nonlinear one $\boldsymbol{\psi}_\tau(\mathbf{X})$, and assume that

$$Y \perp\!\!\!\perp Q_\tau(Y|\mathbf{X})|\boldsymbol{\psi}_\tau(\mathbf{X}). \tag{4}$$

Nonlinear dimension reduction can potentially achieve greater dimension reduction if the data are concentrated on a nonlinear low-dimensional space.

To construct a nonlinear extension of the algorithm presented in Section 2, we need to map the original data into a feature space induced by a kernel function. Then, a linear algorithm in the feature space corresponds to a nonlinear algorithm in the original space; this is called the 'kernel-trick'.

The 'kernel-trick' has been considered by several authors as a method for nonlinear generalization of existing linear algorithms. Generalizations include kernel principal component analysis (Schölkopf et al. 1998, 1999), kernel independent component analysis (Bach and Jordan 2002), kernel Fisher's discriminant analysis (Mika et al. 1999; Baudat and Annouar 2000; Roth and Steinhage 2000), kernel canonical correlation analysis (Lai and Fyfe 2000; Akaho 2001; Fukumizu et al. 2007), kernel sliced inverse regression (Wu 2008; Yeh et al. 2009; Wu et al. 2013), and kernel principal support vector machine (Li et al. 2011). However, to the best of our knowledge, there is no kernel extension of any linear algorithm for performing dimension reduction for conditional quantiles.

Following the ideas from Wu (2008) and Yeh et al. (2009), we map the input space $\mathcal{X} \subset \mathbb{R}^p$ to an isometric isomorphic space $\mathcal{H}_K$ via the transformation Γ given by $\mathbf{X} \to \Gamma(\mathbf{X}) := K(\mathbf{X}, \cdot)$. $\mathcal{H}_K$ is known as the reproducing kernel Hilbert space (RKHS) generated by K and, for a given positive-definite kernel K, it consists of all finite kernel mixtures $\sum_{q=1}^m \lambda_q K(\mathbf{X}, \mathbf{U}_q)$ and their limits, where $m \in \mathbb{N}$, $\mathbf{U}_q \in \mathbb{R}^p$, and $\lambda_q \in \mathbb{R}$, are arbitrary. According to Yeh et al. (2009), the reproducing kernels are assumed to be (1) symmetric and measurable, (2) of trace type, and (3) for $\mathbf{X} \neq \mathbf{U}$, $K(\mathbf{X}, \cdot) \neq K(\mathbf{U}, \cdot)$ in $L_2(\mathcal{X}, \mu)$ sense for Lebesgue measure μ. For the purpose of this paper, we will be writing $\mathcal{H}_{K,\tau}$ to indicate that the constants λ_q can depend on the quantile level τ. Therefore, λ_q are specific to the τth quantile, but we omit the subscript τ for notational convenience.

For $\tau \in (0,1)$, let $H_\tau = \{h_{1,\tau}, \ldots, h_{d_\tau,\tau}\}$ be a collection of elements in $\mathcal{H}_{K,\tau}$, and let $\mathcal{H}_\tau$ be the *linear* subspace spanned by elements in H_τ. The analogous of model (1) in the feature space is

$$Y \perp\!\!\!\perp Q_\tau(Y|\mathbf{X})|\{h_{1,\tau}(\mathbf{X}), \ldots, h_{d_\tau,\tau}(\mathbf{X})\},$$

where $h_{k,\tau}(\mathbf{X}) = \langle h_{k,\tau}(\cdot), K(\mathbf{X}, \cdot)\rangle_{\mathcal{H}_{K,\tau}}, k = 1, \ldots, d_\tau$. Then, $\mathcal{H}_\tau$ is called the feature τth quantile dimension reduction subspace. The smallest feature τth quantile dimension reduction subspace is called the *feature τ-CQS*. For the remaining of this paper, we assume that the feature τ-CQS exists and is spanned by H_τ. We call $h_{k,\tau}$ the feature τ-CQS directions and $h_{k,\tau}(\mathbf{X})$ the feature τ-CQS predictors, for $k = 1, \ldots, d_\tau$.

The linearity condition (LD), defined in Theorem 1, can be stated in the framework of $\mathcal{H}_{K,\tau}$ as follows. For a given τ, the conditional expectation $E\{f_\tau(\mathbf{X})|h_{1,\tau}(\mathbf{X}), \ldots, h_{d_\tau,\tau}(\mathbf{X})\}$ is linear in $\{h_{1,\tau}(\mathbf{X}), \ldots, h_{d_\tau,\tau}(\mathbf{X})\}$, for any $f_\tau \in \mathcal{H}_{K,\tau}$, where $f_\tau(\mathbf{X}) = \langle f_\tau(\cdot), K(\mathbf{X}, \cdot)\rangle_{\mathcal{H}_{K,\tau}}$.

Remark 1 Although the LD in $\mathcal{H}_{K,\tau}$ seems more restrictive than the LD in the classical setting, this is not the case. In fact, kernel functions are flexible and can therefore adequately approximate any smooth function. This stems from Euclidean space linearity being more strict than RKHS linearity; see Wu (2008), Yeh et al. (2009), and Wu et al. (2013) for comments and further details.

3.2 Feature data

The feature space $K(\mathbf{X}, \cdot)$ and the feature τ-CQS directions $h_{k,\tau}$, $k = 1, \ldots, d_\tau$, are in high- or infinite-dimensional space. For practical simplicity, we will use a finite basis and revise Theorem 1 accordingly. For data $\{Y_i, \mathbf{X}_i\}_{i=1}^n$, we form the feature data $K(\mathbf{X}_i, \cdot)$ by considering the finite basis set $\{K(\cdot, \mathbf{X}_i)\}_{i=1}^n$. Then, the feature data become $\{K(\mathbf{X}_i, \mathbf{X}_j)\}_{i,j=1}^n$ and the feature τ-CQS predictors can be expressed as

$$h_{k,\tau}(\mathbf{X}) = \sum_{i=1}^n \alpha_{k,\tau}^i K(\mathbf{X}, \mathbf{X}_i), \quad \alpha_{k,\tau}^1, \ldots, \alpha_{k,\tau}^n \in \mathbb{R},$$

for $k = 1, \ldots, d_\tau$.

Let $\mathbf{T} = (K(\mathbf{X}, \mathbf{X}_1), \ldots, K(\mathbf{X}, \mathbf{X}_n))^\top$ and $\boldsymbol{\alpha}_{k,\tau} = (\alpha_{k,\tau}^1, \ldots, \alpha_{k,\tau}^n)^\top$. The LD can be restated as:

$$E\{\mathbf{a}_\tau^\top \mathbf{T}|\boldsymbol{\alpha}_{1,\tau}^\top \mathbf{T}, \ldots, \boldsymbol{\alpha}_{d_\tau,\tau}^\top \mathbf{T}\} \tag{5}$$

is linear in $\{\boldsymbol{\alpha}_{1,\tau}^\top \mathbf{T}, \ldots, \boldsymbol{\alpha}_{d_\tau,\tau}^\top \mathbf{T}\}$ for every $\mathbf{a}_\tau \in \mathbb{R}^n$. Then, Theorem 1 becomes:

Theorem 2 *For a given $\tau \in (0,1)$, assume that $Y \perp\!\!\!\perp Q_\tau(Y|\mathbf{X})|\{\boldsymbol{\alpha}_{1,\tau}^\top \mathbf{T}, \ldots, \boldsymbol{\alpha}_{d_\tau,\tau}^\top \mathbf{T}\}$, i.e., the feature τ-CQS is spanned by $\{\boldsymbol{\alpha}_{1,\tau}^\top \mathbf{T}, \ldots, \boldsymbol{\alpha}_{d_\tau,\tau}^\top \mathbf{T}\}$. If the LD, stated in (5), holds then*

(a) $\boldsymbol{\zeta}_\tau^* \in span\{\boldsymbol{\alpha}_{1,\tau}^\top \mathbf{T}, \ldots, \boldsymbol{\alpha}_{d_\tau,\tau}^\top \mathbf{T}\}$, *where*

$$(\gamma_\tau^*, \boldsymbol{\zeta}_\tau^*) = \arg \min_{(c_\tau, \mathbf{z}_\tau)} E\{Q_\tau(Y|\boldsymbol{\Gamma}^\top \mathbf{T}) - c_\tau - \mathbf{z}_\tau^\top \mathbf{T}\}^2$$

and span$(\boldsymbol{\Gamma})$ *is the feature CS; see Remark 2.*

(b) If U_τ *is a measurable function of* $\{\boldsymbol{\alpha}_{1,\tau}^\top \mathbf{T}, \ldots, \boldsymbol{\alpha}_{d_\tau,\tau}^\top \mathbf{T}\}$, *then* $E\{Q_\tau(Y|U_\tau)\mathbf{T}\} \in span\{\boldsymbol{\alpha}_{1,\tau}^\top \mathbf{T}, \ldots, \boldsymbol{\alpha}_{d_\tau,\tau}^\top \mathbf{T}\}$, *provided that* $Q_\tau(Y|U_\tau)\mathbf{T}$ *is integrable.*

Proof The proof of Theorem 2 follows directly from Christou (2020). We outline here the basic steps.

(a) Let $R(c_\tau, \mathbf{z}_\tau) = E\{Q_\tau(Y|\boldsymbol{\Gamma}^\top \mathbf{T}) - c_\tau - \mathbf{z}_\tau^\top \mathbf{T}\}^2$ and $\mathcal{A}_\tau = (\boldsymbol{\alpha}_{1,\tau}, \ldots, \boldsymbol{\alpha}_{d_\tau,\tau})$ the $n \times d_\tau$ matrix. Using similar steps as those in Christou (2020), we have

$$\begin{aligned} R(c_\tau, \mathbf{z}_\tau) &= E[E[\{Q_\tau(Y|\boldsymbol{\Gamma}^\top \mathbf{T}) - c_\tau - \mathbf{z}_\tau^\top \mathbf{T}\}^2 | \mathcal{A}_\tau^\top \mathbf{T}]] \\ &\geq E[E\{Q_\tau(Y|\boldsymbol{\Gamma}^\top \mathbf{T}) - c_\tau - \mathbf{z}_\tau^\top \mathbf{T} | \mathcal{A}_\tau^\top \mathbf{T}\}]^2 \\ &= E\{Q_\tau(Y|\mathbf{T}) - c_\tau - \mathbf{z}_\tau^\top E(\mathbf{T}|\mathcal{A}_\tau^\top \mathbf{T})\}^2 \\ &= E\{Q_\tau(Y|\mathbf{T}) - c_\tau - \mathbf{z}_\tau^\top \mathbf{P}_\tau^*(\boldsymbol{\Sigma}_{\mathbf{TT}})^\top \mathbf{T}\}^2 \\ &= R(c_\tau, \mathbf{P}_\tau^*(\boldsymbol{\Sigma}_{\mathbf{TT}})\mathbf{z}_\tau), \end{aligned}$$

where $\mathbf{P}_\tau^*(\boldsymbol{\Sigma}_{\mathbf{TT}}) = \mathcal{A}_\tau(\mathcal{A}_\tau^\top \boldsymbol{\Sigma}_{\mathbf{TT}} \mathcal{A}_\tau)^{-1} \mathcal{A}_\tau^\top \boldsymbol{\Sigma}_{\mathbf{TT}}$ and $\boldsymbol{\Sigma}_{\mathbf{TT}}$ is the covariance matrix of $\mathbf{T}$.

(b) Using similar steps as those in Christou (2020), we have

$$\begin{aligned} E\{Q_\tau(Y|U_\tau)\mathbf{T}\} &= E[E\{Q_\tau(Y|U_\tau)\mathbf{T}|\mathcal{A}_\tau^\top \mathbf{T}\}] = E\{Q_\tau(Y|U_\tau)E(\mathbf{T}|\mathcal{A}_\tau^\top \mathbf{T})\} \\ &= E\{Q_\tau(Y|U_\tau)\mathbf{P}_\tau^*(\boldsymbol{\Sigma}_{\mathbf{TT}})^\top \mathbf{T}\} = \mathbf{P}_\tau^*(\boldsymbol{\Sigma}_{\mathbf{TT}})^\top E\{Q_\tau(Y|U_\tau)\mathbf{T}\}. \end{aligned}$$

Remark 2 According to Definition 1 of Yeh et al. (2009), if $Y \perp\!\!\!\perp \mathbf{X}|\{h_1(\mathbf{X}), \ldots, h_d(\mathbf{X})\}$, where $h_k(\mathbf{X}) = \langle h_k(\cdot), K(\mathbf{X}, \cdot)\rangle_{\mathcal{H}_K}$, $k = 1, \ldots, d$, then the linear subspace spanned by $\{h_1, \ldots, h_d\}$ is called the feature effective dimension reduction (e.d.r) subspace of $\mathcal{H}_K$. The smallest feature e.d.r subspace is called the feature CS. According to Wu (2008) and Yeh et al. (2009), we can estimate the feature e.d.r predictors $h_k(\mathbf{X})$, $k = 1, \ldots, d$, by applying SIR (Li 1991) on the feature data. This is called the kernel SIR (KSIR).

Remark 3 The kernel data $\{K(\mathbf{X}_i, \mathbf{X}_j)\}_{i,j=1}^n$ consist of $n \times n$ observations, a dimension that can pose numerical difficulties especially when n is large. Therefore, a subset $\{\mathbf{X}_i\}_{i=1}^{n'}$ of size $n' < n$ can be used to form the data $\{K(\mathbf{X}_i, \mathbf{X}_j)\}_{n \times n'}$; we will call this the reduced kernel data. For other ways to deal with numerical instabilities see Yeh et al. (2009) and Wu et al. (2013).

3.3 Sample level

Let $\{Y_i, \mathbf{X}_i\}_{i=1}^n$ be iid observations. Form the kernel matrix $\{K(\mathbf{X}_i, \mathbf{X}_j)\}_{i,j=1}^n$ and define the new predictors $\mathbf{T}_i = (K(\mathbf{X}_i, \mathbf{X}_1), \ldots, K(\mathbf{X}_i, \mathbf{X}_n))^\top$. Apply the algorithm of Christou (2020) to the data $\{Y_i, \mathbf{T}_i\}_{i=1}^n$. That is:

1. Use KSIR to estimate the $n \times d$ basis matrix $\mathbf{\Gamma}$ of the feature CS, denoted by $\widehat{\mathbf{\Gamma}}$, and form the new feature e.d.r predictors $\widehat{\mathbf{\Gamma}}^\top \mathbf{T}_i$, $i = 1, \ldots, n$. This will be performed by applying SIR on the data $\{Y_i, \mathbf{T}_i\}_{i=1}^n$; see Wu (2008) and Yeh et al. (2009).
2. For each $i = 1, \ldots, n$, use the local linear conditional quantile estimation method of Guerre and Sabbah (2012) to estimate $Q_\tau(Y|\widehat{\mathbf{\Gamma}}^\top \mathbf{T}_i)$. Specifically, take $\widehat{Q}_\tau(Y|\widehat{\mathbf{\Gamma}}^\top \mathbf{T}_i) = \widehat{q}_\tau(\mathbf{T}_i)$, where $\widehat{q}_\tau(\mathbf{T}_i)$ is given by (2), except we replace $\widehat{\mathbf{A}}$ by $\widehat{\mathbf{\Gamma}}$ and $\mathbf{X}_i$ by $\mathbf{T}_i$.
3. Take $\widehat{\boldsymbol{\zeta}}_\tau$ to be

$$(\widehat{\gamma}_\tau, \widehat{\boldsymbol{\zeta}}_\tau) = \arg \min_{(c_\tau, \mathbf{z}_\tau)} \sum_{i=1}^n \{\widehat{Q}_\tau(Y|\widehat{\mathbf{\Gamma}}^\top \mathbf{T}_i) - c_\tau - \mathbf{z}_\tau^\top \mathbf{T}_i\}^2.$$

4. If $d_\tau = 1$, stop and report $\widehat{\boldsymbol{\zeta}}_\tau$ as the estimated basis vector for the feature τ-CQS. Otherwise, move to Step 5.
5. Set $\widehat{\boldsymbol{\zeta}}_{\tau,0} = \widehat{\boldsymbol{\zeta}}_\tau$.
6. Given j, for $j = 1, \ldots, n-1$,
   (a) form the predictors $\widehat{\boldsymbol{\zeta}}_{\tau,j-1}^\top \mathbf{T}_i$, $i = 1, \ldots, n$, and use the local linear conditional quantile estimation method of Guerre and Sabbah (2012) to estimate $Q_\tau(Y|\widehat{\boldsymbol{\zeta}}_{\tau,j-1}^\top \mathbf{T}_i)$. Specifically, take $\widehat{Q}_\tau(Y|\widehat{\boldsymbol{\zeta}}_{\tau,j-1}^\top \mathbf{T}_i) = \widehat{q}_\tau(\mathbf{T}_i)$, where $\widehat{q}_\tau(\mathbf{T}_i)$ is given in (2), except that we replace $\widehat{\mathbf{A}}$ by $\widehat{\boldsymbol{\zeta}}_{\tau,j-1}$ and $\mathbf{X}_i$ by $\mathbf{T}_i$.
   (b) let $\widehat{\boldsymbol{\zeta}}_{\tau,j} = n^{-1} \sum_{i=1}^n \widehat{Q}_\tau(Y|\widehat{\boldsymbol{\zeta}}_{\tau,j-1}^\top \mathbf{T}_i)\mathbf{T}_i$.
7. Let $\widehat{\mathbf{W}}_\tau$ be the $n \times n$ matrix with column vectors $\widehat{\boldsymbol{\zeta}}_{\tau,j}$, $j = 0, 1, \ldots, n-1$, that is, $\widehat{\mathbf{W}}_\tau = (\widehat{\boldsymbol{\zeta}}_{\tau,0}, \ldots, \widehat{\boldsymbol{\zeta}}_{\tau,n-1})$, and choose the eigenvectors $\widehat{\mathbf{w}}_{\tau,k}$, $k = 1, \ldots, d_\tau$, corresponding to the d_τ largest eigenvalues of $\widehat{\mathbf{W}}_\tau \widehat{\mathbf{W}}_\tau^\top$. Then,

$$\widehat{\mathbf{H}}_\tau = (\widehat{\mathbf{w}}_{\tau,1}, \ldots, \widehat{\mathbf{w}}_{\tau,d_\tau}) \tag{6}$$

   is an estimated basis matrix for the feature τ-CQS.

This algorithm gives an estimated basis matrix for the feature τ-CQS. We can then form the new feature τ-CQS predictors $\{\widehat{\mathbf{H}}_\tau^\top \mathbf{T}_i\}_{i=1}^n$ and use existing nonparametric QR techniques to estimate the conditional quantile function.

4 Simulation Studies

4.1 Computational Remarks

In this section, we demonstrate the finite sample performance of the proposed feature τ-CQS and compare it with that of the τ-CQS of Christou (2020). For the simulations, several parameters and measurements need to be specified.
Kernel choice. 'The choice of kernel type is often not crucial as long as the chosen kernel consistutes suitable building blocks for the underlying functional class' (Yeh et al. 2009, p. 1600). In this work, we employ the commonly used

Gaussian kernel $K(\mathbf{x}, \mathbf{u}) = \exp\{-\gamma \|\mathbf{x} - \mathbf{u}\|^2\}$, where γ needs to be specified. Although the optimal choice of γ has been investigated, the parameter is generally domain-specific (Duan et al. 2003; Keerthi and Lin 2003). Therefore, we investigate the choice of γ for all the simulation examples considered in this work (see Example 1 for details); a γ of 0.01 is concluded as the best choice for all models considered in this section.

Data structure and kernel data. Training and test sets are generated for all simulation examples. Specifically, unless otherwise stated, the sample size for the training set is chosen to be $n_t = 600$. An independent test set of size $n_{te} = 1000$ is also generated. The kernel data is formed by applying the Gaussian kernel with $\gamma = 0.01$ on the original data. To avoid numerical instabilities, a random subset of 10% from the training set is chosen and the reduced kernel data is formed, resulting in $n_t \times n'$ for the training set and in $n_{te} \times n'$ for the test set, where $n' = 0.1n_t$.

Structural dimension. We assume that the true structural dimensions of the feature τ-CQS and of the τ-CQS of Christou (2020) are known for Examples 1-4. However, Example 5 demonstrates the performance of the modified BIC-type criterion (Zhu et al. 2010) for estimating d_τ for the feature τ-CQS. Specifically, d_τ is estimated by $\arg\max_{1 \le k \le n'} G_{n_t}(k)$, where

$$G_{n_t}(k) = n_t \frac{\sum_{i=1}^{k} \widehat{\lambda}_i^2}{\sum_{i=1}^{n'} \widehat{\lambda}_i^2} - C_{n_t} \left\{ \frac{k(k+1)}{2} \right\}, \tag{7}$$

where $\widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_{n'}$ are the eigenvalues of the matrix $\widehat{\mathbf{W}}_\tau \widehat{\mathbf{W}}_\tau^\top$, and $C_{n_t}/n_t \to 0$, $C_{n_t} \to \infty$ as $n_t \to \infty$.

Evaluation. To evaluate the performance of the feature τ-CQS we use the distance correlation between $\widehat{\mathbf{H}}_\tau^\top \mathbf{T}$ and $\boldsymbol{\psi}_\tau(\mathbf{X})$, denoted by $\mathrm{dCor}\{\widehat{\mathbf{H}}_\tau^\top \mathbf{T}, \boldsymbol{\psi}_\tau(\mathbf{X})\}$, where $\widehat{\mathbf{H}}_\tau$ and $\boldsymbol{\psi}_\tau(\mathbf{X})$ are defined in (6) and (4), respectively. This is calculated using the function `dcor` in the `R` package `energy`. To compare the performance of the proposed methodology with that of the τ-CQS of Christou (2020), we calculate the distance correlation between $\widehat{\mathbf{B}}_\tau^\top \mathbf{X}$ and $\mathbf{B}_\tau^\top \mathbf{X}$, where $\widehat{\mathbf{B}}_\tau$ and $\mathbf{B}_\tau$ are defined in (3) and (1), respectively. To avoid overfitting, the distance correlations are evaluated over the independent test set.

Setting. The choice of the models considered in Section 4.2 follows from Wang et al. (2018). The quantiles under consideration are $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$, and the results are based on $N = 100$ iterations.

4.2 Simulation Results

Example 1. To provide reasoning as to how the behavior of the feature τ-CQS directions are affected with the scale parameter γ, we perform an experiment using γ from 0.01 to 0.91, with increments of 0.1. The data are generated according to

$$\text{Model I: } Y = \psi_1(\mathbf{X}) + 0.2\varepsilon, \quad \psi_1(\mathbf{X}) = X_1/\{0.5 + (X_2 + 1.5)^2\},$$
$$\text{Model II: } Y = \psi_1(\mathbf{X}) + 0.2\varepsilon, \quad \psi_1(\mathbf{X}) = \sin(X_1) + \sin(X_2),$$

where $\mathbf{X} = (X_1, \ldots, X_{10})^\top$ and the error ε are generated according to a standard normal distribution. Figure 1 demonstrates the mean, over 100 iterations, distance correlation for the test set for the five different quantile levels. We observe that the mean distance correlation decreases with increasing γ, indicating that a smaller γ yields better performance. Note that, the same pattern was observed for the rest of the models considered in this section. Therefore, a γ of 0.01 was used for the remaining of the examples.

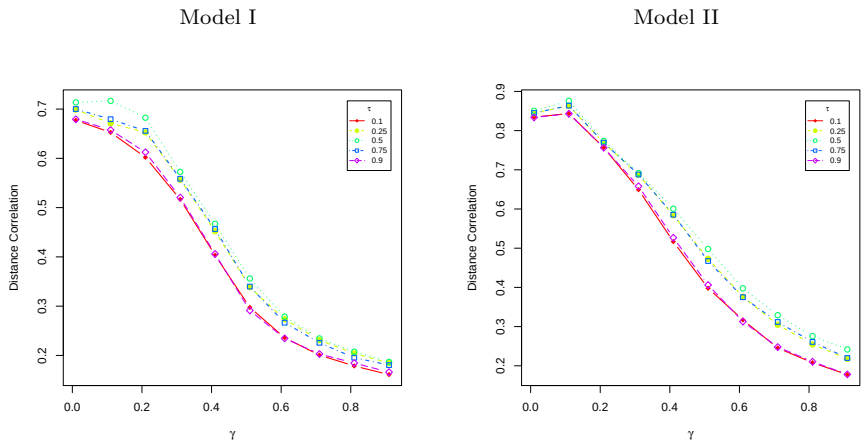

Fig. 1 Mean distance correlation against the scale parameter γ for $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$, for Example 1. A Gaussian kernel is used

Example 2. We now investigate the performance of the proposed algorithm, described in Section 3.3, for different choices of n_t and p. The data are generated according to Model II, where $\mathbf{X} = (X_1, \ldots, X_p)^\top$ and the error ε are generated according to a standard normal distribution. The sample size for the training set is given by $n_t = 200, 400$, or 600, and the number of predictors is $p = 10, 20$, or 40. Table 1 reports the mean and standard deviation of the distance correlation for the test set. First, as expected, we observe that the performance of the proposed method is robust to the specific quantile τ. Moreover, we note that the distance correlation decreases with p and increases with n. However, the relationship between distance correlation and n is not always clear. The reason is that the kernel data has dimension $n_t \times n'$, where $n' = 0.1n_t$, and therefore, the number of columns of the feature data changes with the sample size.

Example 3. This example demonstrates the performance of the proposed algorithm for different number of feature τ-CQS directions used. The data are generated according to the heteroscedastic model

$$\text{Model III: } Y = \psi_1(\mathbf{X}) + \psi_2(\mathbf{X})\varepsilon,$$
$$\psi_1(\mathbf{X}) = \exp(X_1 + X_2) - 1.05, \quad \psi_2(\mathbf{X}) = \exp(X_2)/5,$$

Table 1 Mean (and standard deviation) of the distance correlation for $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$, for Example 2. A Gaussian kernel with $\gamma = 0.01$ is used.

n_t	p	0.1	0.25	0.5	0.75	0.9
200	10	0.876 (0.026)	0.886 (0.024)	0.889 (0.021)	0.885 (0.025)	0.875 (0.030)
	20	0.789 (0.070)	0.798 (0.067)	0.803 (0.066)	0.796 (0.064)	0.786 (0.064)
	40	0.472 (0.119)	0.488 (0.116)	0.496 (0.112)	0.489 (0.119)	0.468 (0.129)
400	10	0.855 (0.027)	0.865 (0.023)	0.869 (0.022)	0.865 (0.025)	0.853 (0.030)
	20	0.854 (0.031)	0.866 (0.026)	0.870 (0.023)	0.864 (0.025)	0.847 (0.031)
	40	0.721 (0.053)	0.733 (0.051)	0.738 (0.046)	0.731 (0.050)	0.721 (0.053)
600	10	0.836 (0.028)	0.850 (0.026)	0.852 (0.023)	0.848 (0.025)	0.835 (0.025)
	20	0.835 (0.024)	0.851(0.021)	0.856 (0.019)	0.850 (0.023)	0.835 (0.033)
	40	0.799 (0.031)	0.813 (0.029)	0.819 (0.029)	0.811 (0.032)	0.794 (0.038)

where $\mathbf{X} = (X_1, \ldots, X_{10})^\top$ and the error ε are generated according to a standard normal distribution. Figure 2 demonstrates the mean distance correlation for the test set and suggests that there are two feature τ-CQS directions. This observation agrees with the model.

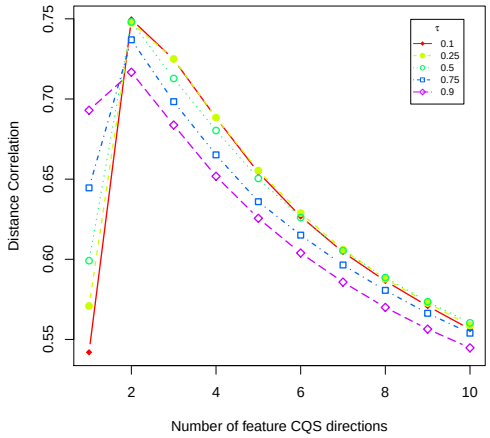

Fig. 2 Mean distance correlation against the dimensionality from one to ten for $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$, for Example 3. A Gaussian kernel with $\gamma = 0.01$ is used

Example 4. We compare the performance of the feature τ-CQS with that of the τ-CQS of Christou (2020). The data are generated according to Models

II, III, and the heteroscedastic Models IV and V as follows:

$$\text{Model IV: } Y = \psi_1(\mathbf{X}) + \psi_2(\mathbf{X})\varepsilon, \quad \psi_1(\mathbf{X}) = X_1^2 + X_2^2, \quad \psi_2(\mathbf{X}) = \sin(X_2)/2,$$
$$\text{Model V: } Y = \psi_1(\mathbf{X})\varepsilon, \quad \psi_1(\mathbf{X}) = (X_1 + X_2)^2,$$

where $\mathbf{X} = (X_1, \ldots, X_{10})^\top$ and the error ε are generated according to a standard normal distribution. Table 2 reports the mean and standard deviation of the distance correlation for the test set of the feature τ-CQS and the τ-CQS of Christou (2020). We observe that the feature τ-CQS reports higher mean distance correlations, demonstrating that the proposed methodology is able to better capture the nonlinear data structure.

For illustration purposes, we also include an example where the subspace spanned by the feature τ-CQS is identical to that spanned by the τ-CQS of Christou (2020). The data are generated according to

$$\text{Model VI: } Y = \psi_1(\mathbf{X}) + \varepsilon, \quad \psi_1(\mathbf{X}) = 3X_1 + X_2,$$

where the setup for $\mathbf{X}$ and ε follows as before. The results are also included in Table 2, where we observe that, as expected, both methods have comparable performance with τ-CQS of Christou (2020) perform better by a small margin.

Table 2 Mean (and standard deviation) of the distance correlation for the feature τ-CQS and for the τ-CQS of Christou (2020) for $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$, for Example 4. A Gaussian kernel with $\gamma = 0.01$ is used.

M	τ-CQS	0.1	0.25	0.5	0.75	0.9
II	Feature	**0.834** (0.033)	**0.844** (0.028)	**0.851** (0.024)	**0.845** (0.026)	**0.834** (0.030)
	Linear	0.727 (0.046)	0.729 (0.049)	0.729 (0.050)	0.730 (0.050)	0.727 (0.043)
III	Feature	**0.750** (0.037)	**0.748** (0.038)	**0.748** (0.041)	**0.737** (0.039)	**0.717** (0.039)
	Linear	0.719 (0.042)	0.720 (0.038)	0.720 (0.039)	0.718 (0.042)	0.714 (0.050)
IV	Feature	**0.778** (0.038)	**0.782** (0.031)	**0.770** (0.028)	**0.752** (0.028)	**0.736** (0.033)
	Linear	0.632 (0.079)	0.518 (0.143)	0.465 (0.138)	0.530 (0.137)	0.588 (0.126)
V	Feature	**0.492** (0.167)	**0.509** (0.187)	**0.446** (0.171)	**0.455** (0.152)	**0.483** (0.174)
	Linear	0.254 (0.188)	0.249 (0.185)	0.226 (0.172)	0.227 (0.160)	0.237 (0.188)
VI	Feature	0.974 (0.009)	0.980 (0.006)	0.981 (0.006)	0.979 (0.006)	0.974 (0.008)
	Linear	**0.998** (0.001)	**0.998** (0.001)	**0.998** (0.001)	**0.998** (0.001)	**0.998** (0.001)

To further comment on the performance of the proposed methodology, we consider methods that estimate the feature CS. For that reason we use KSIR of Wu (2008) and kernel principal support vector machine (KPSVM) of Li et al. (2011) as comparison methods. We repeat Models II - VI and calculate the distance correlation of the feature CS estimated by the two methods. However, for a more direct and fair comparison between the distance correlations of the feature τ-CQS and that of the feature CS, we only report the results for Models II and VI, which include model situations where the two subspaces are identical. Table 3 reports the mean and standard deviation of the distance correlations of the feature CS estimated by KSIR and KPSVM, respectively. Comparing the numbers with those of Table 2 we observe that KSIR performs

better than the proposed methodology by a small margin for both models, while KPSVM performs the best for Model II and has a comparable performance with the proposed methodology for Model VI.

Table 3 Mean (and standard deviation) of the distance correlation for the feature CS estimated by KSIR of Wu (2008) and KPSVM of Li et al. (2011) for Example 4.

M	KSIR	KPSVM
II	0.919 (0.009)	0.942 (0.008)
VI	0.989 (0.002)	0.975 (0.003)

Example 5. In this example we evaluate the performance of the modified BIC-type criterion, defined in (7). The data are generated according to Model V, and the error-only models

$$\text{Model VII: } Y = \psi_1(\mathbf{X})\varepsilon, \quad \psi_1(\mathbf{X}) = (X_1^2 + X_2^2),$$
$$\text{Model VIII: } Y = \psi_1(\mathbf{X})\varepsilon, \quad \psi_1(\mathbf{X}) = (X_1^3 + X_2^3),$$

where $\mathbf{X} = (X_1, \ldots, X_{10})^\top$ and the error ε are generated according to a standard normal distribution. The unknown structural dimension of the feature τ-CQS, d_τ, is estimated using the modified BIC-type criterion. Table 4 reports the number of times, over the 100 iterations, the method correctly estimates the true structural dimension. We can see that all numbers are close to 100, indicating good performance of the BIC-type criterion.

Table 4 Number of times, over 100 iterations, the modified BIC-type criterion correctly estimates the true structural dimension d_τ for $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$, for Example 5. A Gaussian kernel with $\gamma = 0.01$ is used.

Model	0.1	0.25	0.5	0.75	0.9
V	94	88	93	89	97
VII	92	86	88	91	94
VIII	90	83	95	86	93

Example 6. Finally, we provide some insights on the computational time of the proposed methodology. We expect that the feature τ-CQS will be computationally more expensive as it requires the calculation of the kernel data and it applies the algorithm in the higher-dimensional kernel $n \times n'$ data. For this example, we follow the same setup as in the previous examples and simulate data from Models I - VIII. However, to save space, and since the results are similar, we only report the computation time for Models I - III. The time (in seconds) for calculating the extracted directions over 100 simulation runs is computed, and the average time is reported in Table 5. We observe that, on

average, the τ-CQS of Christou (2020) takes between 5 and 7 seconds to run, while the feature τ-CQS takes between 31 and 37 seconds. The runs were carried out on a Dell Poweredge 820 with 256 GB of memory, Intel(R) Xeon(R) CPU E5-4620, and running Ubuntu 18.04 and R version 3.6.0.

Table 5 Average computation time (in seconds) for estimating the feature τ-CQS and the τ-CQS of Christou (2020) directions for $\tau = 0.1, 0.25, 0.5, 0.75, 0.9$, for Example 6.

Model	τ-CQS	0.1	0.25	0.5	0.75	0.9
I	Feature	34.87	33.32	32.77	31.42	30.64
	Linear	6.15	6.57	6.35	5.71	5.47
II	Feature	36.84	33.88	33.85	33.36	32.30
	Linear	6.66	6.89	6.76	5.98	5.69
III	Feature	34.42	32.71	32.41	32.55	32.33
	Linear	6.48	6.52	6.64	6.55	6.18

5 Real Data Analysis

5.1 Data Sets

We utilize several real-world datasets to apply our proposed methodology by focusing on data visualization, classification, and regression. Below is a brief description of each data set considered.

The Ionosphere data set consists of 351 observations on 35 variables and contains information on radar returns collected by the Space Physics Group of the Johns Hopkins University Applied Physics Laboratory (Sigillito et al. 1989). The dependent variable of interest indicates 'good' and 'bad' radar returns, and the other 34 variables describe the 17 discrete values of the real and imaginary parts of an auto-correlation function. One attribute consisting of only 0 values was excluded from the data set.

The Waveform data set consists of 5000 observations on 22 variables. The dependent variable of interest is the class of waves (3 classes) and the other 21 variables are combinations of the waveforms with noise added. Since this data set will be used for visualization purposes, we follow Wu (2008)'s suggestion and consider a sample of 600 observations. This will provide a more readable plot.

The Wine data set consists of 178 observations on 14 variables and is the result of a chemical analysis of 3 Italian wines from the same region. The dependent variable of interest is the type of wine (3 classes) and the other 13 variables are different elements of a wine, such as alcohol, flavanoids, color intensity, etc.

The Breast Cancer data set consists of 682 observations on 10 variables. The dependent variable of interest indicates either a benign or malignant diagnosis and the other 9 variables describe characteristics of the cell.

The Prostate data set consists of 97 observations on 9 variables. The dependent variable of interest is the seminal vesicle invasion (SVI), which is defined as the presence of prostate cancer, and the other 8 variables are clinical measures, such as prostate weight and Gleason score.

The Auto MPG data set consists of 392 observations on 8 variables. The dependent variable of interest is mpg, the city-cycle fuel consumption in miles per gallon, and the other 7 variables describe characteristics of the car, such as horsepower and weight.

The Boston Housing data set consists of 506 observations on 14 variables. The dependent variable of interest is medv, the median value of owner-occupied homes in $1000s and the other 13 variables are statistical measurements on the 506 census tracts in suburban Boston from the 1970 census. The Charles River dummy variable and the index of accessibility to radial highways are excluded from the analysis. Moreover, based on previous suggestions, e.g., Opsomer and Ruppert (1998) and Wu et al. (2010), the logarithmic transformation of TAX and LSTAT is taken. The data were originally published by Harrison and Rubinfeld (1978).

The Machine CPU data set consists of 209 observations on 8 variables. The dependent variable of interest is estperf, the estimated performance of computer CPUs, and the other 7 variables are measurements on the machine, e.g., the minimum and maximum main memory and the minimum and maximum number of channels. The data were also considered by previous investigators, e.g., Takeuchi et al. (2006).

Refer to the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html) for the *Ionosphere, Waveform, Wine, Breast Cancer,* and *Auto MPG* data sets. The *Prostate* data set can be found in the `lasso2` library in `R`, while *Boston Housing* and *Machine CPU* data sets can be found in the `MASS` library in `R`.

Tables 6, 7, 8 show a summary of characteristics of the datasets considered for visualization, classification, and regression, respectively. The R.S. column reports the proportion of stratified random subset for reduced kernel data, and the γ column reports the choice of the scale parameter for the Gaussian kernel. The choice of these two quantities is based on previous investigations, i.e., Wu (2008) and Yeh et al. (2009). According to Yeh et al. (2009), 'a 10% reduced set is often enough for median-sized data. For smaller sized data sets, one may increase the reduction ratio.'

Table 6 Characteristics of the data sets for visualization.

Data Set	Classes	Training Size	Test Size	Attributes	R.S.	γ
Ionosphere	2	234	117	33	0.2	0.05
Waveform	3	200	400	21	0.2	0.05
Wine	3	59	119	13	0.3	0.05

Table 7 Characteristics of the data sets for classification.

Data Set	Classes	Size	Attributes	R.S.	γ
Breast Cancer	2	682	9	0.3	0.05
Prostate	2	97	8	0.3	0.1

Table 8 Characteristics of the data sets for regression.

Data Set	Size	Attributes	R.S.	γ
Auto MPG	392	7	0.1	0.05
Housing	506	11	0.1	0.05
CPUs	209	7	0.1	0.001

5.2 Data Visualization

The purpose of this application is to show how the main data structure is captured by a low-dimensional subspace. Specifically, we use a training set to extract the linear and feature τ-CQS directions and then plot the first two directions using the test set. Figure 3 shows that the feature τ-CQS has a superior performance over the τ-CQS of Christou (2020) with respect to discriminative and visualization purposes. Observe that, the plots for Ionosphere and Wine demonstrate situations where the data are visually separable in the two-dimensional subspace provided by both methods. However, feature τ-CQS achieves patterns that are most distinctly separated. This is especially evident in the Waveform plots, where, although there is a great amount of overlapping for the τ-CQS of Christou (2020), the data form more distinct clusters for the feature τ-CQS.

5.3 Classification

When using the proposed methodology for classification purposes we first extract the linear and feature τ-CQS directions. We then use them to run a learning algorithm which will be computationally less expensive on the low-dimensional subspace. We use the training set to extract the linear and feature τ-CQS and apply the learning algorithm, and the test set to evaluate the performance using the correct classification accuracy. The learning algorithm used here is the group lasso for binary classification by Hashem et al. (2016). Specifically, Hashem et al. (2016) assumed a linear QR model and estimated the regression coefficients using a Bayesian Gibbs sampling procedure, where

Y is a binary response variable. Following, as the main interest consists of predicting Y for a specific set of realizations $\mathbf{x}$, the classification is based on estimating $P(Y = 1|\mathbf{x})$. The estimation of this probability is performed using the estimated regression coefficients from the binary QR model above; see Section 4 from Hashem et al. (2016) for more details. However, Kordas (2006) suggested estimating $P(Y = 1|\mathbf{x})$ by averaging over different quantile levels. We follow the suggestion and average over $\tau = 0.1, 0.25, 0.5, 0.75$, and 0.9. Finally, for a threshold t, a new object $\mathbf{x}$ is classified to class 1 if $P(Y = 1|\mathbf{x}) > t$. We use $t = 0.5$ for equal misclassification costs (Hashem et al. 2016).

We use a training set to fit a binary linear QR model and obtain the estimated regression coefficients for the different quantile levels, and a test set to classify the response. We then calculate the correct classification accuracy. We use a 10-fold cross validation for the training and test sets and report the average classification accuracy over the 10 replications of the 10-fold partition. The number of the linear and feature τ-CQS directions is estimated using the modified BIC-type criterion defined in (7).

Table 9 reports the mean and standard deviation of the correct classification accuracy, over the 10 replications of the 10-fold partition, for the different data sets. For comparison purposes, we also report the average classification accuracy of the group lasso binary classification algorithm without performing dimension reduction, i.e., with the original predictor variables. For the Breast Cancer data set, we observe that the feature τ-CQS has the higher correct classification proportion. For the Prostate data set, we observe that the performance of the proposed method is not much different from the others, but never fall below.

Table 9 Mean (and standard deviation), over 10 replications, of the classification accuracy for the group lasso binary classification without dimension reduction, and with the feature τ-CQS directions and the τ-CQS directions of Christou (2020).

Data Set	No dimension reduction	Linear τ-CQS	Feature τ-CQS
Breast Cancer	0.820 (0.048)	0.856 (0.082)	**0.927** (0.043)
Prostate	0.806 (0.147)	0.804 (0.156)	**0.807** (0.155)

5.4 Regression

We now evaluate the estimation of the τth conditional quantile using the low-dimensional linear and feature τ-CQS directions. Specifically, we use the training set to extract the linear directions $\widehat{\mathbf{B}}_\tau$ and the feature directions $\widehat{\mathbf{H}}_\tau$, and to form the new sufficient predictors $\widehat{\mathbf{B}}_\tau^\top \mathbf{X}_{train}$ and $\widehat{\mathbf{H}}_\tau^\top \mathbf{T}_{train}$. Following, we fit linear QR models using the sufficient predictors (i.e. $Q_\tau(Y|\mathbf{x}) = \alpha_{0,\tau} + \boldsymbol{\alpha}_{1,\tau}^\top \widehat{\mathbf{B}}_\tau^\top \mathbf{X}_{train}$ and $Q_\tau(Y|\mathbf{x}) = \alpha_{0,\tau} + \boldsymbol{\alpha}_{1,\tau}^\top \widehat{\mathbf{H}}_\tau^\top \mathbf{T}_{train}$) to obtain $\widehat{\alpha}_{0,\tau}$ and $\widehat{\boldsymbol{\alpha}}_{1,\tau}$. We then estimate the conditional quantiles on the test set (i.e.

$\widehat{\alpha}_{0,\tau} + \widehat{\boldsymbol{\alpha}}_{1,\tau}^{\top}\widehat{\mathbf{B}}_{\tau}^{\top}\mathbf{X}_{test}$ and $\widehat{\alpha}_{0,\tau} + \widehat{\boldsymbol{\alpha}}_{1,\tau}^{\top}\widehat{\mathbf{H}}_{\tau}^{\top}\mathbf{T}_{test}$) and evaluate the performance using

$$R^2(\tau) = 1 - \frac{\sum_{i=1}^{n_{te}} \rho_\tau(Y_i - \widehat{\alpha}_{0,\tau} - \widehat{\boldsymbol{\alpha}}_{1,\tau}^{\top}\widehat{\mathbf{B}}_{\tau}^{\top}\mathbf{X}_{test})}{\sum_{i=1}^{n_{te}} \rho_\tau(Y_i - \widehat{Y}^0)},$$

where $\widehat{Y}^0$ is the τth conditional quantile estimate for the only-intercept model (Koenker and Machado 1999). Similarly, we calculate $R^2(\tau)$ using $\widehat{\alpha}_{0,\tau}$ and $\widehat{\boldsymbol{\alpha}}_{1,\tau}^{\top}\widehat{\mathbf{H}}_{\tau}^{\top}\mathbf{T}_{test}$. We use a 10-fold cross validation for the training and test sets and report the average $R^2(\tau)$ over 10 replications. Note that we also estimated the conditional quantiles using nonparametric techniques, but the degree by which feature τ-CQS outperforms the τ-CQS of Christou (2020) was the same. The number of the linear and feature τ-CQS directions is estimated using the modified BIC-type criterion, defined in (7). We report the results for $\tau = 0.1, 0.25, 0.5, 0.75$, and 0.9.

Table 10 presents the average $R^2(\tau)$ for the different data sets. The superior performance of the feature τ-CQS is apparent.

Table 10 $R^2(\tau)$ for feature τ-CQS and τ-CQS of Christou (2020) for the regression data sets.

Data Set	Method	0.1	0.25	0.5	0.75	0.9
Auto MPG	Feature	**0.587**	**0.621**	**0.656**	**0.646**	**0.593**
	Linear	0.493	0.550	0.598	0.597	0.579
Housing	Feature	**0.587**	**0.583**	**0.622**	**0.661**	**0.679**
	Linear	0.515	0.485	0.472	0.499	0.550
CPUs	Feature	**0.690**	**0.732**	**0.798**	**0.867**	**0.884**
	Linear	0.423	0.516	0.674	0.793	0.854

6 Discussion

In this work, we presented the first nonlinear extension of a linear algorithm for performing dimension reduction for conditional quantiles. The method utilized the so-called 'kernel-trick' which allows for nonlinear extension of existing linear algorithms. Simulation results and real data applications demonstrated the performance of the proposed methodology and showed the ability of nonlinear dimension reduction techniques to better explore conditional quantiles of complex high-dimensional data. Moreover, the feature τ-CQS demonstrated data visualization capabilities, promising classification rates, and accurate regression estimates.

Acknowledgements We wish to thank the authors of Hashem et al. (2016) for providing us the `R` code for performing group lasso for binary classification. We also want to thank the anonymous referees, whose comments lead to improvements in the presentation of this paper.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Aizerman M, Braverman E, Rozonoer L (1964) Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 25:821–837.
2. Akaho S (2001) Kernel method for canonical correlation analysis. In Proceedings of the International Meeting of the Psychometric Society (IMPS2001).
3. Aronszajn N (1950) Theory of Reproducing Kernels. Transactions of the American Mathematical Society, 68(3):337–404.
4. Bach FR, Jordan MI (2002) Kernel independent component analysis. Journal of Machine Learning Research, 3:1–48.
5. Baudat G, Annouar F (2000) Generalized discriminant analysis using a kernel approach. Neural Computation, 12(10):2385–2404.
6. Chaudhuri P (1991) Nonparametric estimates of regression quantiles and their local Bahadur representation. Ann Stat 19(2):760–777.
7. Christou E (2020) Central Quantile Subspace. Stat Comput 30:677-695.
8. Christou E, Akritas MG (2016) Single index quantile regression for heteroscedastic data. J Multivar Anal 150:169–182.
9. Duan K, Keerthi SS, Poo AN (2003) Evaluation of Simple Performance Measures for Tuning SVM Hyperparameters. Neurocomputing, 51:41–59.
10. Fukumizu K, Bach FR, Gretton A (2007) Statistical consistency of kernel canonical correlation analysis. Journal of Machine Learning Research, 8:361–383.
11. Guerre E, Sabbah C (2012) Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function. Econom Theory 28(1):87–129.
12. Harrison D, Rubinfeld DL (1978) Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management, 5:81–102.
13. Hashem H, Vinciotti V, Alhamzawi R, Keming Y (2016) Quantile regression with group lasso for classification. Advances in Data Analysis and Classification, 10:375–390.
14. Keerthi SS, Lin CJ (2003) Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. Neural Computation, 15:1667–1689.
15. Koenker R, Bassett G (1978) Regression quantiles. Econometrica, 46(1):33–50.
16. Koenker R, Machado J (1999) Goodness of Fit and Related Inference Processes for Quantile Regression. J Am Stat Assoc 94(448):1296–1310.
17. Kong E, Linton O, Xia Y (2010) Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model. Econom Theory 26(5):1529–1564.
18. Kong E, Xia Y (2012) A single-index quantile regression model and its estimation. Econom Theory 28(4):730–768.
19. Kong E, Xia Y (2014) An adaptive composite quantile approach to dimension reduction. Ann Stat 42(4):1657–1688.
20. Kordas G (2006) Smoothed binary regression quantiles. Journal of Applied Econometrics, 21(3):387–407.
21. Lai PL, Fyfe C (2000) Kernel and nonlinear canonical correlation analysis. International Journal of Neural Systems, 10(5):365–377.
22. Li KC (1991) Sliced inverse regression for dimension reduction. J Am Stat Assoc 86(414):316–327.
23. Luo W, Li B, Yin X (2014) On efficient dimension reduction with respect to a statistical functional of interest. Ann Stat 42(1):382–412.
24. Li B, Artemiou A, Li L (2011) Principal Support Vector Machines for linear and nonlinear sufficient dimension reduction. Ann Stat 39(6):3182–3210.
25. Mika S, Rätsch G, Weston J, Schölkopf B, Müller KR (1999) Fisher discriminant analysis with kernel. In Proceedings of IEEE Neural Networks for Signal Processing Workshop, 9:41–48.

26. Opsomer JD, Ruppert D (1998) A Fully Automated Bandwidth Selection Method for Fitting Additive Models. J Am Stat Assoc 93(442):605–619.
27. Roth V, Steinhage V (2000) Nonlinear discriminant analysis using kernel functions. In Advances in Neural Information Processing Systems, pages 568–574. Cambridge, MA: MIT Press.
28. Schölkopf B, Smola AJ, Müller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10(5):1299–1319.
29. Schölkopf B, Smola AJ, Müller KR (1999) Kernel principal component analysis. Advances in kernel methods: Support vector learning, pages 327–352. Cambridge, MA: MIT Press.
30. Schölkopf B, Tsuda K, Vert JP (eds.) (2004), Kernel Methods in Computational Biology, Cambridge, MA: MIT Press.
31. Sigillito VG, Wing SP, Hutton LV, Baker KB (1989) Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest, 10:262–266.
32. Takeuchi I, Le QV, Sears T, Smola AJ (2006) Nonparametric quantile regression. Journal of Machine Learning Research, 7:1231–1264.
33. Truong YK (1989) Asymptotic properties of kernel estimators based on local medians. Ann Stat 17(2):606–617.
34. Wang C, Shin SJ, and Wu Y (2018) Principal quantile regression for sufficient dimension reduction with heteroscedasticity. Electronic Journal of Statistics 12:2114–2140.
35. Wu HM (2008) Kernel sliced inverse regression with applications to classification. J Comput Graph Stat 17(3):590–610.
36. Wu Q. Liang F, Mukherjee S (2013) Kernel sliced inverse regression: Regularization and consistency. Abstract and Applied Analysis, Volume 2013, Special Issue, Article ID 540725, 11 pages.
37. Wu TZ, Yu K, Yu Y (2010) Single-index quantile regression. J Multivar Anal 101(7):1607–1621.
38. Yeh YR, Huang SY, Lee YJ (2009) Nonlinear dimension reduction with kernel sliced inverse regression. IEEE Transactions on Knowledge and Data Engineering, 21(11):1590–1603.
39. Yu K, Jones MC (1998) Local linear quantile regression. J Am Stat Assoc 93(441):228–237.
40. Zhu LP, Zhu LX, Feng ZH (2010) Dimension reduction in regression through cumulative slicing estimation. J Am Stat Assoc 105(492):1455–1466.

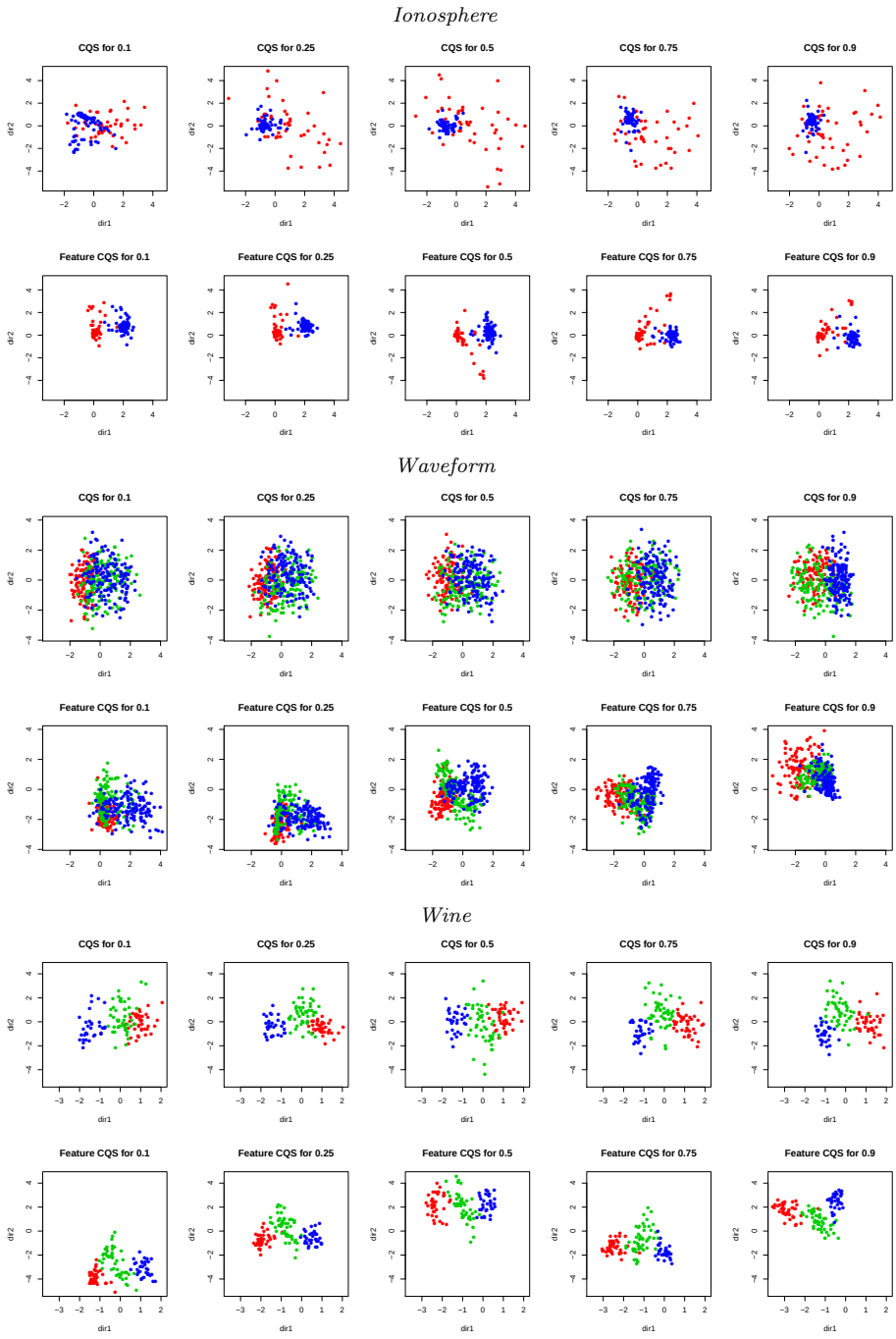

Fig. 3 *The first two extracted directions of the Ionosphere, Waveform, and Wine data sets. For each data set, the first row represents the first two feature τ-CQS directions and the second row represents the first two τ-CQS directions of Christou (2020). The different colors represent the different classes of the data set*