Electronic Journal of Statistics
Vol. 12 (2018) 2114–2140
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1432

Principal quantile regression for sufficient dimension reduction with heteroscedasticity

Chong Wang

Department of Statistics
North Carolina State University
Raleigh, North Carolina, U.S.A.
e-mail: cwang19@ncsu.edu

Seung Jun Shin

Department of Statistics
Korea University
Seoul, South Korea
e-mail: sjshin@korea.ac.kr

Yichao Wu

Department of Mathematics, Statistics and Computer Science
University of Illinois at Chicago
Chicago, IL, U.S.A.
e-mail: yichaowu@uic.edu

Abstract: Sufficient dimension reduction (SDR) is a successful tool for reducing data dimensionality without stringent model assumptions. In practice, data often display heteroscedasticity which is of scientific importance in general but frequently overlooked since a primal goal of most existing statistical methods is to identify conditional mean relationship among variables. In this article, we propose a new SDR method called principal quantile regression (PQR) that efficiently tackles heteroscedasticity. PQR can naturally be extended to a nonlinear version via kernel trick. Asymptotic properties are established and an efficient solution path-based algorithm is provided. Numerical examples based on both simulated and real data demonstrate the PQR's advantageous performance over existing SDR methods. PQR still performs very competitively even for the case without heteroscedasticity.

MSC 2010 subject classifications: 60K35.
Keywords and phrases: Heteroscedasticity, kernel quantile regression, principal quantile regression, sufficient dimension reduction.

Received March 2017.

Contents

1. Introduction

1.1. Background on sufficient dimension reduction

Dimension reduction is often of primary interest in high dimensional data analysis. The principal component analysis (PCA) is widely used in this regard, but it suffers when identifying relationship between response and covariates. Variable selection can be regarded as another type of dimension reduction. However, it often relies on specific model assumptions which can possibly be violated in practice.

Sufficient dimension reduction (SDR) has recently received much attention thanks to its promising performance in reducing data dimensionality under relatively mild model assumptions. SDR achieves dimension reduction of p-dimensional predictor $\mathbf{X}$ by finding a matrix $\mathbf{B} = (\mathbf{b}_1, \cdots, \mathbf{b}_d) \in \mathbb{R}^{p\times d}$ which satisfies

$$Y \perp \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}. \tag{1.1}$$

Under model (1.1), dimension reduction is naturally achieved as long as $d < p$. Compared to the traditional parametric models such as linear regression and generalized linear models, (1.1) is less stringent in the sense that it does not impose a specific model between Y and $\mathbf{X}$. Unlike PCA, it also preserves information about the association between Y and $\mathbf{X}$.

Notice that $\mathbf{B}$ satisfying (1.1) is not unique. Hence the goal is to identify the central subspace [2], which is defined as the intersection of all subspaces spanned by columns of $\mathbf{B}$ satisfying (1.1). The central subspace uniquely exists

under mild conditions (Cook [3]; Yin, Li and Cook [38]) and is typically denoted by $\mathcal{S}_{Y|\mathbf{X}}$. Thus, we assume its existence throughout this article and further $\text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{X}}$ to facilitate the estimation. The dimension of $\mathcal{S}_{Y|\mathbf{X}}$, d is called the structural dimension which is another important quantity to be inferred from the data.

The SDR model (1.1) is often called the linear SDR since the dimension reduction is achieved through finding a linear mapping, $\mathbf{B}^\top\mathbf{X}$. The linear SDR can naturally be extended in a nonlinear fashion by assuming

$$Y \perp \mathbf{X} \mid \boldsymbol{\phi}(\mathbf{X}), \tag{1.2}$$

where $\boldsymbol{\phi} : \mathbb{R}^p \to \mathbb{R}^d$ is an arbitrary function of $\mathbf{X}$ defined on a Hilbert space denoted by $\mathcal{H}$ [5]. Similarly as $\mathbf{B}$, $\boldsymbol{\phi}$ is not unique and we assume that $\boldsymbol{\phi}$ is a unique modulo injective transformation to guarantee its uniqueness [18].

A variety of SDR methods have been developed in the literature. Sliced inverse regression [SIR, 15] and sliced average variance estimation [SAVE, 8] are the two seminal works for the linear SDR and they are still widely-used in many applications. Other linear SDR methods include, but are not limited to, principal Hessian direction [pHd, 17, 4], partial least squares [PLS, 11], inverse regression[6], contour regression[23], directional regression[22], and composite quantile outer-product of gradients method [qOPG, 13]. Toward the nonlinear SDR, several methods are proposed by exploiting reproducing kernel Hilbert space for $\mathcal{H}$ to estimate the nonlinear function $\boldsymbol{\phi}$ via kernel trick. See, for example, Wu [34], Wu, Liang and Mukherjee [35], and Yeh, Huang and Lee [37].

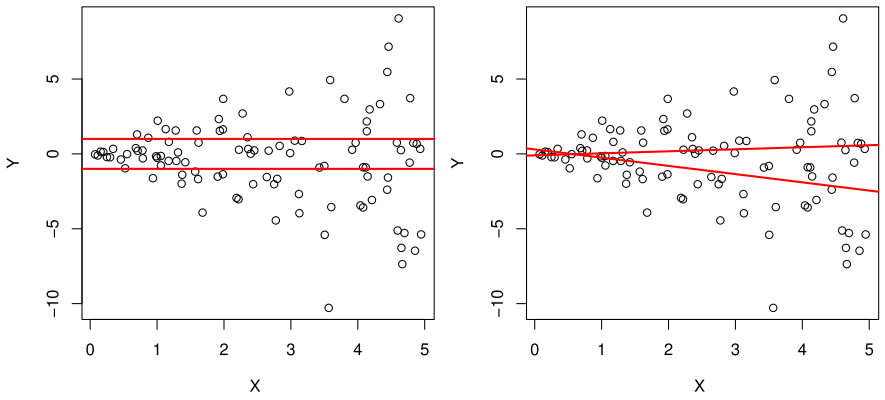
The principal support vector machine [PSVM, 18] is the first attempt to tackle both linear and nonlinear SDR in a unified framework. PSVM connects SDR to the support vector machine [SVM, 33] and the idea is as follows. First dichotomize the continuous response Y based on its value. A pseudo binary variable is introduced as $\tilde{Y} = 1$ if Y is larger than a pre-specified cutoff c and -1 otherwise. Next, find a hyperplane of the standardized $\mathbf{X}$ that separates the two classes by training a linear SVM with respect to $(\mathbf{X}, \tilde{Y})$. Then it turns out that the normal of the hyperplane lies on $\mathcal{S}_{Y|\mathbf{X}}$ and hence SDR naturally follows. The PSVM can be readily extended to nonlinear SDR via kernel trick, just like the SVM.

1.2. Motivation

In practice, it is often observed that the data display heteroscedasticity. The heteroscedasticity itself can be of scientific importance, but is often overlooked since most existing statistical methods primarily focus only on conditional mean relationship. In principle, SDR only requires the conditional independence assumption as shown in (1.1) or (1.2), and there is no difficulty to uncover underlying heteroscedasticity in the data. However, most of SDR methods are designed mainly for the mean relationship and may be inefficient to uncover heteroscedasticity. This is illustrated in the upcoming toy example.

The quantile regression (QR) is known as a reasonable alternative to the least squares (LS) regression when errors are non-iid or have heteroscedastic variance. The QR explores the entire conditional distribution of Y given $\mathbf{X}$ by controlling target quantile levels while the LS regression focuses only on the conditional mean $E(Y \mid \mathbf{X})$. Kong and Xia [13] showed that the gradients of regression quantiles lie on $\mathcal{S}_{Y|\mathbf{X}}$ and proposed an associated SDR method, qOPG. Compared with moment-based methods such as SIR and SAVE, qOPG requires less restrictive assumptions and can identify all dimension reduction directions including those associated with a heteroscedastic structure.

The QR also has a very close connection to the SVM due to the similarity of their loss functions. However, the SVM solution depends only on a part of data, data points either close to the classification boundary or misclassified, while the QR takes into account all the data points for estimating the conditional quantile function. For this reason, the SVM is not proper to capture the heteroscedasticity and this makes PSVM inefficient for extracting information from the heteroscedasticity. Motivated by this, we propose the principal quantile regression (PQR) for SDR with heteroscedasticity.



(a) Support Vector Machine (b) Quantile Regression

FIG 1. *A motivating toy example: red solid lines are the classification boundaries estimated by the SVM in panel (a) and the regression functions estimated by the QR in panel (b).*

In order to illustrate how the SVM and QR behave differently with heteroscedastic data, we consider the following toy example with error term only: $Y_i = X_i\epsilon_i$ where $\epsilon_i \overset{iid}{\sim} N(0, 0.2)$ and $X_i \overset{iid}{\sim} \text{Uniform}(0, 5), i = 1, \cdots, 100$. Notice that the conditional variance of Y given X depends on X. For the PSVM, the continuous response Y_i is artificially dichotomized based on a given c as $\tilde{Y}_{i,c} = 1$ if $Y_i > c$ and -1 otherwise. We set two values of c_1 and c_2 as the 33.3% and 66.6% sample percentiles of Y_is, respectively and apply the SVM to $\{(X_i, \tilde{Y}_{i,c_h}), i = 1, \cdots, n\}$ for the two different $c_h, h = 1, 2$. Next, the QR is applied to $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$ with two different quantile levels, 33.3%

and 66.6%. The results are depicted in Figure 1. Panel (a) shows the SVM classification boundaries and panel (b) plots the regression functions estimated by the QR. As shown in the left panel (a), the two classification boundaries estimated from the SVM (red solid lines) are parallel with both slopes nearly zero. It clearly shows that their normals are not affected very much by the heteroscedasticity and fails the PSVM. On the other hand, the QR produces different (non-zero coefficient) slope estimates for different quantile levels since it takes into account the behaviors of all data points. See panel (b) of Figure 1. This simple example justifies the use of quantile regression as an alternative of the SVM in the presence of heteroscedasticity and provides a clear motivation of the PQR.

The rest of article is organized as follows. In Section 2, the PQR for linear SDR, which is referred to as the linear PQR, is developed and its theoretical properties and computational issues are described in details. The kernel PQR (KPQR) is proposed in Section 3 as a nonlinear extension of the linear PQR for nonlinear SDR. Finite sample performance of the proposed methods are investigated via simulation in Section 4 and real data analysis in Section 5. Discussions follow in Section 6. All the technical proofs are relegated in Appendix.

2. Linear principal quantile regression

2.1. Principal quantile regression

Let us begin with a brief introduction of the linear QR model: $Y = \alpha + \boldsymbol{\beta}^\top \mathbf{X} + \epsilon$, where the random error ϵ satisfies $P(\epsilon \leq 0 \mid \mathbf{X}) = \tau$ for a given target quantile level $\tau \in (0,1)$. The regression function $\alpha + \boldsymbol{\beta}^\top \mathbf{X}$ thus represents the τth conditional quantile of Y given $\mathbf{X}$, and α and $\boldsymbol{\beta}$ are parameters of interest. The QR does not require the error term to be *i.i.d.* and hence is often regarded as an attractive alternative to the conventional mean regression in the presence of heteroscedasticity. At the population level, the QR solves

$$(\alpha_0, \boldsymbol{\beta}_0^\top)^\top = \underset{\alpha,\boldsymbol{\beta}}{\operatorname{argmin}} E\left[\rho_\tau(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X})\right], \tag{2.1}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ denotes the check loss function. Sample estimates of α_0 and $\boldsymbol{\beta}_0$ are obtained by minimizing the empirical counterpart of (2.1).

Motivated by Li, Artemiou and Li [18] and Shin et al. [28], the τth PQR objective function at the population level is defined by

$$\Lambda_\tau(\boldsymbol{\theta}) = \boldsymbol{\beta}^\top \boldsymbol{\Sigma} \boldsymbol{\beta} + \lambda E\left[\rho_\tau\left(Y - \alpha - \boldsymbol{\beta}^\top \{\mathbf{X} - E(\mathbf{X})\}\right)\right], \tag{2.2}$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^\top)^\top$, $\boldsymbol{\Sigma} = \operatorname{cov}(\mathbf{X})$ and $\lambda > 0$ is the regularization parameter which balances the data fitting and the model complexity. Let $\boldsymbol{\theta}_{0,\tau} = (\alpha_{0,\tau}, \boldsymbol{\beta}_{0,\tau}^\top)^\top$ be the minimizer of (2.2). Then it can be shown that $\boldsymbol{\beta}_{0,\tau}$ is unbiased for the linear SDR (1.1) as follows.

Theorem 1. *Under the linearity condition that* $E(\mathbf{X} \mid \mathbf{B}^\top \mathbf{X})$ *is a linear function of* $\mathbf{B}^\top \mathbf{X}$, $\boldsymbol{\beta}_{0,\tau} \in \mathcal{S}_{Y|\mathbf{X}}$ *for any given* $\tau \in (0,1)$.

The linearity condition plays an essential role in many SDR methods. It implies $E(\boldsymbol{\beta}^\top \mathbf{X}|\mathbf{B}^\top \mathbf{X}) = \boldsymbol{\beta}^\top \mathbf{P}_{\mathbf{B}}^\top(\boldsymbol{\Sigma})\mathbf{X}$ when $\mathbf{X}$ is centered such that $E(\mathbf{X}) = 0$, where $\mathbf{P}_{\mathbf{B}}(\boldsymbol{\Sigma}) = \mathbf{B}(\mathbf{B}^\top \boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^\top \boldsymbol{\Sigma}$. It is known that the linearity condition holds when $\mathbf{X}$ is elliptically symmetric [20, 19] or when p is large [9]. It is important to note that the linearity condition is only for the marginal distribution of $\mathbf{X}$, not the conditional distribution of $Y \mid \mathbf{X}$.

The proposed PQR shares a fundamental similarity to Kong and Xia [13] that recovers $\mathcal{S}_{Y|\mathbf{X}}$ from derivatives of the conditional quantile $Y|\mathbf{X}$ with respect to $\mathbf{X}$, since $\boldsymbol{\beta}_{0,\tau}$ can be viewed as a linear approximation of derivative of the τth quantile of $Y|\mathbf{X}$. We admit Theorem 1 is a similar but weaker result than Lemma 1 of Kong and Xia [13] in the sense that the linear PQR requires the linearity condition and do not recover $\mathcal{S}_{Y|\mathbf{X}}$ exhaustively. Both drawbacks are due to the global linearity of the target function, $\alpha + \boldsymbol{\beta}^T\{\mathbf{X} - E(\mathbf{X})\}$, and are finally resolved when kernel PQR is employed. However, the linear target quantile function works reasonably well in many real applications and brings in substantial amount of computational savings as demonstrated in Section 4.1.

Since linear PQR do not possess exhaustiveness, we assume the coverage condition that $\text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{X}}$. See Cook and Ni [7] for the practical impact of the coverage condition. Now, one can recover $\mathcal{S}_{Y|\mathbf{X}}$ by obtaining multiple solutions of $\boldsymbol{\beta}_{0,\tau}$ for different values of τ. More explicitly, we consider H distinctive values of $\tau_h, h = 1, \cdots, H$, where H is assumed to be larger than d to fully recover a d-dimensional subspace. The selection of d will be further discussed in section 2.4. Let $\boldsymbol{\theta}_{0,h} = (\alpha_{0,h}, \boldsymbol{\beta}_{0,h}^\top)^\top$ denote the sequence of minimizers of (2.2) for different values of $\tau_h, h = 1, \cdots, H$, then $\mathcal{S}_{Y|\mathbf{X}}$ is estimated by the eigenvectors of $\mathbf{M}_0$ associated with non-zero eigenvalues, where

$$\mathbf{M}_0 = \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h}\boldsymbol{\beta}_{0,h}^\top.$$

Notice that unlike PSVM [18], an additional step of dichotomizing the response is not required for PQR. Instead, PQR obtains multiple solutions $\boldsymbol{\beta}_{0,h}$ by varying the quantile level parameter τ that controls the shape of the loss function ρ_τ. This makes PQR fundamentally different from PSVM whose objective function varies as the (pseudo) binary response changes while the loss function remains fixed. In fact, the idea is much similar to the principal weighted SVM [28].

In order to see the connection to the QR, suppose that $\mathbf{X}$ has $E(\mathbf{X}) = \mathbf{0}_p$ and $\text{cov}(\mathbf{X}) = \mathbf{I}_p$ where $\mathbf{0}_p$ and $\mathbf{I}_p$ are the p-dimensional zero vector and identity matrix, respectively. Then (2.2) reduces to

$$\boldsymbol{\beta}^\top \boldsymbol{\beta} + \lambda E\left[\rho_\tau\left(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X}\right)\right], \tag{2.3}$$

which can be viewed as a population version of the L_2-penalized linear QR. That is, the L_2-penalized linear QR coefficient $\boldsymbol{\beta}$ is unbiased for linear SDR

when $\mathbf{X}$ is standardized. This provides intuitive explanation why the PQR is advantageous in handling heteroscedasticity.

2.2. Finite sample estimation

Given a set of data $\{(\mathbf{X}_i, Y_i), i = 1, \cdots, n\}$, the sample PQR objective function is given by

$$\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta}) = \boldsymbol{\beta}^\top \hat{\boldsymbol{\Sigma}}_n \boldsymbol{\beta} + \frac{\lambda}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \alpha - \boldsymbol{\beta}^\top (\mathbf{X}_i - \bar{\mathbf{X}}_n) \right), \tag{2.4}$$

where $\bar{\mathbf{X}}_n$ and $\hat{\boldsymbol{\Sigma}}_n$ are sample mean and covariance matrix, respectively. Denote $\hat{\boldsymbol{\theta}}_{n,h} = (\hat{\alpha}_{n,h}, \hat{\boldsymbol{\beta}}_{n,h}^\top)^\top = \operatorname{argmin}_{\boldsymbol{\theta}} \hat{\Lambda}_{n,\tau_h}(\boldsymbol{\theta}), h = 1, \cdots, H$, for a given a grid $0 < \tau_1 < \cdots < \tau_H < 1$. The PQR candidate matrix of the linear SDR is

$$\widehat{\mathbf{M}}_n = \sum_{h=1}^{H} \hat{\boldsymbol{\beta}}_{n,h} \hat{\boldsymbol{\beta}}_{n,h}^\top. \tag{2.5}$$

The first d leading eigenvectors of (2.5) denoted by $\hat{\mathbf{V}}_n = (\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_d)$ can be considered as a PQR estimate of $\mathbf{B}$ in (1.1), the basis set of $\mathcal{S}_{Y|\mathbf{X}}$.

We remark that the optimization should be repeatedly carried out for different values of τ to obtain (2.5), which can be too computationally intensive especially when n and/or H is large. To improve computational efficiency, we consider the following transformation $\boldsymbol{\eta} = \hat{\boldsymbol{\Sigma}}_n^{1/2} \boldsymbol{\beta}$ and $\mathbf{Z}_i = \hat{\boldsymbol{\Sigma}}_n^{-1/2} (\mathbf{X}_i - \bar{\mathbf{X}}_n)$, which makes (2.4) equivalent to the (linear) kernel quantile regression [KQR, 31, 21] as follows:

$$\boldsymbol{\eta}^\top \boldsymbol{\eta} + \frac{\lambda}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \alpha - \boldsymbol{\eta}^\top \mathbf{Z}_i \right). \tag{2.6}$$

It is shown that the solution of (2.6) moves in a piecewise linear manner as τ varies, which enables us to develop an efficient algorithm that computes the entire solution trajectories of $\boldsymbol{\eta}$ (and hence $\boldsymbol{\beta}$) as a function of $\tau \in (0, 1)$ with the same computational complexity of solving a single optimization problem for a single τ [30, 27]. In order to illustrate the solution paths for the aforementioned toy example, we add four additional noise variables $X_2, \cdots, X_5 \overset{iid}{\sim} \text{Uniform}(0, 5)$ irrelevant to the response, i.e., $Y_i = (\mathbf{e}_1^\top \mathbf{X}_i)\varepsilon_i$ where $\mathbf{e}_1 = (1, 0, 0, 0, 0)^\top, \mathbf{X}_i = (X_{i1}, \cdots, X_{i5},)^\top, i = 1, \cdots, 100$. The five paths of $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_5$ as a function of τ are depicted in Figure 2. Notice that X_1 is the only signal variable whose corresponding solution profile of β_1 (red solid line) shows significantly larger variability compared to those of $\beta_2, \cdots, \beta_5$ (black dashed lines) corresponding to noise variables $X_2, \cdots, X_5$.

2.3. Large sample properties

Without loss of generality, we assume that $E(\mathbf{X}) = \mathbf{0}_p$ in this section. We define

$$m_\tau(\boldsymbol{\theta}, \mathbf{Z}) = \boldsymbol{\theta}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{\theta} + \lambda \{\rho_\tau(Y - \boldsymbol{\theta}^\top \tilde{\mathbf{X}})\},$$

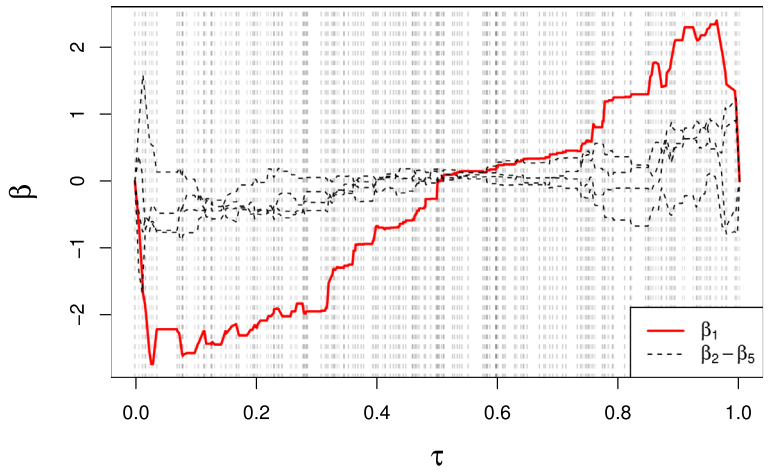


FIG 2. *Illustration of piecewise linear solution paths of the PQR: solution path of* β_1 *(red solid line) corresponding to* X_1*, the only signal variable, shows significantly larger variability compared to those of* $\beta_2, \cdots, \beta_5$ *(black dashed lines) corresponding to the noise variables* $X_2, \cdots, X_5$*.*

where $\mathbf{Z} = (\tilde{\mathbf{X}}^\top, Y)^\top$, $\tilde{\mathbf{X}} = (1, \mathbf{X}^\top)^\top$, and $\tilde{\mathbf{\Sigma}} = diag(0, \mathbf{\Sigma})$. Notice that $\Lambda_\tau(\boldsymbol{\theta}) = E\left[m_\tau(\boldsymbol{\theta}, \mathbf{Z})\right]$. For notational simplicity, we omit subscript τ if τ is fixed, and let $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$ be the minimizers of $\Lambda_\tau(\boldsymbol{\theta})$ and $\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta})$, respectively, where $\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta})$ denotes the empirical version of $\Lambda_\tau(\boldsymbol{\theta})$ based on a sample.

We first establish consistency of $\hat{\boldsymbol{\theta}}_n$ for an arbitrary given τ.

Theorem 2. *Suppose* $var(\mathbf{X}) = \Sigma$ *is positive definite,*

$$\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0 \quad \textit{in probability}$$

In order to derive asymptotic distribution of PQR solution $\hat{\boldsymbol{\theta}}_n$, a Bahadur representation is given in Theorem 3

Theorem 3. *Under the assumptions (C1)–(C4) in the Appendix*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = -n^{-1/2}\mathbf{H}_{\boldsymbol{\theta}_0}^{-1}\sum_{i=1}^{n}\mathbf{D}_{\boldsymbol{\theta}_0}(\mathbf{Z}_i) + o_p(1), \tag{2.7}$$

where

$$\mathbf{D}_\theta(\mathbf{Z}) = 2\tilde{\mathbf{\Sigma}}\boldsymbol{\theta} - \lambda[\tilde{\mathbf{X}}(\tau - \mathbb{1}\{Y \le \boldsymbol{\theta}^\top\tilde{\mathbf{X}}\})],$$
$$\mathbf{H}_\theta = 2\tilde{\mathbf{\Sigma}} - \lambda E_Y\left[f_{U|Y}(y - \alpha \mid y)E(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top|U = y)\right] \quad \textit{with } U = \boldsymbol{\theta}^\top\tilde{\mathbf{X}}.$$

As a consequence of Theorem 3, a Bahadur representation of $\hat{\boldsymbol{\beta}}_{n,h}$ is given by

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{n,h} - \boldsymbol{\beta}_{0,h}) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z}_i) + o_p(1), \tag{2.8}$$

where $\mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) = \mathbf{F}_{\boldsymbol{\theta}_0} \mathbf{D}_{\boldsymbol{\theta}_{0,h}}(\mathbf{Z})$ with $\mathbf{F}_{\boldsymbol{\theta}_0}$ denoting the last p rows of $\mathbf{H}_{\boldsymbol{\theta}_0}^{-1}$ and $\mathbf{D}_{\boldsymbol{\theta}_{0,h}}(\mathbf{Z})$ being the value of $\mathbf{D}_\theta(\mathbf{Z})$ with $\theta = \theta_0$ and $\tau = \tau_h$ for $h = 1, \cdots, H$. With the above Bahadur representation, we can establish the asymptotic normality of $\widehat{\mathbf{M}}_n$ as follows.

Theorem 4. *Under conditions (C1)–(C4),*

$$\sqrt{n}[vec(\widehat{\mathbf{M}}_n - \mathbf{M}_0)] \to N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{M}}), \tag{2.9}$$

where $\mathbf{M}_0 = \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h} \boldsymbol{\beta}_{0,h}^\top$ *and the asymptotic covariance matrix* $\boldsymbol{\Sigma}_M$ *is explicitly provided in the Appendix.*

Finally a basis of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ is estimated by the first d leading eigenvectors of $\widehat{\mathbf{M}}_n$, denoted by $\widehat{\mathbf{V}}_n = (\widehat{\mathbf{v}}_1, \cdots, \widehat{\mathbf{v}}_d)$. The asymptotic normality of $\widehat{\mathbf{V}}_n$ is established in the following corollary which is a direct consequence of Theorem 4 and Bura and Pfeiffer [1].

Corollary 1. *Let rank*$(\mathbf{M}_0) = d$ *and* $\mathbf{V}_0 = (\mathbf{v}_1, \cdots, \mathbf{v}_d)$ *be* $p \times d$ *matrix whose columns are the eigenvectors of* $\mathbf{M}_0$ *corresponding to the nonzero eigenvalues. Then*

$$\sqrt{n} vec(\widehat{\mathbf{V}}_n - \mathbf{V}_0) \to N(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{V}}),$$

where $\boldsymbol{\Sigma}_{\mathbf{V}} = (\mathbf{D}^{-1}\mathbf{V}_0^\top \otimes \mathbf{I}_p)\boldsymbol{\Sigma}_{\mathbf{M}}(\mathbf{V}_0\mathbf{D}^{-1} \otimes \mathbf{I}_p)$ *with* $\mathbf{D}$ *being a d-dimensional diagonal matrix whose diagonal elements are nonzero eigenvalues of* $\mathbf{M}_0$*. Here the operator* $\otimes$ *denotes Kronecker product.*

2.4. Determination of structural dimension

In practice the structural dimension d is typically unknown, and should be inferred from data. Following the spirit of Li, Artemiou and Li [18], we estimate it by $\hat{d}$ that maximizes

$$G_n(k; \rho, \widehat{\mathbf{M}}_n) = \sum_{j=1}^{k} \upsilon_j - \rho \frac{k \log n}{\sqrt{n}} \upsilon_1, \tag{2.10}$$

where υ_j is the jth leading eigenvalue of the candidate matrix $\widehat{\mathbf{M}}_n$ and $\rho > 0$ is a tuning parameter. By Theorem 4 and Theorem 8 of Li, Artemiou and Li [18], we can proof that $\hat{d}$ is a consistent estimator of d, i.e., $\lim_{n\to\infty} P(\hat{d} = d) = 1$.

In order to tune ρ, we propose the following procedure based on cross-validation. First, randomly split the data into the training and test sets, which are denoted by $\{(\mathbf{X}_j^{\text{tr}}, Y_j^{\text{tr}}) : j = 1, \cdots, n_{\text{tr}}\}$ and $\{(\mathbf{X}_{j'}^{\text{ts}}, Y_{j'}^{\text{ts}}) : j' = 1, \cdots, n_{\text{ts}}\}$, respectively. Note $n_{\text{ts}} = n - n_{\text{tr}}$. Then apply the linear PQR to the training set $\{(\mathbf{X}_j^{\text{tr}}, Y_j^{\text{tr}}) : j = 1, \cdots, n_{\text{tr}}\}$. Let $\widehat{\mathbf{M}}_n^{\text{tr}}$ denote the corresponding candidate matrix. For an appropriate grid of ρ given, repeat the steps 1–4 below.

1. Compute $\hat{d}_{\text{tr}} = \text{argmax}_{k \in \{1, \cdots, p\}} G_n(k; \rho, \widehat{\mathbf{M}}_n^{\text{tr}})$.
2. Transform the training predictors by $\tilde{\mathbf{X}}_j^{\text{tr}} = (\widehat{\mathbf{V}}_n^{\text{tr}})^\top \mathbf{X}_j^{\text{tr}}$, where $\widehat{\mathbf{V}}_n^{\text{tr}} = (\widehat{\mathbf{v}}_1^{\text{tr}}, \cdots, \widehat{\mathbf{v}}_{\hat{d}_{\text{tr}}}^{\text{tr}})$ are the first $\hat{d}_{\text{tr}}$ leading eigenvectors of $\widehat{\mathbf{M}}_n^{\text{tr}}$.

3. For the given $\tau_h, h = 1, \cdots, H$, apply the linear QR to $\{(\tilde{\mathbf{X}}_j^{\text{tr}}, Y_j^{\text{tr}}) : j = 1, \cdots, n_{\text{tr}}\}$ to obtain τ_hth conditional quantile function estimates of $Y|\mathbf{X}$, denoted by $\hat{f}_{\tau_h}(\mathbf{X})$.
4. Compute the total test quantile loss

$$TC(\rho) = \sum_{h=1}^{H} \sum_{j'=1}^{n_{\text{ts}}} \rho_{\tau_h} \left(Y_{j'}^{\text{ts}} - \hat{f}_{\tau_h}(\mathbf{X}_{j'}^{\text{ts}}) \right).$$

We repeat the above procedure on each fold in the cross validation and select ρ^* to be the minimizer of the sum of $TC(\rho)$ across different folds.

Finally, we propose to choose $\hat{d}$ which maximizes $G_n(k; \rho^*, \widehat{\mathbf{M}}_n)$ using the full data. This tuning method is named cross validation Bayesian information criterion (CVBIC). In Section 4.3, we investigate the numerical performance of CVBIC under a variety of combinations of p, d, and n using different models.

3. Kernel PQR for nonlinear dimension reduction

3.1. Kernel PQR (KPQR)

As a nonlinear generalization of (2.2), we replace the linear function $\boldsymbol{\beta}^\top \mathbf{X}$ with the nonlinear one $\psi(\mathbf{X})$ and propose the following objective function:

$$\Lambda_\tau(\alpha, \psi) = \text{var}(\psi(\mathbf{X})) + \lambda E\{\rho_\tau(Y - \alpha - \bar{\psi}(\mathbf{X}))\}, \tag{3.1}$$

where $\bar{\psi}(\mathbf{X}) = \psi(\mathbf{X}) - E\psi(\mathbf{X})$ for $\psi \in \mathcal{H}$. Let $(\alpha_{0,\tau}, \psi_{0,\tau})$ minimize the objective function (3.1), Theorem 5 states the unbiasedness of $\psi_{0,\tau}(\mathbf{X})$ for the nonlinear SDR (1.2) and provides a foundation of the KPQR.

Theorem 5. *Consider the identity mapping from a function in $\mathcal{H}$ to a function in $L_2(P_{\mathbf{X}})$ where $L_2(P_{\mathbf{X}}) = \{f : \int |f|^2 dP_{\mathbf{X}} < \infty\}$ with $P_{\mathbf{X}}$ the probability measure induced by $\mathbf{X}$. Assume that the mapping is continuous and $\mathcal{H}$ is a dense subset of $L_2(P_{\mathbf{X}})$. Then $\psi_{0,\tau}(\mathbf{X})$ has a one-to-one transformation that is measurable with respect to $\sigma\{\boldsymbol{\phi}(\mathbf{X})\}$, where $\sigma\{\boldsymbol{\phi}(\mathbf{X})\}$ denotes the σ-field generated by $\boldsymbol{\phi}(\mathbf{X})$.*

The concept of unbiasedness of nonlinear SDR is firstly introduced by Li, Artemiou and Li [18]. See also Lee, Li and Chiaromonte [14] for a general theory for nonlinear SDR.

3.2. Sample estimation

The objective function (3.1) is not as easy to minimize as in the linear case since the space $\mathcal{H}$ is of infinite dimensionality. To tackle this issue, we employ the reproducing kernel Hilbert space (RKHS) theory and let $\mathcal{H}_K$ be the RKHS associated with a positive definite kernel $K(\cdot, \cdot)$. Common choices of the kernel

function include the radial basis kernel $K(\mathbf{X}, \mathbf{X}') = \exp(-r\|\mathbf{X} - \mathbf{X}'\|^2), r > 0$ and the polynomial kernel $(c + \mathbf{X}^\top \mathbf{X}')^q$ with a positive integer q and $c \geq 0$. Based on the RKHS theory, the minimizer of the empirical version of (3.1) has a finite representation [12]. Namely, the solution can always be represented by the following form: $\psi(\cdot) = \boldsymbol{\alpha}^\top \mathbf{k}_n(\cdot)$ where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^\top$ and $\mathbf{k}_n = \{K(\cdot, \mathbf{X}_i) : i = 1, \cdots, n\}^\top$. However as pointed out by Li, Artemiou and Li [18] in the context of PSVM, such a finite form cannot be directly used for the KPQR because it always overfits the training data. Instead, Li, Artemiou and Li [18] introduce the following alternative:

$$\psi(\mathbf{X}) = \boldsymbol{\gamma}^\top \boldsymbol{\omega}(\mathbf{X}), \tag{3.2}$$

where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_b)^\top$, $\boldsymbol{\omega}(\mathbf{X}) = \{\omega_j(\mathbf{X}) : j = 1, \cdots, b\}^\top$, and $\omega_j(\mathbf{X})$ is the jth leading eigenfunction of the sample covariance operator Σ_n defined by $\langle \psi_1, \Sigma_n \psi_2 \rangle_{\mathcal{H}_K} = \mathrm{cov}_n\{\psi_1(\mathbf{X}), \psi_2(\mathbf{X})\}$ for $\psi_1, \psi_2 \in \mathcal{H}_K$. Here $\mathrm{cov}_n(\mathbf{X}, \mathbf{X}')$ denotes the sample covariance between $\mathbf{X}$ and $\mathbf{X}'$. By Proposition 2 of Li, Artemiou and Li [18], we have

$$\omega_j(\mathbf{X}) = \{\mathbf{k}_n(\mathbf{X})\}^\top \mathbf{w}_j / \lambda_j, \quad j = 1, \cdots b, \tag{3.3}$$

where $\mathbf{w}_j$ and λ_j are the jth leading eigenvector and eigenvalue of the matrix $(\mathbf{I}_n - \mathbf{J}_n/n)\mathbf{K}_n(\mathbf{I}_n - \mathbf{J}_n/n)$, respectively. Here $\mathbf{K}_n$ is the kernel matrix whose (i, j)th element is $K(\mathbf{X}_i, \mathbf{X}_j), i, j = 1, \cdots, n$, and $\mathbf{J}_n$ denotes the n-dimensional square matrix whose elements are all one. In fact, $\omega_j(\mathbf{X}_i)$ is closely related to the kernel principle component analysis [26] for the covariates $\mathbf{X}_i, i = 1, \cdots, n$ on the RKHS generated by $K(\cdot, \cdot)$. Essentially, the representation (3.2) proposes to restrict the full solution space corresponding to $K(\cdot, \mathbf{X}_i), i = 1, \cdots, n$ by focusing on its subspace spanned by the first b principal directions on the RKHS, to avoid over-fitting. For the choice of b, any integer between $n/3$ and $2n/3$ can be used [18]. For our KPQR, we propose to use $b = n/4$ based on our limited numerical experience.

Finally, the sample version of (3.1) is given by

$$\hat{\Lambda}_{n,\tau}(\alpha, \gamma) = \gamma^\top \boldsymbol{\Omega}^\top \boldsymbol{\Omega} \gamma + \lambda \sum_{i=1}^{n} \{\rho_\tau(Y_i - \alpha + \boldsymbol{\Omega}_i \boldsymbol{\gamma})\}, \tag{3.4}$$

where $\boldsymbol{\Omega}$ is a $(n \times b)$ matrix whose (i, j) element is $\omega_j(\mathbf{X}_i) - n^{-1} \sum_{m=1}^{n} \omega_j(\mathbf{X}_m)$ for $i = 1, \cdots, n$ and $j = 1, \cdots, b$ and $\boldsymbol{\Omega}_i$ is the ith row of $\boldsymbol{\Omega}$. The dual problem of (3.4) is even simpler quadratic programming as shown in Theorem 6.

Theorem 6. *Let* $\hat{\mathbf{v}} = (\hat{v}_1, \cdots, \hat{v}_n)^\top$ *and* $\hat{\eta} = (\hat{\eta}_1, \cdots, \hat{\eta}_n)^\top$ *denote the maximizer of the following quadratic programming*

$$\max_{v_1, \cdots, v_n, \eta_1, \cdots, \eta_n} \sum_{i=1}^{n} (v_i - \eta_i) Y_i - \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} (v_i - \eta_i)(v_j - \eta_j) P_{\boldsymbol{\Omega}}^{\{i,j\}}$$

subject to $0 \leq v_i \leq \lambda\tau$ *and* $0 \leq \eta_i \leq \lambda(1 - \tau)$ *and* $\sum_{i=1}^{n} (v_i - \eta_i) = 0$, *where* $P_{\boldsymbol{\Omega}}^{\{i,j\}}$ *is the* (i, j)*th element of* $P_{\boldsymbol{\Omega}} = \boldsymbol{\Omega}(\boldsymbol{\Omega}^\top \boldsymbol{\Omega})^{-1} \boldsymbol{\Omega}^\top$. *Then the minimizer of*

(3.4) *is given by*

$$\hat{\gamma}_n = \frac{\lambda}{2}\sum_{i=1}^{n}\hat{\nu}_i(\boldsymbol{\Omega}^\top\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}_i^\top,$$

where $\hat{\nu}_i = \hat{v}_i - \hat{\eta}_i$.

Similar to the linear PQR we can obtain the entire solution profile $\hat{\boldsymbol{\nu}} = (\hat{\nu}_1, \cdots, \hat{\nu}_n)^\top$. For a given grid $\tau_1 < \cdots < \tau_H$, we extract from the solution profile a sequence of kernel PQR solutions, $(\hat{\alpha}_{n,h}, \hat{\boldsymbol{\gamma}}_{n,h})$ corresponding to τ_h and obtain the first d leading eigenvectors, $\hat{\mathbf{V}}_n = (\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_d)$, of $\sum_{h=1}^{H}\hat{\boldsymbol{\gamma}}_{n,h}\hat{\boldsymbol{\gamma}}_{n,h}^\top$. Then the estimated d sufficient predictors evaluated at $\mathbf{X}$ is given by $\hat{\boldsymbol{\phi}}(\mathbf{X}) = \{\hat{\phi}_1(\mathbf{X}), \cdots, \hat{\phi}_d(\mathbf{X})\}^\top = \hat{\mathbf{V}}_n^\top\{\omega_1(\mathbf{X}), \cdots, \omega_b(\mathbf{X})\}^\top$.

At the best of our knowledge, there is no method developed for the estimation of the structural dimension d in nonlinear SDR. In this article, we have skipped developing the estimation of d for the kernel PQR since it requires the theoretical analysis for RKHS, which is somewhat beyond of the scope of this paper. In practice, we can consider the cross-validation idea proposed by Xia et al. [36] that selects an optimal d that results in the best (cross-validated) prediction accuracy on the reduced space. Although the idea is originally proposed in the linear SDR context, we believe that it can be readily extended to the nonlinear SDR as long as the notion of the reduced space is clearly defined at the sample level, which is the case for the kernel PQR.

4. Simulation studies

We carried out simulations to investigate the finite sample performance of the proposed linear and kernel PQR. We assume that the true structural dimension d is known for both linear and kernel PQR in Sections 4.1 and 4.2, respectively. In Section 4.3, we demonstrate the performance of the CVBIC procedure for estimating d for the linear PQR.

4.1. Linear sufficient dimension reduction

We consider the following six models:

(L1) $Y = X_1 + 0.5X_2\epsilon$;
(L2) $Y = X_1 + 0.5\exp(0.15X_2)\epsilon$;
(L3) $Y = X_1^3 + 0.5\exp(X_2)\epsilon$;
(L4) $Y = exp(X_1) - 1.05 + 0.5\exp(X_2)\epsilon$;
(L5) $Y = \frac{X_1}{0.5+(X_2+1.5)^2} + \epsilon$;
(L6) $Y = X_1(X_1 + X_2 + 1) + (X_3 + 2)\epsilon$,

where $\epsilon \sim N(0, 1)$, $p = 10$, 20, 30 and the sample size n is taken to be 300. Note that only two predictors X_1 and X_2 are informative (i.e., $d = 2$) in models (L1)-(L5) and there is a third informative predictor X_3 (i.e., $d = 3$) in (L6).

Additional uninformative predictors are included and the total number of predictors is denoted by p. Each predictor is independently generated from $U(0.1, 5)$ for models (L1) and (L2), from $U(-1, 1)$ for (L3) and (L4), and from $N(0, 1)$ for (L5) and (L6). In particular, model (L5) is taken from Li [16] and model (L6) is from an example of Li [16] by introducing heteroscedasticity.

We compare the linear PQR with SIR, PSVM and qOPG. To evaluate the performance of each method, we use the following distance measure

$$dist(\mathbf{B}_1, \mathbf{B}_2) = ||\mathbf{P}_{\mathbf{B}_1} - \mathbf{P}_{\mathbf{B}_2}||_F, \tag{4.1}$$

where $\mathbf{P}_{\mathbf{A}}$ denotes the orthogonal projection matrix onto span($\mathbf{A}$) and $\|\mathbf{A}\|_F$ is the Frobenius norm of a matrix $\mathbf{A}$.

The number of slices is fixed at 10 for SIR and PSVM. Our choice is in line with the usual practice in the SDR literature for such a sample size. For linear PQR and qOPG, to be fair we set the number of different quantile levels to be 10 as well. For simplicity, we also refer to the number of τ as number of slices from now on. Note that both PSVM and PQR involve a cost parameter λ. We tried different values of λ and observed that the performance is not overly sensitive to the value of λ. The reported values are from the best case that each method can achieve. Table 1 contains the averaged distance measure (4.1) over 100 independent repetitions.

We observe that the linear PQR performs consistently better than SIR and PSVM for all scenarios under consideration. The improvement over SIR and PSVM is dramatic for models (L1)–(L4). Note that models (L1)–(L4) have heteroscedasticity and PQR shows promising performance as expected. Even for model (L5) without heteroscedasticity, PQR performs very competitively in comparison to SIR and PSVM. Although the improvement compared to the qOPG turns out not to be significant, it reduces the computational time dramatically. The qOPG is more computationally intensive than all other methods. It takes the qOPG 478.1 minutes to estimate the central subspace when $p = 10$, and up to 3645 minutes for $p = 30$, while the PQR takes less than 5 minutes. Figure 3 shows computing time of the two methods for different p on model (L2). Under model (L6) the linear PQR performs unsatisfactory. This is because the regression function is approximately symmetric about the origin, in which both SIR and the linear PSVM fail badly as well. This motivates the kernel version.

In the review process, one referee asked us to include comparison to SAVE [8], contour regression[CR, 23], directional regression[DR, 22]. The performance of these three methods are summarized in the last three columns of Table 1. It shows that our new method PQR outperforms all these three methods as well.

4.2. Nonlinear sufficient dimension reduction

To investigate performance of nonlinear SDR methods, we consider the following six models:

TABLE 1. *Performance of the linear SDR methods: Averaged Frobenius norm distances* (4.1) *over 100 independent repetitions. Corresponding standard deviations are given in parentheses.*

Model	p	SIR	PSVM	qOPG	PQR	CR	SAVE	DR
(L1)	10	0.814 (.206)	1.130 (.228)	0.314(.074)	0.280 (.074)	1.225 (.179)	1.299 (.171)	0.908 (.249)
	20	1.115 (.178)	1.185 (.169)	0.498(.094)	0.491 (.096)	1.363 (.095)	1.894 (.073)	1.282 (.164)
	30	1.245 (.145)	1.253 (.113)	0.667(.095)	0.672 (.096)	1.426 (.073)	1.948 (.036)	1.412 (.110)
(L2)	10	1.306 (.131)	1.332 (.098)	0.821(.183)	0.796 (.186)	1.314 (.111)	1.339 (.099)	1.319 (.109)
	20	1.380 (.065)	1.371 (.084)	1.054(.145)	1.055 (.146)	1.392 (.055)	1.852 (.108)	1.403 (.046)
	30	1.399 (.052)	1.400 (.052)	1.186(.118)	1.174 (.124)	1.420 (.035)	1.959 (.031)	1.430 (.031)
(L3)	10	0.700 (.132)	0.986 (.244)	0.514(.094)	0.487 (.086)	1.000 (.209)	1.529 (.198)	1.012 (.240)
	20	0.990 (.113)	1.159 (.162)	0.774(.109)	0.759 (.116)	1.351 (.143)	1.893 (.074)	1.466 (.188)
	30	1.229 (.136)	1.274 (.120)	0.963(.102)	0.952 (.099)	1.503 (.111)	1.943 (.038)	1.633 (.141)
(L4)	10	0.830 (.216)	1.156 (.237)	0.419(.095)	0.408 (.089)	1.118 (.198)	1.313 (.164)	0.960 (.253)
	20	1.084 (.180)	1.203 (.181)	0.636(.092)	0.642 (.096)	1.317 (.116)	1.886 (.079)	1.300 (.154)
	30	1.248 (.137)	1.268 (.121)	0.799(.096)	0.801 (.099)	1.407 (.086)	1.955 (.035)	1.426 (.092)
(L5)	10	1.137(0.223)	1.189(0.210)	1.007(.255)	1.077(0.185)	1.283 (.197)	1.731 (.150)	1.264 (.247)
	20	1.384(0.153)	1.371(0.370)	1.440(.116)	1.223(0.336)	1.522 (.118)	1.904 (.073)	1.587 (.179)
	30	1.533(0.121)	1.444(0.189)	1.580(.075)	1.469(0.181)	1.639 (.106)	1.943 (.047)	1.713 (.123)
(L6)	10	1.567(0.146)	1.696(0.143)	1.131(.243)	1.409(0.154)	1.598 (.184)	1.763 (.144)	1.588 (.201)
	20	1.835(0.119)	1.895(0.100)	1.559(.141)	1.737(0.106)	1.874 (.102)	2.077 (.121)	1.902 (.117)
	30	1.980(0.099)	1.992(0.114)	1.839(.111)	1.895(0.096)	2.016 (.076)	2.175 (.097)	2.042 (.083)

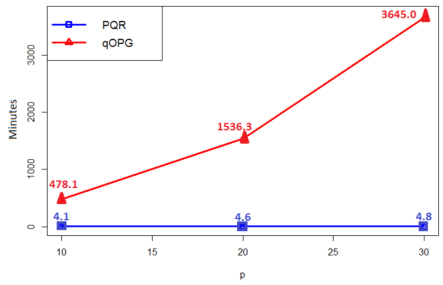


FIG 3. *Comparison of the execution time (in minutes) for the qOPG (red solid line) and PQR (blue solid line) under model (L2).*

(Single-index location-only models with $d = 1$)

(N11) $Y = \sqrt{\phi_1}\log(\sqrt{\phi_1}) + 0.5\epsilon$ with $\phi_1 = \sqrt{X_1^2 + X_2^2}$,
(N12) $Y = \phi_1 + 0.2\epsilon$ with $\phi_1 = X_1/\{0.5 + (X_2 + 1.5)^2\}$,
(N13) $Y = \phi_1 + 0.2\epsilon$ with $\phi_1 = \sin(X_1) + \sin(X_2)$.

(Two-index location-scale models with $d = 2$)

(N21) $Y = \phi_1 + \phi_2\epsilon$ with $\phi_1 = \sin(X_1 + X_2)$ and $\phi_2 = \exp(X_2)/10$,
(N22) $Y = \phi_1 + \phi_2\epsilon$ with $\phi_1 = \exp(X_1 + X_2) - 1.05$ and $\phi = \exp(X_2)/5$,
(N23) $Y = \phi_1 + \phi_2\epsilon$ with $\phi_1 = X_1^2 + X_2^2$ and $\phi_2 = \sin(X_2)/2$.

We set $p = 10$ and $n = 100$. The predictor $\mathbf{X}$ is independently generated from $U(-1, 1)$ for models (N22) and (N23) and from standard normal distribution for the rest four models. We compare the kernel PQR (KPQR) with the kernel SIR [KSIR, 34] and the kernel PSVM [KPSVM, 18]. For all of these kernel-based methods, we employ Gaussian kernel $K(\mathbf{X}, \mathbf{X}') = \exp\{-||\mathbf{X}-\mathbf{X}'||^2/(2\sigma^2)\}$ with σ chosen as the median pairwise distance of predictors in the data.

It is not appropriate to use (4.1) to evaluate the performance of nonlinear SDR methods and we consider two alternatives. The first is the distance correlation [29] between $\hat{\boldsymbol{\phi}}(\mathbf{X})$ and $\boldsymbol{\phi}(\mathbf{X})$, denoted by $\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\phi}(\mathbf{X})\}$. To avoid potential overfitting, we generate an independent test set of size 1000, and the distance correlations are evaluated over the test set. The second one measures prediction performance of the regression model built on the dimension reduction space. Toward this we fit a nonparametric regression model of Y on $\hat{\boldsymbol{\phi}}(\mathbf{X})$ via local constant smoothing from the training set. Then the coefficient of determination, $R^2 = \mathrm{Cor}(Y, \hat{Y})$ over the independent test set can regarded as a reasonable performance measure for nonlinear SDR methods. Here $\hat{Y}$ denotes predicted value of the test Y from the nonparametric regression model.

For both measure, the larger values indicate better performance in terms of SDR. For both KPQR and KPSVM, we try different values of cost parameter λ and set it to the value which gives best performance. Results over 100 independent repetitions are reported in Table 2. It is observed that the proposed KPQR outperforms KSIR and KPSVM under all scenarios under consideration

including both heteroscedastic and homoscedastic models. In the review process, one referee pointed out that it is not easy to tell whether the improvement over KPSVM is significant. While revising, we have performed paired t tests, which show that the improvement over KPSVM are significant with p-value less than .0001 for all models and for both distance correlation and coefficient of determination criteria.

TABLE 2
Performance of nonlinear SDR methods: averaged distance correlation and R^2 measures for independent test sets over 100 independent repetitions. Corresponding standard deviations are given in parentheses.

Model	$\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\phi}(\mathbf{X})\}$			R^2		
	KSIR	KPSVM	KPQR	KSIR	KPSVM	KPQR
(N11)	.203 (.074)	.494 (.062)	.508 (.061)	.042 (.094)	.302 (.085)	.322 (.085)
(N12)	.697 (.046)	.683 (.030)	.700 (.029)	.399 (.113)	.451 (.068)	.482 (.067)
(N13)	.855 (.150)	.854 (.022)	.869 (.019)	.759 (.164)	.738 (.038)	.761 (.034)
(N21)	.679 (.105)	.857 (.016)	.868 (.015)	.712 (.071)	.828 (.025)	.834 (.022)
(N22)	.620 (.134)	.809 (.018)	.821 (.017)	.670 (.121)	.732 (.044)	.747 (.043)
(N23)	.328 (.108)	.549 (.047)	.557 (.048)	.250 (.218)	.458 (.070)	.464 (.071)

To provide better snapshot of the PQR with heteroscedasticity, we additionally consider error-only models (i.e. $d = 1$) as follows.

(N31): $Y = (X_1 + X_2 + 0.5)^2\epsilon$,
(N32): $Y = (X_1 + X_2)^2\epsilon$,
(N33): $Y = (X_1^2 + X_2^2)\epsilon$,
(N34): $Y = (X_1^3 + X_2^3)\epsilon$,

where $\epsilon \sim N(0, 0.5)$ and $\mathbf{X} \sim N_p(\mathbf{0}_p, \mathbf{I}_p)$ with $p = 10$ and $n = 100$. Averaged distance correlation $\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\phi}(\mathbf{X})\}$ for test sets over 100 repetitions are reported in Table 3. Note that R^2 measure does not make sense because there is no mean relations between Y and $\mathbf{X}$. KPQR shows clear improvement for capturing signals from heteroscedasticity.

4.3. Estimation of structural dimension

We now investigate the performance of the proposed CVBIC procedure to estimate the structural dimension from the linear PQR candidate matrix. Table 4 reports the empirical probabilities (in percentage) that the CVBIC correctly estimates d. One can observe that the numerical performance of the proposed CVBIC approach for the linear PQR is quite promising especially when n is large enough.

5. Real data analysis

We apply the proposed method to the Boston housing data [10]. The dependent variable Y is the logarithm of the median value of owner occupied homes in each

TABLE 3
Performance of nonlinear SDR methods: averaged distance correlation measures over 100 independent repetitions for error-only models. Corresponding standard deviations are given in parentheses.

Model	$\mathrm{dCor}\{\hat{\boldsymbol{\phi}}(\mathbf{X}), \boldsymbol{\phi}(\mathbf{X})\}$		
	KSIR	KPSVM	KPQR
(N31)	.291 (.144)	.169 (.094)	.493 (.055)
(N32)	.204 (.124)	.151 (.082)	.426 (.067)
(N33)	.192 (.084)	.161 (.083)	.451 (.061)
(N34)	.256 (.116)	.191 (.084)	.373 (.048)

TABLE 4
Empirical probabilities (in percentage) that the CVBIC correctly estimates true d over 100 independent repetitions.

d	Model	n	$p = 10$	$p = 20$	$p = 30$
2	(L1)	200	70%	48%	24%
		500	89%	79%	42%
	(L2)	200	69%	50%	29%
		500	89%	79%	51%
	(L3)	200	66%	45%	24%
		500	89%	79%	56%
	(L4)	200	79%	47%	33%
		500	96%	76%	53%
	(L5)	200	97%	72%	58%
		500	100%	100%	98%
3	(L6)	200	84%	65%	41%
		500	93%	80%	57%

of the 506 census tracts in Boston Standard Metropolitan Statistical Areas. There are 13 predictors (Table 5). We do not consider the houses with tract bounds the Charles river, so X_4 is excluded in the following analysis. The sample size ends up to be 471 with each observation/case representing a census tract.

Y	median value of owner-occupied homes in \$1000's
X_1	per capita crime rate by town
X_2	proportion of residential land zoned for lots over 25,000 sq.ft.
X_3	proportion of non-retail business acres per town
X_4	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
X_5	nitric oxides concentration (parts per 10 million)
X_6	average number of rooms per dwelling
X_7	proportion of owner-occupied units built prior to 1940
X_8	weighted distances to five Boston employment centres
X_9	index of accessibility to radial highways
X_{10}	full-value property-tax rate per $\$10,000$
X_{11}	pupil-teacher ratio by town
X_{12}	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
X_{13}	% lower status of the population

TABLE 5
Variables in Boston housing data

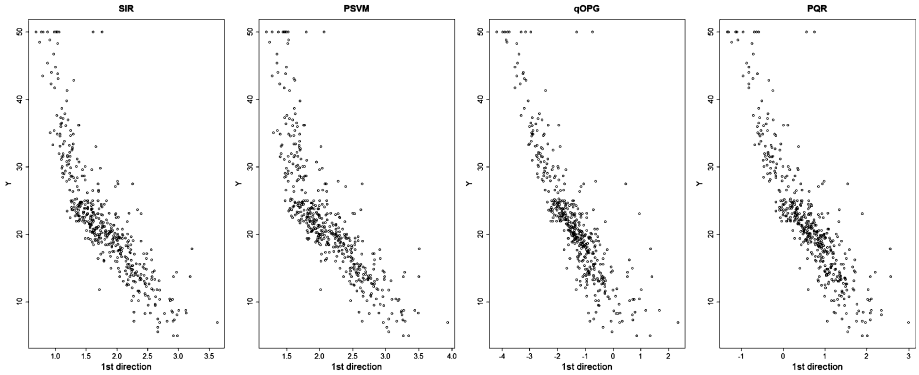


FIG 4. *Scatter plots of the response variable vs the first SDR predictor estimated by SIR, PSVM($\lambda = 0.1$), qOPG and PQR($\lambda = 0.01$).*

Method	λ							
	1e-6	.01	.05	0.1	0.3	1	10	100
PSVM	.809	.813	.834	**.836**	.828	.809	.738	.723
PQR	.843	**.869**	.866	.866	.864	.864	.864	.864

TABLE 6

The distance correlations between the dependent variable Y and the first SDR predictor estimated by PSVM and PQR for different values of λ.

We first apply linear SDR methods considered in Section 4. Detailed settings of these methods are the same as done for simulated data in Section 4. In order to estimate the structure dimension d in the PKQR, we employ the four-fold CVBIC approach and select an optimal ρ as 0.02. The corresponding BIC-type criterion in (2.10) is maximized at $k = 1$ and gives $\hat{d} = 1$. Since the true $\mathbf{B}$ is not available for the real data, we use the distance correlation between sufficient predictors projected onto the estimated $\mathcal{S}_{Y|\mathbf{X}}$, $\hat{\mathbf{b}}_1^\top \mathbf{X}$ and Y, $\mathrm{dCor}(\hat{\mathbf{b}}_1 \mathbf{X}, \mathrm{Y})$ as a performance measure. Table 6 reports the distance correlations from the SDR results of PSVM and PQR under different values of λ. The performance of PQR is not overly sensitive to λ, and their best performances are 0.836 and 0.869 respectively. The distance correlations of SIR and qOPG are 0.856 and 0.865, respectively. Figure 4 depicts the scatter plots of Y against $\hat{\mathbf{b}}_1 \mathbf{X}$ obtained from different SDR methods. Based on the distance measurements and scatter plots, we see that PQR achieves a slightly better prediction comparing with SIR and PSVM. Comparing with qOPG, PQR has a great computational advantage. It takes 6.9 seconds for PQR while it takes 9.7 minutes for qOPG to estimate $\mathbf{B}$.

Toward nonlinear SDR, we also apply KPQR, KSIR and KPSVM with the Gaussian kernel to the Boston housing data. First, we randomly divide the dataset into nonoverlapping training and test data sets with 300 and 171 observations respectively. The performance measures used in Section 4.2 for nonlinear SDR methods are then computed. We repeat this process on 100 random

	KSIR	KPSVM	KPQR
R^2	**.507** (.193)	**.580** (.064)	**.608** (.065)
dCor	**.591** (.040)	**.777** (.039)	**.791** (.036)

TABLE 7

The Boston housing data: averaged R^2 and distance correlations computed from kernel SDR methods.

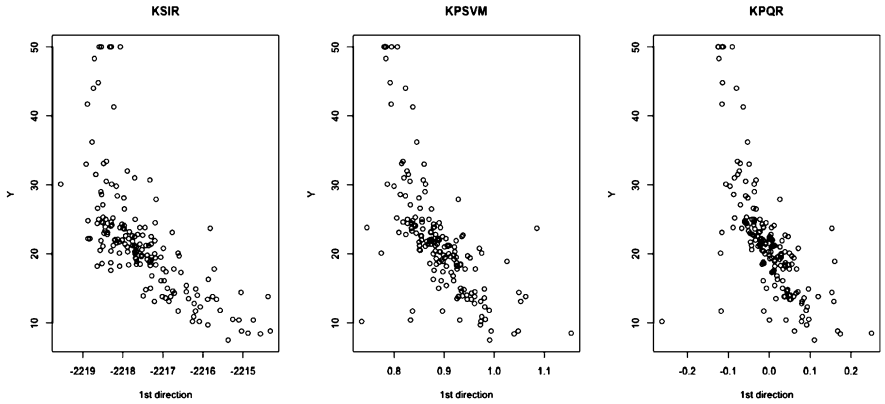


FIG 5. *The Boston housing data: scatter plots of the response variable vs the estimated first SDR predictor in the test set for KSIR (left), KPSVM (center), and KPQR (right).*

splitting of training and test sets under a wide range of cost parameters and with the number of slices fixed at 10. Table 7 reports the averaged R^2 and $\mathrm{dCor}\{\hat{\phi}(\mathbf{X}), Y\}$ for test set under selected cost parameter values. The paired t tests show that the improvements of our new method KPQR over both KSIR and KPSVM are significant in terms of both distance correlation and coefficient of determination criteria. Figure 5 depict the scatter plots of Y against $\hat{\boldsymbol{\phi}}(\mathbf{X})$, and its corresponding predicted value $\hat{Y}$ over the test set for one random splitting. They indicate that PQR achieves a slightly better prediction compared with KSIR and KPSVM. Between the linear and kernel methods, we think that linear methods give a better prediction for the Boston housing data set.

6. Conclusion

In this paper we proposed a new SDR method, PQR by exploiting QR that is particularly useful in the presence of heteroscedasticity. Compared to PSVM, PQR greatly improves the performance in capturing the conditional variance structure of $Y|\mathbf{X}$. Thanks to the flexibility of SDR under very mild assumption, PQR possess a wide range of applicability in a variety of applications in which heteroscedasticity itself is of interest. Our limited numerical experience shows that PQR still performs very competitively even for the case without heteroscedasticity.

Appendix

Proof of Theorem 1

Assume $E(\mathbf{X}) = \mathbf{0}$ without loss of generality. The linear PQR objective function (2.2) is

$$\Lambda_\tau(\boldsymbol{\theta}) = var(\boldsymbol{\beta}^\top \mathbf{X}) + \lambda E\left[\rho_\tau\left(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X}\right)\right].$$

The first term is decomposed as

$$var(\boldsymbol{\beta}^\top \mathbf{X}) = var[E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X})] + E[var(\boldsymbol{\beta}^\top \mathbf{X} | \mathbf{B}^\top \mathbf{X})] \geq var[E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X})]. \tag{A.1}$$

Now, the second term is

$$\begin{aligned} E\{\rho_\tau(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X})\} &= E[E\{\rho_\tau(Y - \alpha - \boldsymbol{\beta}^\top \mathbf{X}) | \mathbf{B}^\top \mathbf{X}, Y\}] \\ &\geq E[\rho_\tau(Y - \alpha - E(\boldsymbol{\beta}^\top \mathbf{X} | \mathbf{B}^\top \mathbf{X})). \end{aligned} \tag{A.2}$$

The last inequality holds since $\rho_\tau(\cdot)$ is a convex function. Thus, the (possibly nonunique) minimum is achieved at $E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X})$. By the Linearity condition, we have $E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}) = \boldsymbol{\beta}^\top \mathbf{P}_{\mathbf{B}}^\top(\boldsymbol{\Sigma})\mathbf{X}$ and hence $E(\boldsymbol{\beta}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}) \in \mathcal{S}_{Y|\mathbf{X}}$.

Suppose $\tilde{\boldsymbol{\beta}} \notin \mathcal{S}_{Y|\mathbf{X}}$, then $var(\tilde{\boldsymbol{\beta}}^\top \mathbf{X} \mid \mathbf{B}^\top \mathbf{X}) > 0$, and the inequality in (A.1) must be strict. Therefore $\tilde{\boldsymbol{\beta}}$ cannot be the minimizer. ■

Proof of Theorem 2

It is obvious that both $\Lambda_\tau(\boldsymbol{\theta})$ and $\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta})$ are strictly convex functions of $\boldsymbol{\theta}$ since $\boldsymbol{\Sigma}$ is positive definite and the check loss is the strictly convex. $\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta}) \to \Lambda_\tau(\boldsymbol{\theta})$ in probability for each θ. Applying Convexity Lemma of Pollard [25], we have $sup|\hat{\Lambda}_{n,\tau}(\boldsymbol{\theta}) - \Lambda_\tau(\boldsymbol{\theta})| \to 0$, in probability. By Theorem 2.1 of Newey and McFadden [24], the consistency of $\boldsymbol{\theta}_n$ follows. ■

The following regularity conditions are required to study the asymptotic properties of the linear PQR.

(C1) $\mathbf{X}$ has an open and convex support and satisfies that $E(||\mathbf{X}||^2) < \infty$.
(C2) The conditional distribution $\mathbf{X}|Y = y$ is dominated by the Lebesgue measure for $y \in \mathbb{R}$.
(C3) For arbitrary vector $\boldsymbol{\beta}, \boldsymbol{\delta} \in \mathbb{R}^p$, define $U = \boldsymbol{\beta}^\top \mathbf{X}$ and $V = \boldsymbol{\delta}^\top \mathbf{X}$. A map $u \to E(\mathbf{X} \mid U = u, V, Y) f_{U|V,Y}(U = u \mid V, Y)$ is continuous for any given $V \in \mathbb{R}$, where $f_{U|V,Y}$ denotes the conditional density of U given V and Y.
(C4) Given $U = u$, there exists a nonnegative $\mathbb{R}^{p+1}$-valued function $\mathbf{c}(V, Y)$ such that $E[\mathbf{c}(V, Y)] < \infty$ and $E(\mathbf{X} \mid U = u, V, Y) f_{U|V,Y}(U = u \mid V, Y) < \mathbf{c}(V, Y)$ where the inequality holds component-wisely.

Proof of Theorem 3

Notice that $\rho_\tau(a) = \tau a - [a]_+$ with $[a]_+ = \max(a, 0)$, and hence $\Lambda_\tau(\boldsymbol{\theta}) = E[m_\tau(\boldsymbol{\theta}, \mathbf{Z})]$ where $m_\tau(\boldsymbol{\theta}, \mathbf{Z}) = \boldsymbol{\theta}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{\theta} + \lambda\{(Y - \boldsymbol{\theta}^\top \tilde{\mathbf{X}})\tau + [\boldsymbol{\theta}^\top \tilde{\mathbf{X}} - Y]_+\}$. We first claim the following:

(a) $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$ satisfies the Lipschitz condition with respect to $\boldsymbol{\theta}$. That is, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ there exists an integrable function $Q(\mathbf{Z})$ such that

$$|m_\tau(\boldsymbol{\theta}_1, \mathbf{Z}) - m_\tau(\boldsymbol{\theta}_2, \mathbf{Z})| \le Q(\mathbf{Z})||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||. \tag{A.3}$$

(b) For every $\boldsymbol{\theta}$, $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$ is differentiable for almost every $\mathbf{Z}$.

(c) $\Lambda_\tau(\boldsymbol{\theta})$ is twice differentiable with respect to $\boldsymbol{\theta}$ with the gradient vector $\mathbf{D}_\theta$ and Hessian matrix $\mathbf{H}_\theta$ given by

$$\mathbf{D}_\theta(\mathbf{Z}) = 2\tilde{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \lambda[\tilde{\mathbf{X}}(\tau - \mathbb{1}\{Y \le \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\})],$$
$$\mathbf{H}_\theta = 2\tilde{\boldsymbol{\Sigma}} - \lambda E_Y\left[f_{U|Y}(y - \alpha \mid y) E(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top | U = y)\right] \text{ with } U = \boldsymbol{\theta}^\top \tilde{\mathbf{X}}.$$

Finally Theorem 3 follows from the consistency established in Theorem 2 and the claims (a)-(c) by applying Theorem 5.23 of Van der Vaart [32].

– Proof of (a)

It suffice to show that the second term of $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$ satisfies the Lipschitz condition. Let $m^*_\tau(\boldsymbol{\theta}, \mathbf{Z}) = (Y - \boldsymbol{\theta}^\top \tilde{\mathbf{X}})\tau + [\boldsymbol{\theta}^\top \tilde{\mathbf{X}} - Y]_+$. Then for any $\boldsymbol{\theta}_i = (\alpha_i, \boldsymbol{\beta}_i) \in \Theta$, $i = 1, 2$, we have

$$\begin{aligned}
& m^*_\tau(\boldsymbol{\theta}_1, \mathbf{Z}) - m^*_\tau(\boldsymbol{\theta}_2, \mathbf{Z}) \\
= \;& (\alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X} - \alpha_1 - \boldsymbol{\beta}_1^\top \mathbf{X})\tau + [\alpha_1 + \boldsymbol{\beta}_1^\top \mathbf{X} - Y]_+ - [\alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X} - Y]_+ \\
\le \;& (\alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X} - \alpha_1 - \boldsymbol{\beta}_1^\top \mathbf{X})\tau + |\alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X} - \alpha_1 - \boldsymbol{\beta}_1^\top \mathbf{X}| \\
\le \;& |\alpha_2 + \boldsymbol{\beta}_2^\top \mathbf{X} - \alpha_1 - \boldsymbol{\beta}_1^\top \mathbf{X}|(1 + \tau) \\
\le \;& (1 + ||\mathbf{X}||^2)^{1/2}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||(1 + \tau).
\end{aligned}$$

The first inequality holds because $u_+ - v_+ \le |u_+ - v_+| \le |u - v|$ and the last inequality holds by Cauchy Schwarz inequality. By (C1), $E[(1 + ||\mathbf{X}||^2)^{1/2}||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||(1 + \tau)] < \infty$ and hence $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$ satisfies the Lipschitz condition.

– Proof of (b)

The first term $\boldsymbol{\theta}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{\theta}$ is differentiable. Let $N_\theta(m^*_\tau)$ be the set of $\mathbf{z}$ for which the function m^*_τ is not differentiable at $\boldsymbol{\theta}$, i.e.,

$$N_\theta(m^*_\tau) = \{m^*_\tau(\cdot, \mathbf{z}) \text{ is not differentiable at } \theta\}.$$

Under condition (C2),

$$P[\mathbf{Z} \in N_\theta(m^*_\tau)] = \int_{-\infty}^{\infty} f(y) P(\mathbf{X} \in \{x : \alpha + \boldsymbol{\beta}^\top x = Y\} | Y = y) dy = 0.$$

Therefore, $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$ is almost sure differentiable with respect to any $\boldsymbol{\theta} \in \Theta$.

– **Proof of (c)**
Under (a) and (b), we can compute the gradient vector of $m_\tau(\boldsymbol{\theta}, \mathbf{Z})$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \Lambda_\tau(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} E[m_\tau(\boldsymbol{\theta}, \mathbf{Z})] \\
&= E[\frac{\partial}{\partial \boldsymbol{\theta}} m_\tau(\boldsymbol{\theta}, \mathbf{Z})] \\
&= E[\frac{\partial}{\partial \boldsymbol{\theta}} \{\boldsymbol{\theta}^\top \tilde{\boldsymbol{\Sigma}} \boldsymbol{\theta} + \lambda\{(Y - \boldsymbol{\theta}^\top \tilde{\mathbf{X}})\tau + [\boldsymbol{\theta}^\top \tilde{\mathbf{X}} - Y]_+\}\}] \\
&= 2\tilde{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \lambda E[\tilde{\mathbf{X}}\tau - \tilde{\mathbf{X}} \mathbb{1}\{Y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\}]
\end{aligned}$$

by applying Lemma 2 of Li, Artemiou and Li [18]. The second-order derivative is given by

$$\begin{aligned}
\mathbf{H}_\theta = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \Lambda_\tau(\boldsymbol{\theta}) =& \frac{\partial}{\partial \boldsymbol{\theta}} \{2\tilde{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \lambda E[\tilde{\mathbf{X}}\tau - \tilde{\mathbf{X}} \mathbb{1}\{Y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\}]\} \\
=& 2\tilde{\boldsymbol{\Sigma}} + \lambda \frac{\partial}{\partial \boldsymbol{\theta}} E[\tilde{\mathbf{X}} \mathbb{1}\{Y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\}] \\
=& 2\tilde{\boldsymbol{\Sigma}} + \lambda \frac{\partial}{\partial \boldsymbol{\theta}} E\{E[\tilde{\mathbf{X}} \mathbb{1}\{y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\} \mid Y = y]\} \\
=& 2\tilde{\boldsymbol{\Sigma}} + \lambda \int_{-\infty}^{\infty} f(y) \frac{\partial}{\partial \boldsymbol{\theta}} E[\tilde{\mathbf{X}} \mathbb{1}\{y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\} \mid Y = y] dy.
\end{aligned}$$

Applying Lemma 4 and 5 of Li, Artemiou and Li [18] under (C3)–(C4), we have

$$\frac{\partial}{\partial \boldsymbol{\theta}} E[\tilde{\mathbf{X}} I\{y \leq \boldsymbol{\theta}^\top \tilde{\mathbf{X}}\} \mid Y = y] = -f_{\beta^\top \mathbf{X} \mid Y}(y - \alpha \mid Y = y) E[\tilde{\mathbf{X}} \tilde{\mathbf{X}}^\top \mid \boldsymbol{\theta}^\top \tilde{\mathbf{X}} = y]$$

and hence

$$\mathbf{H}_\theta = 2\tilde{\boldsymbol{\Sigma}} - \lambda E_Y \left[f_{U \mid Y}(y - \alpha \mid y) E[\tilde{\mathbf{X}} \tilde{\mathbf{X}}^\top \mid U = Y] \right],$$

where $U = \theta^\top \tilde{\mathbf{X}}$ and $f_{U \mid Y}$ is the conditional distribution of U given Y. ■

Proof of Theorem 4

Let $\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z}_i)$, the sample average of $\mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z})$. Since $E[\mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z})] = \mathbf{0}$, $\bar{\mathbf{S}}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) = O_p(n^{-1/2})$. By (12),

$$\begin{aligned}
&vec(\widehat{\mathbf{M}}_n) - vec(\mathbf{M}_0) \\
&= \sum_{h=1}^H \boldsymbol{\beta}_{n,h} \otimes \boldsymbol{\beta}_{n,h} - \sum_{h=1}^H \boldsymbol{\beta}_{0,h} \otimes \boldsymbol{\beta}_{0,h}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{h=1}^{H} \{\boldsymbol{\beta}_{0,h} - \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) + o_p(n^{-1/2})\} \otimes \{\boldsymbol{\beta}_{0,h} - \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) + o_p(n^{-1/2})\} \\
&\quad - \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h} \otimes \boldsymbol{\beta}_{0,h} \\
&= - \sum_{h=1}^{H} \{\boldsymbol{\beta}_{0,h} \otimes \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) + \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) \otimes \boldsymbol{\beta}_{0,h}\} \\
&\quad + \sum_{h=1}^{H} \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) \otimes \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) + o_p(n^{-1/2}) \\
&= - \sum_{h=1}^{H} \{\boldsymbol{\beta}_{0,h} \otimes \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) + \bar{S}_n(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) \otimes \boldsymbol{\beta}_{0,h}\} + o_p(n^{-1/2}),
\end{aligned}
$$

where operator $\otimes$ denotes Kronecker product. Define a communication matrix $\mathbf{T}_{u,v} \in R^{uv \times uv}$ that satisfies:

- $\mathbf{T}_{i_1,i_2} = \mathbf{T}_{i_2,i_1}^{\top}$,
- $\mathbf{A} \otimes \mathbf{B} = \mathbf{T}_{i_1,i_3}(\mathbf{B} \otimes \mathbf{A})\mathbf{T}_{i_4,i_2}$ for $\mathbf{A} \in \mathbb{R}^{i_1,i_2}$ and $\mathbf{B} \in \mathbb{R}^{i_3,i_4}$,

and hence $\mathbf{T}_{u,v} vec(\mathbf{A}) = vec(\mathbf{A}^{\top})$ for $\mathbf{A} \in \mathbb{R}^{u \times v}$. Therefore

$$
\sqrt{n}\{vec(\widehat{\mathbf{M}}_n) - vec(\mathbf{M}_0)\} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{(\mathbf{I}_{p^2} + \mathbf{T}_{p,p}) \sum_{h=1}^{H} \boldsymbol{\beta}_{0,h} \otimes \mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z}_i)\} + o_p(1), \tag{A.4}
$$

where $\mathbf{I_{p^2}}$ is a p^2-dimensional identity matrix. Finally, the covariance matrix $\mathbf{\Sigma_M}$ is given by

$$
\mathbf{\Sigma_M} = (\mathbf{I}_{p^2} + \mathbf{T}_{p,p}) \sum_{h=1}^{H} \sum_{h'=1}^{H} \{\boldsymbol{\beta}_{0,h} \boldsymbol{\beta}_{0,h'}^{\top} \otimes E[\mathbf{S}(\boldsymbol{\theta}_{0,h}, \mathbf{Z}) \mathbf{S}^{\top}(\boldsymbol{\theta}_{0,h'}, \mathbf{Z})]\} (\mathbf{I}_{p^2} + \mathbf{T}_{p,p}). \tag{A.5}
$$

■

Proof of Theorem 5

Similar to what we did in the proof of Theorem 1, we have

$$
\begin{aligned}
& var\{\psi(\mathbf{X})\} \\
= \; & var\{\psi(\mathbf{X}) - E\psi(\mathbf{X})\} \\
= \; & var\{E[\psi(\mathbf{X}) - E\psi(\mathbf{X}) \mid \phi(\mathbf{X})]\} + E\{var[\psi(\mathbf{X}) - E\psi(\mathbf{X}) \mid \phi(\mathbf{X})]\} \\
\geq \; & var\{E[\psi(\mathbf{X}) - E\psi(\mathbf{X}) | \phi(\mathbf{X})]\}
\end{aligned}
$$

and

$$
E\{\rho_\tau(Y - (\psi(\mathbf{X}) - E\psi(\mathbf{X})))\} = E[E\{\rho_\tau(Y - (\psi(\mathbf{X}) - E\psi(\mathbf{X})))\} | Y, \phi(\mathbf{X})]
$$

$$\geq E[\rho_\tau(Y - E\{\psi(\mathbf{X}) - E\psi(\mathbf{X})|\phi(\mathbf{X})\})].$$

Therefore, for any $\alpha \in R$,

$$\Lambda_\tau(\alpha, \psi) \geq \Lambda_\tau(\alpha, \mathcal{L}(\psi)), \tag{A.6}$$

where $\mathcal{L}(\psi) = E[\psi(\mathbf{X}) - E\psi(\mathbf{X})|\phi(\mathbf{X})]$. If there is a version of ψ measurable with respect to $\sigma\{\phi(\mathbf{X})\}$(i.e. unbiased), then

$$var\{E[\psi(\mathbf{X}) - E\psi(\mathbf{X}) \mid \phi(\mathbf{X})]\} = 0$$

and

$$E[\psi(\mathbf{X}) - E\psi(\mathbf{X}) \mid \phi(\mathbf{X})] = \psi(\mathbf{X}) - E\{\psi(\mathbf{X})\}$$

and the equality in (A.6) holds. Suppose a function $\tilde{\psi}$ has no version measurable with respect to $\sigma\{\phi(\mathbf{X})\}$, then $\Lambda_\tau(\alpha, \tilde{\psi}) > \Lambda_\tau(\alpha, \mathcal{L}(\psi))$. Notice that $\mathcal{H} \subseteq L_2(P_X)$ and hence $\mathcal{L}(\psi) \in L_2(P_X)$ for all $\psi \in \mathcal{H}$. Thus, for any given $\alpha \in R$, we can choose $\psi' \in \mathcal{H}$ such that

$$\Lambda_\tau(\alpha, \tilde{\psi}) > \Lambda_\tau(\alpha, \psi') > \Lambda_\tau(\alpha, \mathcal{L}(\psi))$$

since the map $\psi \to \Lambda_\tau(\alpha, \psi)$ is continuous in ψ with respect to $L_2(P_X)$-norm. Therefore, $\tilde{\psi}$ cannot be the minimizer. ■

Proof of Theorem 6

Define $\xi_i = [Y_i - \alpha - \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i]_+$ and $\xi_i^* = [\alpha + \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i - Y_i]_+$. The sample objective function (3.4) can be rewritten in an equivalent way as

$$\boldsymbol{\gamma}^\top \boldsymbol{\Omega}^\top \boldsymbol{\Omega} \boldsymbol{\gamma} + \lambda \sum_{i=1}^{n} \{\tau \xi_i + (1-\tau)\xi_i^*\}$$

subject to

$$-\xi_i^* \leq Y_i - f(\mathbf{X}_i) \leq \xi_i$$

and

$$\xi_i, \xi_i^* \geq 0, \quad i = 1, ..., n,$$

where

$$f(\mathbf{X}_i; \alpha, \psi) = \alpha + \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i \quad i = 1, ..., n.$$

Then the foregoing setting gives the Lagrangian primal function,

$$\begin{aligned} & L_p(\alpha, \gamma, \xi_i, \cdots, \xi_n, \xi_i^*, \cdots, \xi_i^*) \\ = \; & \boldsymbol{\gamma}^\top \boldsymbol{\Omega}^\top \boldsymbol{\Omega} \boldsymbol{\gamma} + \lambda \sum_{i=1}^{n} \{\tau \xi_i + (1-\tau)\xi_i^*\} + \sum_{i=1}^{n} v_i (Y_i - \alpha - \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i - \xi_i) \\ & - \sum_{i=1}^{n} \eta_i (Y_i - \alpha - \boldsymbol{\gamma}^\top \boldsymbol{\Omega}_i + \xi_i^*) - \sum_{i=1}^{n} k_i \xi_i - \sum_{i=1}^{n} \rho_i \xi_i^*, \end{aligned}$$

where v_i, η_i, k_i and ρ_i are nonnegative Lagrange multipliers. Setting the derivatives of L_p to 0, we arrive at

$$\begin{aligned}
\frac{\partial}{\partial \gamma}: &\quad \gamma = \frac{1}{2}\sum_{i=1}^{n}(v_i - \eta_i)(\mathbf{\Omega}^\top \mathbf{\Omega})^{-1}\mathbf{\Omega}_i, \\
\frac{\partial}{\partial \alpha}: &\quad \sum_{i=1}^{n} v_i = \sum_{i=1}^{n} \eta_i, \\
\frac{\partial}{\partial \xi_i}: &\quad v_i = \lambda\tau - k_i, \\
\frac{\partial}{\partial \xi_i^*}: &\quad \eta_i = \lambda(1-\tau) - \rho_i.
\end{aligned}$$

Because the Lagrange multiplier must be nonnegative, we can conclude that both $0 \le v_i \le \lambda\tau$ and $0 \le \eta_i \le \lambda(1-\tau)$. Plugging the constraints into L_p, the corresponding dual problem is given by

$$\max_{v_1,\ldots,v_n,\eta_1,\ldots,\eta_n} \sum_{i=1}^{n}(v_i - \eta_i)Y_i - \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}(v_i - \eta_i)(v_j - \eta_j)\mathbf{\Omega}_i^\top(\mathbf{\Omega}^\top\mathbf{\Omega})^{-1}\mathbf{\Omega}_j.$$

Note that $\mathbf{\Omega}_i^\top(\mathbf{\Omega}^\top\mathbf{\Omega})^{-1}\mathbf{\Omega}_j$ is the same as $P_{\mathbf{\Omega}}^{\{i,j\}}$.

Acknowledgments

We thank two reviewers, an associate editor, and the editor for their most helpful comments. Shin is partially supported by National Research Foundation of Korea (NRF) grant No. 2015R1C1A1A01054913. Wu is partiallly supported by National Science Foundation grants DMS-1055210 and DMS-1812354.

References

[1] Bura, E. and Pfeiffer, C. A. (2008). On the distribution of the left singular vectors of a random matrix and its applications. *Statistics and Probability Letters* **78** 2275–2280. MR2462662

[2] Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association* **89** 177–189. MR1266295

[3] Cook, R. (1998a). Regression Graphics: Ideas for Studying Regressions Through Graphics. 1998. *USA: A Wiley-Interscience Pubilcation CrossRef Google Scholar*. MR1645673

[4] Cook, R. D. (1998b). Principal Hessian directions revisited. *Journal of the American Statistical Association* **93** 84–94. MR1614584

[5] Cook, R. D. (2007). Fisher Lecture: Dimension Reduction in Regression. *Statistical Science* **22** 1–26. MR2408655

[6] Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression. *Journal of the American Statistical Association* **100**. MR2160547

[7] Cook, R. D. and Ni, L. (2006). Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika* 65–74. MR2277740

[8] Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86** 28–33. MR1137117

[9] Hall, P. and Li, K. C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *The Annals of Statistics* **21** 867–889. MR1232523

[10] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management* **5** 81–102.

[11] Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics* **2** 97–114. MR1085924

[12] Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications* **33** 82–95. MR0290013

[13] Kong, E. and Xia, Y. (2014). An adaptive composite quantile approach to dimension reduction. *The Annals of Statistics* **42** 1657–1688. MR3262464

[14] Lee, K.-Y., Li, B. and Chiaromonte, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *The Annals of Statistics* **41** 221–249. MR3059416

[15] Li, K. C. (1991a). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86** 316–342. MR1137117

[16] Li, K.-C. (1991b). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86** 316–327. MR1137117

[17] Li, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association* **87** 1025–1039. MR1209564

[18] Li, B., Artemiou, A. and Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics* **39** 3182–3210. MR3012405

[19] Li, B. and Dong, Y. (2009). dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics* **37** 1272–1298. MR2509074

[20] Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics* **17** 1009–1052. MR1015136

[21] Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association* **102** 255–268. MR2293307

[22] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102** 997–1008. MR2354409

[23] Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of statistics* 1580–1616. MR2166556

[24] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In *Hnadbook of Economietrics IV* 2113–2245. MR1315971

[25] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory* **7** 186–199. MR1128411

[26] Schölkopf, B., Smola, A. and Müller, K.-R. (1997). Kernel principal component analysis. In *Artificial Neural Networks–ICANN'97* 583–588. Springer.

[27] Shin, S. J., Zhang, H. H. and Wu, Y. (2017). A nonparametric survival function estimator via censored kernel quantile regressions. *Statistica Sinica* **27** 457-478. MR3618178

[28] Shin, S. J., Wu, Y., Zhang, H. H. and Liu, Y. (2017). Principal Weighted Support Vector Machines for Sufficient Dimension Reduction in Binary Classification. *Biometrika* **104** 67-81. MR3626475

[29] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics* **35** 2769–2794. MR2382665

[30] Takeuchi, I., Nomura, K. and Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation* **21** 533–559. MR2477869

[31] Takeuchi, I., Le, Q. V., Sears, T. D. and Somolar, A. J. (2006). Nonparametric quantile estimation. *Journal of machine learning research* **7** 1231–1264. MR2274404

[32] Van der Vaart, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press. MR1652247

[33] Vapnik, V. (1996). *The Nature of Statistical Learning Theory*. Cambridge University Press: New York. MR1719582

[34] Wu, H. M. (2008). Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics* **17** 590–610. MR2528238

[35] Wu, Q., Liang, F. and Mukherjee, S. (2013). Kernel sliced inverse regression: regularization and consistency. *Abstract and Applied Analysis* **2013** 1–11. Article ID 540725. MR3081598

[36] Xia, Y., Tong, H., Li, W. and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64** 363–410. MR1924297

[37] Yeh, Y. R., Huang, S. Y. and Lee, Y. Y. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Transactions on Knowledge and Data Engineering* **21** 1590–1603.

[38] Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis* **99** 1733–1757. MR2444817