ESAIM: PS 20 (2016) 1–17
DOI: 10.1051/ps/2015018

ESAIM: Probability and Statistics
www.esaim-ps.org

PARTIALLY LINEAR ESTIMATION USING SUFFICIENT DIMENSION REDUCTION

Takuma Yoshida[1]

Abstract. In this paper, we study estimation for partial linear models. We assume radial basis functions for the nonparametric component of these models. To obtain the estimated curve with fitness and smoothness of the nonparametric component, we first apply the sufficient dimension reduction method to the radial basis functions. Then, the coefficients of the transformed radial basis functions are estimated. Finally, the coefficients in the parametric component can be estimated. The above procedure is iterated and hence the proposed method is based on an alternating estimation. The proposed method is highly versatile and is applicable not only to mean regression but also quantile regression and general robust regression. The $\sqrt{n}$-consistency and asymptotic normality of the estimator are derived. A simulation study is performed and an application to a real dataset is illustrated.

Mathematics Subject Classification. 62F12, 62J02, 62G07.

Received April 21, 2015. Revised November 27, 2015.

1. Introduction

Partial linear models generalize multiple linear regression and nonparametric regression, and combine their two regressions. That is, the partial linear model has the following structure

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + g(\boldsymbol{Z}) + \varepsilon, \tag{1.1}$$

where Y is the scalar response, $\boldsymbol{X}, \boldsymbol{Z}$ are p and q dimensional predictors, $\boldsymbol{\beta}$ is an unknown parametric vector, g is an unknown smooth function and ε is the error. Many authors have developed partial linear regression, including Heckman [18], Chen [3], Bhattacharya and Zhao [1] and Xia and Härdle [37]. Härdle [14] has written recent book on partial linear models. The partial linear model is an efficient technique not only for mean regression but also for quantile regression. Quantile regression estimates the conditional $100\tau\%$ points of Y given predictors and was suggested by Koenker and Bassett [23]. He and Shi [17], Lee [24] and Sun [32] studied partial linear models with quantile regression. A typical problem of partial linear regression is that nonparametric components are subject to the curse of dimensionality and hence their estimates become complicated. To avoid this, an additive structure in the nonparametric component is often considered [16]. It is known that the additive model is a very simple and useful model. However it is based on a somewhat unsuitable assumption that there is no interaction

Keywords and phrases. Partial linear model, robust regression, sliced average variance estimation, sliced inverse regression, sufficient dimension reduction.

[1] Graduate School of Science and Engineering, Kagoshima University, Kagoshima 890-8580, Japan.
yoshida@sci.kagoshima-u.ac.jp

Article published by EDP Sciences

between components of the predictor vector. This has motivated us to investigate the surface smoothing problem in the nonparametric component of partial linear models.

In this paper, we consider (1.1), investigate the conditional mean and quantile of Y given predictors, and propose a smooth estimation of g. When $q \geq 2$, the computational cost of the estimation of g grows and estimation is unstable if nonparametric smoothing methods such as the kernel or spline method, are used. This is particularly prominent in quantile regression. Therefore to simplify the estimation of g, we approximate g using the radial basis function models defined as

$$S(\boldsymbol{z}) = \sum_{k=1}^{K} c_k \phi_k(\boldsymbol{z}) = \boldsymbol{c}^T \boldsymbol{\phi}(\boldsymbol{z}),$$

where $\boldsymbol{\phi}(\boldsymbol{z}) = (\phi_1(\boldsymbol{z}), \ldots, \phi_K(\boldsymbol{z}))^T$ is a known radial basis function vector and $\boldsymbol{c} = (c_1, \ldots, c_K)^T$ is an unknown parameter vector. We estimate $c_1, \ldots, c_K$ instead of estimating g directly. Although (1.1) is regarded as a fully parametric model when using the radial basis function models, the least squares estimation leads a wiggly curve being obtained. Therefore we should improve the estimation of g to have both fitness and smoothness. Wood [36] proposed thin plate regression splines, which is the penalized least squares method with an integral of the square of the mth derivative of $S(\boldsymbol{z})$. Although his method is useful, it can only be applied to mean regression. We want to construct an efficient estimator that works even with for quantile regression. In general, we consider M-estimation including mean, quantile and general robust regression. We propose a new estimation method using the sufficient dimension reduction (SDR) method.

Before describing the proposed method, we now briefly introduce the SDR method. For the response $Y \in \mathbb{R}$ and the predictor $\boldsymbol{Z} \in \mathbb{R}^q$, consider the matrix $B \in \mathbb{R}^{q \times d}(d < q)$ satisfying

$$Y \perp\!\!\!\perp \boldsymbol{Z} \mid B^T \boldsymbol{Z}. \tag{1.2}$$

In other words, the conditional distribution of Y given $\boldsymbol{Z}$ is similar to that of Y given $B^T\boldsymbol{Z}$. Thus, the dimension of the predictor is reduced from q to d. The matrix B satisfying (1.2) and its common space are called a dimension reduction subspace. There are a variety of SDR methods in the literature, including sliced inverse regression [26], sliced average variance estimation [6], minimum average variance estimation [38], directional regression [27] and related methods by several other authors. The SDR method for partial linear models has also been developed by Chiaromonte *et al.* [5], Wang *et al.* [35] and Feng *et al.* [12]. The partial dimension reduction subspace is defined as the matrix B such that

$$Y \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}, B^T \boldsymbol{Z} \tag{1.3}$$

and its common spaces. Roughly speaking, the partial linear model combined with the SDR method is given as

$$Y = \boldsymbol{X}^T \boldsymbol{\beta} + g(B^T \boldsymbol{Z}) + \varepsilon.$$

When the number of rows of B is $d = 1$, the model can be regarded as a partial linear single index model. For this case, Carroll *et al.* [2], Yu and Ruppert [39], Xia and Härdle [37] and Ding *et al.* [8] studied the estimation methods and properties of mean regression. Wang *et al.* [35] proposed the SDR method for partial linear single index models. However to find B satisfying (1.3) for $d = 1$ or greater, the nonparametric regression with response $\boldsymbol{X}$ and predictor $\boldsymbol{Z}$ is needed. Therefore when the dimension of $\boldsymbol{X}$ is larger than 2, the estimation becomes complicated. This motivates us to apply the SDR method (1.2) (but not (1.3)) to the nonlinear component of the partial linear models.

Our idea is simple and our proposed method is based on alternating estimation. If $\boldsymbol{\beta}$ is know in (1.1), then the problem is reduced to the nonlinear regression between $Y - \boldsymbol{X}^T\boldsymbol{\beta}$ and $g(\boldsymbol{Z})$. Together with radial basis functions, we consider the matrix $\Theta \in \mathbb{R}^{K \times d}(d < K)$ satisfying

$$Y - \boldsymbol{X}^T \boldsymbol{\beta} \perp\!\!\!\perp \boldsymbol{\phi}(\boldsymbol{Z}) \;\mid\; \Theta^T \boldsymbol{\phi}(\boldsymbol{Z}).$$

Using this Θ, (1.1) is modified to

$$Y - \boldsymbol{X}^T\boldsymbol{\beta} = h(\Theta^T\boldsymbol{\phi}(\boldsymbol{Z})) + \varepsilon,$$

where h is redefined as the function from $\mathbb{R}^d$ to $\mathbb{R}$. Since $\Theta^T\boldsymbol{\phi}(\boldsymbol{Z})$ is the linear transformation of the nonlinear predictor, we expect to estimate g well even if h has a simple form. Actually, in this paper, h is assumed to be the linear model $h(x_1,\dots,x_d) = w_1x_1 + \dots + w_dx_d$, where $w_1,\dots,w_d$ are unknown parameters. In practice, the pilot estimator of $\boldsymbol{\beta}$ is needed to estimate Θ. Therefore by first using an initial $\boldsymbol{\beta}_0$, we estimate Θ and g. Then, $\boldsymbol{\beta}$ is estimated *via* multiple linear regression with $Y - \hat{g}$ versus $\boldsymbol{X}$, where $\hat{g}$ is the estimator of g obtained in the previous step. Thus, the alternating estimation is proceed. To find and estimate Θ, an existing method can be used. In our paper, a comparison with the SDR method is not discussed and we mainly use the sliced inverse regression (SIR). The proposed method can be applied to mean regression, quantile regression and general robust regression. Futher, it can also be applied it to composite quantile regression [42]). A composite quantile regression (CQR) assumes that there exist common covariate effects in a range of quantiles such that the quantile levels only differ in terms of the intercept. Zou and Yuan [42] showed that the relative efficiency of the CQR is 70% greater than that of the least squares estimators. Furthermore CQR performs well even if the error distribution is non-normal. These efficient properties of the CQR motivate us to develop the proposed method for CQR.

The remainder of this paper is organized as follows. In Section 2, we elaborate on the new estimation method for partial linear models. We then use the general convex loss function and hence construct the estimator based on the literature of robust regression literature. The estimation is based on alternation. Section 3 investigates the $\sqrt{n}$-consistency and asymptotic normality of the estimator. Section 4 describes an example of the regression problem to which the proposed method is applicable. In Section 5, we report the results of a simulation study and provide an application to real data. Discussion and future study are given in Section 6. Proofs of the main theorem of this paper are in the Appendix.

2. Estimation

We suppose that an independent and identically distributed (i.i.d.) random sample $\{(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) : i = 1,\dots,n\}$ of $(Y, \boldsymbol{X}, \boldsymbol{Z})$ is modeled by a partial linear model

$$Y_i = \mu + \boldsymbol{X}_i^T\boldsymbol{\beta} + g(\boldsymbol{Z}_i) + \varepsilon_i, \tag{2.1}$$

where Y is an one dimensional response, $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Z} \in \mathbb{R}^q$ are predictor, $\boldsymbol{\beta} = (\beta_1,\dots,\beta_p)^T$ is an unknown parameter vector, $g : \mathbb{R}^q \to \mathbb{R}$ is an unknown function and ε_i is the error and is assumed to be independent of $(\boldsymbol{X}, \boldsymbol{Z})$.

Then, we assume the function g can be written by radial basis function models as follows

$$g(\boldsymbol{z}) = g(\boldsymbol{\phi}(\boldsymbol{z})) = \sum_{k=1}^{K} u_k\phi_k(\boldsymbol{z}) = \boldsymbol{u}^T\boldsymbol{\phi}(\boldsymbol{z}), \tag{2.2}$$

where μ is an unknown intercept parameter, $\boldsymbol{u} = (u_1,\dots,u_K)^T \in \mathbb{R}^K$ is an unknown vector, $\boldsymbol{\phi}(\boldsymbol{z}) = (\phi_1(\boldsymbol{z}),\dots,\phi_K(\boldsymbol{z}))^T$, $\phi_k(\boldsymbol{z}) = \phi(||\boldsymbol{z} - \boldsymbol{\kappa}_k||)$ and $\boldsymbol{\kappa}_1,\dots,\boldsymbol{\kappa}_K \in \mathbb{R}^q$ are q-dimensional fixed knots. Here for any vector $\boldsymbol{a}$, $||\boldsymbol{a}|| = \sqrt{\boldsymbol{a}^T\boldsymbol{a}}$. Although several ϕ can be considered, we use the thin plate spline basis, which is of the form

$$\phi_k(\boldsymbol{z}) = \phi(||\boldsymbol{z} - \boldsymbol{\kappa}_k||) = \begin{cases} ||\boldsymbol{z} - \boldsymbol{\kappa}_k||^{2m-q}, & q \text{ odd}, \\ ||\boldsymbol{z} - \boldsymbol{\kappa}_k||^{2m-q}\log||\boldsymbol{z} - \boldsymbol{\kappa}_k||, & q \text{ even}, \end{cases}$$

where m is an integer satisfying $2m - q > 0$ that controls the smoothness of ϕ.

Applying this, (2.1) can be written as

$$Y_i = \mu + \boldsymbol{X}_i^T\boldsymbol{\beta} + \boldsymbol{u}^T\boldsymbol{\phi}(\boldsymbol{Z}_i) + \varepsilon_i.$$

We wish is to estimate $(\boldsymbol{\beta}, \boldsymbol{u})$ in the context of general robust methods. Using the objective function $\rho : \mathbb{R} \to \mathbb{R}$, the ordinary estimator $(\tilde{\mu}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{u}})$ of $(\mu, \boldsymbol{\beta}, \boldsymbol{u})$ is defined as

$$(\tilde{\mu}, \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{u}}) = \underset{\mu, \boldsymbol{\beta}, \boldsymbol{u}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \rho(Y_i - (\mu + \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{u}^T \boldsymbol{\phi}(\boldsymbol{Z}_i))) \right\}. \tag{2.3}$$

However in this method, the estimator $\tilde{g}(\boldsymbol{z}) = \tilde{\boldsymbol{u}}^T \phi(\boldsymbol{z})$ of g is not suitable for prediction since $\tilde{g}$ will have a wiggly curve with large K. Furthermore the controlling the number and location of the knots is a very serious problem. Wood [36] proposed the penalized least squares method, but this method is only useful for mean regression.

To improve this issue, we propose to obtain an estimator of $(\boldsymbol{\beta}, \boldsymbol{u})$ with fitness and smoothness. We apply the SDR method to the nonlinear term g. Let $F(y - \boldsymbol{\beta}^T \boldsymbol{x} | \boldsymbol{Z})$ be the conditional distribution of $Y - \boldsymbol{\beta}^T \boldsymbol{X}$ given $\boldsymbol{Z}$. We then consider the equation

$$F(y - \boldsymbol{\beta}^T \boldsymbol{x} | \boldsymbol{\phi}(\boldsymbol{Z})) = F(y - \boldsymbol{\beta}^T \boldsymbol{x} | \varTheta^T \boldsymbol{\phi}(\boldsymbol{Z})), \tag{2.4}$$

where $\varTheta$ is a $K \times d$ matrix. Suppose there exists $\varTheta$ such that (2.4) is satisfied. Then (2.1) can be written as

$$Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta} = \mu + h(\varTheta^T \boldsymbol{\phi}(\boldsymbol{Z}_i)) + \varepsilon_i,$$

where $h : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$h(\varTheta^T \boldsymbol{\phi}(\boldsymbol{Z})) = h(\boldsymbol{\theta}_1^T \boldsymbol{\phi}(\boldsymbol{Z}), \ldots, \boldsymbol{\theta}_d^T \boldsymbol{\phi}(\boldsymbol{Z})),$$

with $\varTheta = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d]$ and $\boldsymbol{\theta}_j \in \mathbb{R}^K, j = 1, \ldots, d$. Furthermore we assume that

$$h(x_1, \ldots, x_d) = \sum_{j=1}^{d} w_j x_j,$$

where $\boldsymbol{w} = (w_1, \ldots, w_d)^T$ is an unknown parameter vector. We note that the function h is reduced to (2.2) when $d = K$. In other words, the proposed method becomes similar to (2.3). Therefore we generally consider $d \ll K$. We rewrite $\varTheta = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d]$

Let $\varXi = \{\varTheta \in \mathbb{R}^{K \times d} | F(y - \boldsymbol{x}^T \beta | \boldsymbol{\phi}(\boldsymbol{Z})) = F(y - \boldsymbol{x}^T \beta | \varTheta^T \boldsymbol{\phi}(\boldsymbol{Z}))\}$ Then the parameters are estimated *via*

$$\min_{\mu, \boldsymbol{\beta}, \boldsymbol{w}} \sum_{i=1}^{n} \rho \left(Y_i - \mu - \boldsymbol{\beta}^T \boldsymbol{X}_i - \boldsymbol{w}^T \varTheta^T \boldsymbol{\phi}(\boldsymbol{Z}_i) \right) \quad \text{subject to} \quad \varTheta \in \varXi. \tag{2.5}$$

However the optimization of the solution of (2.5) is difficult and hence the estimation is based on mutual iteration. If $\boldsymbol{\beta}$ is known, then the model is

$$Y_i - \boldsymbol{\beta}^T \boldsymbol{X}_i = \mu + \boldsymbol{w}^T \varTheta^T \boldsymbol{\phi}(\boldsymbol{Z}_i) + \varepsilon_i.$$

and we can then estimate $\varTheta$ by the SDR method and $(\mu, \boldsymbol{w})$ by minimizing the loss function ρ. On the other hand, $\boldsymbol{\beta}$ can be estimated *via* the ordinary parametric method when $\varTheta$ and $(\mu, \boldsymbol{w})$ are known. Thus, we estimate the parameters in accordance with the model at each stage. The algorithm is that for a given initial $\boldsymbol{\beta}_{(0)}$ and $t = 1, 2, \ldots$, we iterate

$$\begin{aligned} \varTheta_{(t)} &= \text{Applying the SDR method to } \{(Y_i - \boldsymbol{\beta}_{(t-1)}^T \boldsymbol{X}_i, \boldsymbol{\phi}(\boldsymbol{Z}_i)) : i = 1, \ldots, n\}, \\ (\mu_{(t)}, \boldsymbol{w}_{(t)}) &= \underset{\mu, \boldsymbol{w}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \rho(Y_i - \hat{\mu} - \boldsymbol{\beta}_{(t-1)}^T \boldsymbol{X}_i - \boldsymbol{w}^T (\varTheta_{(t)}^T \boldsymbol{\phi}(\boldsymbol{Z}_i)))^T) \right\}, \\ \boldsymbol{\beta}_{(t)} &= \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \rho(Y_i - \mu_{(t)} - \boldsymbol{\beta}^T \boldsymbol{X}_i - \mu_{(t)} - \boldsymbol{w}_{(t)}^T \varTheta_{(t)}^T \boldsymbol{\phi}(\boldsymbol{Z}_i)) \right\} \end{aligned}$$

until $(\boldsymbol{\beta}_{(t)}, \mu_{(t)}, \boldsymbol{w}_{(t)}, \Theta_{(t)})$ converges. For the SDR method, sliced inverse regression (SIR), sliced average variance estimation (SAVE), minimum average variance estimation (MAVE) and others can be used.

Remark 2.1. We assume that the joint distribution of $(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $(Y_j, \boldsymbol{X}_j, \boldsymbol{Z}_j)$ are independent and same. This i.i.d. setting is needed to use the SDR method in the estimation of the nonlinear component. In the estimation of the parametric component, this i.i.d. assumption would be relaxed. However for this, the additional discussion is needed and this beyond the scope of this paper.

Remark 2.2. In our method, $(\mu, \boldsymbol{\beta}, \boldsymbol{w})$ and Θ are estimated separately. It is known that for such alternating method, there is the uncertainty of the estimator in estimation of Θ is not incorporated in the estimation of $(\mu, \boldsymbol{\beta}, \boldsymbol{w})$ and *vice versa*. From this, it is possible to encounter the problem that the estimation may be bias or the standard error of the estimators of $(\mu, \boldsymbol{\beta}, \boldsymbol{w})$ or Θ may be under-estimated. Therefore additional step to incorporate the estimation uncertainty in the first step is desired but it may be difficult since the relationship between Θ and $\boldsymbol{\beta}$ is nonlinear. On the other hand, we expect that the approximation of bias and variance of the estimator can be calculated using bootstrap method. However in each replication of bootstrap, the dimension d of Θ is different and hence this should be adjusted appropriately. Thus the problem of the estimation uncertainty of the proposed method is important but this is posited as future study in this paper.

Remark 2.3. The intercept μ should be included in the nonlinear component g but not the parametric component. Otherwise, the nonlinear smoothing $Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta}_{(t)}$ versus $\{\Theta_{(t)}\}^T \boldsymbol{\phi}(\boldsymbol{Z}_i)$ is not able to work well.

Remark 2.4. The algorithm is stopped when $||\boldsymbol{\beta}_{(t)} - \boldsymbol{\beta}_{(t-1)}||_\infty < \delta$ for some constant δ, where $|| \cdot ||_\infty$ is the maximum norm. In Section 5, we set $\delta = 10^{-6}$. On the other hand, it is sufficient that the iteration number of the estimation is low when a fixed d and a good initial estimate $\boldsymbol{\beta}_{(0)}$ are used. In fact, $\hat{\boldsymbol{\beta}}$ in (2.3) works reasonably well, in contrast to the nonlinear component. In the numerical study in Section 5, we see that a low iteration number leads to good performance although it is not further discussed in this paper since the discussion of the optimal iteration time is beyond its scope.

Remark 2.5. If a different d is used in each iteration, one should determine that the stopping rule of the iteration depends only on $\boldsymbol{\beta}$. In this case, the convergence of $\boldsymbol{w}$ may be pointless since the dimension of $\boldsymbol{w}$ is different in each iteration. Incidentally, several techniques to determine d have been proposed. In particular, Zhu *et al.* [41] proposed an elegant method using a so-called Bayesian information criterion (BIC). Their method is also useful in our models (see Sect. 5).

Remark 2.6. One of related technique to our method is that the single index models (SIM). In SIM, the linear transformation of the predictor vector is considered. Then the dimension of the predictor is reduced to one dimension and the nonparametric smoothing is applied to the scalar response and the transformed predictor. On the other hand, the proposed method considers the linear transformation of the nonlinear function of the predictor. Next the multiple linear regression method is applied. The disadvantage of SIM is that the transformed predictor is always one dimension and hence the loss of information may be occurred. In our method, there is no such restriction. Not that the second step estimation becomes complicated since we use the multiple linear regression. Thus, our method is easy to use. In Section 5, we compare the proposed method with SIM and other existing methods.

3. Theoretical result

In this section, theoretical properties of the proposed method are studied. In particular, the $\sqrt{n}-$*consistency* and asymptotic normality of $(\boldsymbol{\beta}_{(t)}, \mu_{(t)}, \boldsymbol{w}_{(t)}, \Theta_{(t)})$ are derived. The estimators $\boldsymbol{\beta}_{(t)}$ and $\mu_{(t)}, \boldsymbol{w}_{(t)}$ are obtained by minimizing the loss function. However, the estimating system of $\Theta_{(t)}$ is different from minimizing ρ. First, we consider the consistency of $\Theta_{(t)}$ although it has already been shown to be consistent for SIR (see, [26]). Next the asymptotic normality of $(\boldsymbol{\beta}_{(t)}, \mu_{(t)}, \boldsymbol{w}_{(t)})$ is derived.

For simplicity, for any function U, we write $E[U(Y,\boldsymbol{x},\boldsymbol{z})] \equiv E[U(Y,\boldsymbol{X},\boldsymbol{Z})|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=\boldsymbol{z}]$. Furthermore when U is a vector, $\mathrm{Cov}[U(Y,\boldsymbol{x},\boldsymbol{z})]$ is the covariance matrix of U. We define the true parameters $\boldsymbol{\beta},\mu,\boldsymbol{w},\Theta$. Note that Θ is concerned with SIR. We define the dimension reduction subspace $\boldsymbol{\Theta} = \{\Theta \in \mathbb{R}^{K\times d} | F(y-\boldsymbol{x}^T\boldsymbol{\beta}_0|\boldsymbol{\phi}(\boldsymbol{Z})) = F(y-\boldsymbol{x}^T\boldsymbol{\beta}_0|\Theta^T\boldsymbol{\phi}(\boldsymbol{Z}))\}$. To use SIR, let $\Lambda = \mathrm{Cov}[\boldsymbol{\phi}(\boldsymbol{Z}) - E[\boldsymbol{\phi}(\boldsymbol{Z})]|Y-\boldsymbol{X}^T\boldsymbol{\beta}_0]$ and $G = \mathrm{Cov}[\boldsymbol{\phi}(\boldsymbol{Z})]$. Then we consider the eigen equation

$$\Lambda\boldsymbol{\theta}_j = \lambda_j G\boldsymbol{\theta}_j, \quad j=1,\dots,K,$$

where $\lambda_1 \geq \dots \geq \lambda_K \geq 0$ are eigenvalues and $\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_K$ are corresponding eigenvectors. We then define $\Theta = [\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_d]$ for $d \leq K$. The parameter $\boldsymbol{\beta}_0$ is not specified here although a brief discussion is presented in Remark 8. Using the above Θ, the true $(\boldsymbol{\beta},\mu,\boldsymbol{w})$ is defined as the minimizer of

$$E[\rho(Y-\mu-\boldsymbol{x}^T\boldsymbol{\beta}-\boldsymbol{w}^T\Theta^T\boldsymbol{\phi}(\boldsymbol{z}))]. \tag{3.1}$$

For the asymptotic consistency of $\Theta_{(t)}$, the following assumptions are needed.

Assumptions A.

A1. For any $\boldsymbol{c} \in \mathbb{R}^K$, $E[\boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{Z})|\Theta^T\boldsymbol{\phi}(\boldsymbol{Z})]$ is linear in $\{\boldsymbol{\theta}_1^T\boldsymbol{\phi}(\boldsymbol{Z}),\dots,\boldsymbol{\theta}_d^T\boldsymbol{\phi}(\boldsymbol{Z})\}$.
A2. For any $s \in \mathbb{R}$, the function $m(s) = E[\boldsymbol{\phi}(\boldsymbol{Z})|Y-\boldsymbol{X}^T\boldsymbol{\beta}_0 = s]$ has continuous first order differentials.
A3. The matrix $G = E[\{\boldsymbol{\phi}(\boldsymbol{Z}) - E[\boldsymbol{\phi}(\boldsymbol{Z})]\}^2]$ is positive definite.

Assumption A1 is important because it requires the vectors $\{\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_d\}$ to be contained in the dimension reduction subspace. Assumptions A2 and A3 are needed to construct the $\sqrt{n}$-consistent estimator of $\Theta \in \boldsymbol{\Theta}$.

Assumption B. For t-steps of iteration, $\boldsymbol{\beta}_{(t-1)} = \boldsymbol{\beta}_0 + O_P(n^{-1/2})$.

Theorem 3.1. *Under the assumptions* A *and* B*, for* $n \to \infty$,

$$\Theta_{(t)} = \Theta + O_P(n^{-1/2})$$

and $\Theta^{(t)}$ *converges to an dimension reduction subspace at order* $\sqrt{n}$.

Using Theorem 3.1, we derive the asymptotics for $\boldsymbol{w}_{(t)}$ and $\boldsymbol{\beta}_{(t)}$. Then we give some additional assumptions for the loss function ρ. To simplify, we define $r(\boldsymbol{x},\boldsymbol{z}) = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{w}^T\Theta\boldsymbol{\phi}(\boldsymbol{z})$. Let $\Psi(s) = E[\rho(Y-r(\boldsymbol{x},\boldsymbol{z})-s)]$, $\Psi'(s) = d\Psi(s)/ds$ and $\Psi''(s) = d^2\Psi(s)/ds^2$.

Assumptions C.

C1. $\Psi(0|X)$ is not 0.
C2. $\Psi(s)$, $\Psi'(s)$ and $\Psi''(s)$ as functions of s are bounded and continuous for the neighborhood of 0.
C3. There exists $\gamma > 0$ such that for any $\boldsymbol{x}$ and $\boldsymbol{z}$,

$$E[|\rho'(Y-r(\boldsymbol{x},\boldsymbol{z}))|^{2+\gamma}] < \infty.$$

C4.

$$\lim_{s\to 0}\frac{1}{s^2}E[\{\rho\left(Y-r(\boldsymbol{x},\boldsymbol{z})-s\right) - \rho\left(Y-r(\boldsymbol{x},\boldsymbol{z})\right) - \rho'\left(Y-r(\boldsymbol{x},\boldsymbol{z})\right)s\}^2] = 0.$$

Assumptions C are important for the asymptotic theory of robust regression (see, [11, 13]) To show the asymptotic normality of $\boldsymbol{w}_{(t)}$ and $\boldsymbol{\beta}_{(t)}$, we present some new notation. For the function ν of Y, let $\Sigma_x(\nu) = \mathrm{Cov}[\sqrt{E[\nu(Y)|\boldsymbol{X},\boldsymbol{Z}]}\boldsymbol{X}]$ and let $\Sigma_\phi(\nu)$ be the covariance matrix of $\sqrt{E[\nu(Y)|\boldsymbol{X},\boldsymbol{Z}]}[1\ \{\Theta\boldsymbol{\phi}(\boldsymbol{Z})\}^T]^T$.

Theorem 3.2. *Under the Assumptions* A, B *and* C*, as* $n \to \infty$,

$$\sqrt{n}\begin{bmatrix}\mu_{(t)}-\mu \\ \boldsymbol{w}_{(t)}-\boldsymbol{w}\end{bmatrix} \xrightarrow{D} N(\boldsymbol{0}, \Sigma_\phi(\rho'')^{-1}\Sigma_\phi(\{\rho'\}^2)\Sigma_\phi(\rho'')^{-1}),$$

$$\sqrt{n}(\boldsymbol{\beta}_{(t)}-\boldsymbol{\beta}) \xrightarrow{D} N(\boldsymbol{0}, \Sigma_x(\rho'')^{-1}\Sigma_x(\{\rho'\}^2)\Sigma_x(\rho'')^{-1}).$$

We describe the additional discussion about the proposed estimator. From SDR method, the regression model can be expressed as

$$E[Y-\boldsymbol{\beta}^T\boldsymbol{X}|\Theta^T\boldsymbol{\phi}(\boldsymbol{Z})]=\mu+h(\Theta^T\boldsymbol{\phi}(\boldsymbol{Z}))=\mu+\boldsymbol{w}^T\Theta^T\boldsymbol{\phi}(\boldsymbol{Z})$$

under Assumption A1 and the following Remark 3.6. If the dimension d equals K, the K-vector $\boldsymbol{w}$ can be replaced to $\boldsymbol{w}=\Theta^T(\Theta\Theta^T)^{-1}\boldsymbol{u}$ for $\boldsymbol{u}\in\mathbb{R}^K$. Thus, the nonlinear regression model is reduced to $h(\Theta^T\boldsymbol{\phi}(\boldsymbol{Z}))=\boldsymbol{u}^T\boldsymbol{\phi}(\boldsymbol{Z})=g(\boldsymbol{Z})$. Therefore the proposed estimator with $d=K$ is similar to the ordinary estimator $\tilde{g}(\boldsymbol{z})$ and this tends to have an overfitting curve. It is generally known that the complicated model becomes more overfitting. To decrease the overfitting of the estimator, it is good to reduce the number of adjustable parameter, i.e, the dimension of $\boldsymbol{w}$. Therefore when $d<K$ is used, the proposed estimator has smooth curve rather than the ordinary estimator. In other words, the overfitting curve has a large variance. Consequently, the proposed method can be regarded as the variance reduction method.

Remark 3.3. In this section, we showed the asymptotic results of the estimator under the assumption that the true nonparametric function $g(\boldsymbol{z})$ is equivalent to the radial basis function basis model $S(\boldsymbol{z})$. I other words, we assume that the model bias between $g(\boldsymbol{z})$ and $S(\boldsymbol{z})$ is 0. It seems that deriving the model bias is unsolved problem and is challenging. Thus, it is posited as future work.

Remark 3.4. The definition of $\boldsymbol{\beta}_0$ is important for obtaining Theorem 3.1. In practice, Assumption B is natural although it may seems very strong. For the initial estimates $\boldsymbol{\beta}_{(0)}$, we can use (2.3). Although Assumptions C are necessary, $\boldsymbol{\beta}_{(0)}$ then converges to $\boldsymbol{\beta}_0$, which is the minimizer of $E[\rho(Y-\boldsymbol{x}^T\boldsymbol{\beta}-\boldsymbol{u}^T\boldsymbol{\phi}(\boldsymbol{z}))]$. In other words, the initial estimator $\boldsymbol{\beta}_{(0)}$ is already consistent estimator of $\boldsymbol{\beta}_0$. For this $\boldsymbol{\beta}_0$, Theorems 3.1 and 3.2 are satisfied. Next for $t=1,2,\ldots$, $\boldsymbol{\beta}_0$ is defined as the minimizer of (3.1). Thus, in each iteration, $\boldsymbol{\beta}_0$ can be defined so that it satisfies Assumption B.

Remark 3.5. We use SIR as the SDR method in this paper. On the other hand, we could have used SAVE, MAVE or other SDR method. However SAVE does not have $\sqrt{n}$-consistency unless there is the additional assumption of the so-called constant variance condition (see [29]) and a bias correction is performed. When $d=1$, the MAVE has $\sqrt{n}$-consistency. However when $d>1$, there is no result related to the MAVE. Particularly for $d\geq 3$, MAVE needs further study. Although the principle Hessian direction [26] has $\sqrt{n}$-consistency, it is useful only for mean regression and is not suitable for quantile regression. Thus from the viewpoint of $\sqrt{n}$-consistent estimates and utilizing quantile regression, we use SIR.

Remark 3.6. Assumption A1 is needed in order to SAVE is included in the dimension reduction subspace. In real data analysis, it is very difficult to decide whether $\boldsymbol{\phi}(\boldsymbol{Z})$ satisfies Assumption A1. However, if the distribution of $\boldsymbol{\phi}(\boldsymbol{Z})$ is elliptically symmetric, then this condition is guaranteed (see, [7]). Furthermore Diaconis and Freedman [7] showed that all low-dimensional projections of high-dimensional data are approximately normal. In our setting, we should choose a large K to capture the smooth function g. In fact we set $K=81$ in the simulation study in Section 5. In this sense, it appears that $\boldsymbol{\phi}(\boldsymbol{Z})$ satisfies Assumption A1 naturally.

4. Example

The proposed method depends on the functional form of the loss function ρ. We describe some examples in this section.

4.1. Mean regression

When $\rho(u)=u^2$ is used, our purpose is to obtain the estimator of the conditional mean function of Y given $(\boldsymbol{X},\boldsymbol{Z})$. Then we assume that ε in (2.1) has mean 0 and variance $\sigma^2<\infty$. In this case, the exact form of the estimator $\boldsymbol{\beta}_{(t)}$ and $\boldsymbol{w}_{(t)}$ can be obtained and can be expressed as

$$(\mu_{(t)},\boldsymbol{w}_{(t)}^T)^T=(\Theta_{(t)}^T\Phi^T(I-P_x)\Phi\Theta_{(t)})^{-1}\Theta_{(t)}^T\Phi^T(I-P_x)\boldsymbol{y},$$
$$\boldsymbol{\beta}_{(t)}=(X^T(I-P_z(t))X)^{-1}X^T(I-P_z(t))\boldsymbol{y},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $\Phi = (\mathbf{1}_n, \boldsymbol{\phi}(\boldsymbol{z}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{z}_n))^T$, $P_x = X(X^TX)^{-1}X^T$ and $P_z(t) = \Phi\Theta_{(t)}\left(\Theta_{(t)}^T\Phi^T\Phi\Theta_{(t)}\right)^{-1}\Theta_{(t)}^T\Phi^T$.

In Theorem 3.2, we observe that $\Sigma_\phi(\rho'') = \Sigma_\phi(1)$ and $\Sigma_\phi(\{\rho'\}^2) = \sigma^2\Sigma_\phi(1)$, where $\sigma^2 = V[Y|\boldsymbol{X}, \boldsymbol{Z}]$. Thus, the asymptotic variance of $[\mu_{(t)}, \boldsymbol{w}_{(t)}^T]^T$ is reduced to $\sigma^2\Sigma_\phi(1)^{-1}$. Similarly, the covariance matrix of $\boldsymbol{\beta}_{(t)}$ becomes $\sigma^2\Sigma_\phi(1)^{-1}$.

4.2. Quantile regression

Quantile regression estimates the conditional $100\tau\%$ quantile of the conditional Y given $\boldsymbol{X}$ and $\boldsymbol{Z}$:

$$\mu(\tau) + \boldsymbol{\beta}(\tau)^T\boldsymbol{x} + \boldsymbol{w}(\tau)^T\Theta^T\boldsymbol{\phi}(\boldsymbol{z}), \quad \tau \in (0, 1).$$

To estimate $\mu(\tau), \boldsymbol{\beta}(\tau), \boldsymbol{w}(\tau)$, we use $\rho(u) = \rho_\tau(u) = (\tau - I(u < 0))u$, which is a so called check function (see [22]). Quantile regression can be more efficient than mean regression when the variance of Y is large (or ∞) and an outlier is included in the dataset. Since the exact form of the coefficients can not be written, the minimizer of ρ_τ can be calculated using a linear programming algorithm. Several optimizing methods and related packages have been developed in statistical software R by many authors.

In Theorem 3.2, we obtain $\Sigma_\phi(\rho'') = \Sigma_\phi(1)/f_\varepsilon(b_\tau)$ and $\Sigma_\phi(\{\rho'\}^2) = \tau(1-\tau)\Sigma_\phi(1)$, where f_ε is the density of ε and b_τ is the $100\tau\%$ percentile of ε. Therefore the asymptotic variance of $[\mu_{(t)}, \boldsymbol{w}_{(t)}^T]^T$ can be written as $(\tau(1-\tau)/f_\varepsilon(b_\tau))\Sigma_\phi(1)^{-1}$. By similar arguments, the covariance matrix of $\boldsymbol{\beta}_{(t)}$ is $(\tau(1-\tau)/f_\varepsilon(b_\tau))\Sigma_\phi(1)^{-1}$.

4.3. Composite quantile regression

Composite quantile regression (CQR) as proposed by Zou and Yuan [42] has good asymptotic efficiency properties compared with least squares in mean regression. The benefit of CQR is that the coefficients of the predictor are the same across different quantile levels. The features of the differences between each quantile level are observed from slope. Let $0 < \tau_1 < \ldots < \tau_L < 1$ be $\tau_\ell = \ell/(1+L)(\ell = 1, \ldots, L)$. Then the estimation algorithm for $\boldsymbol{w}$ and $\boldsymbol{\beta}$ are as follows:

$$\begin{aligned}
&(\mu_{1,(t)}, \ldots, \mu_{L,(t)}, \boldsymbol{w}_{(t)}) \\
&= \underset{\mu_1,\ldots,\mu_L,\boldsymbol{w}}{\operatorname{argmin}} \left\{\sum_{\ell=1}^{L}\sum_{i=1}^{n}\rho_{\tau_\ell}(Y_i - \mu_\ell - \boldsymbol{\beta}_{(t-1)}^T\boldsymbol{X}_i - \boldsymbol{w}^T\Theta_{(t)}^T\boldsymbol{\phi}(\boldsymbol{Z}_i))\right\}, \\
&(b_{1,(t)}, \ldots, b_{L,(t)}, \boldsymbol{\beta}) \\
&= \underset{b_1,\ldots,b_L,\boldsymbol{\beta}}{\operatorname{argmin}} \left\{\sum_{\ell=1}^{L}\sum_{i=1}^{n}\rho_{\tau_\ell}(Y_i - b_\ell - \boldsymbol{\beta}^T\boldsymbol{X}_i - \boldsymbol{w}_{(t)}^T\Theta_{(t)}^T\boldsymbol{\phi}(\boldsymbol{Z}_i))\right\},
\end{aligned}$$

where μ_ℓ is the conditional $100\tau_\ell\%$ percentile of $Y - \boldsymbol{x}^T\boldsymbol{\beta}$ given $\boldsymbol{Z}$ and $b_{(\ell)}$ is the conditional $100\tau_\ell\%$ percentile of $Y - \boldsymbol{w}^T\Theta^T\boldsymbol{\phi}(\boldsymbol{z})$ given $\boldsymbol{X}$. The estimation of Θ is not changed. Then we define $\hat{g}(\boldsymbol{z}) = \mu_{(t)} + \boldsymbol{w}_{(t)}^T\Theta_{(t)}^T\boldsymbol{\phi}(\boldsymbol{z})$, where $\mu_{(t)} = (1/L)\sum_{\ell=1}^{L}\mu_{\ell,(t)}$. However this $\mu_{(t)}$ is not suitable when ε does not have symmetric density. In the Monte Carlo simulation of Section 5.1, since we assume that ε has symmetric density, the performance of our method for CQR is evaluated. However in the data example of Section 5.2, this ε's assumption is uncertain. Therefore we will not use CQR in our example and its improvement is beyond the scope of this paper.

In terms of optimization methods, the MM algorithm proposed by Hunter and Lange [19] is empirically efficient. Although this MM algorithm is for quantile regression, it can easily be extended to CQR.

From the proof of Theorem 2.1 of Zou and Yuan [42], we obtain

$$E[\{\rho'(Y - h(\boldsymbol{x}, \boldsymbol{z}))\}^2|\boldsymbol{X}, \boldsymbol{Z}] = \sum_{k,\ell=1}^{K}\min(\tau_k, \tau_\ell)(1 - \max(\tau_k, \tau_\ell))$$

and $E[\rho''(Y - h(\boldsymbol{x}, \boldsymbol{z}))|\boldsymbol{X}, \boldsymbol{Z}] = 1/\sum_{k=1}^{K} f_\varepsilon(\mu_k)$. The asymptotic variance of $[\mu_{(t)}, \boldsymbol{w}_{(t)}^T]^T$ and $\boldsymbol{\beta}_{(t)}$ can be expressed as $B(\tau)\Sigma_\phi(1)^{-1}$ and $B(\tau)\Sigma_x(1)^{-1}$, where $B(\tau) = \sum_{k,\ell=1}^{L} \min(\tau_k, \tau_\ell)(1 - \max(\tau_k, \tau_\ell))/(\sum_{k=1}^{L} f_\varepsilon(\mu_k))^2$.

4.4. M-type robust regression

One robust estimation method is M-type robust regression. In M-type robust regression, we use the Huber loss function

$$\rho_c(u) = \begin{cases} u^2, & |u| \le c, \\ 2c|u| - c^2, & |u| > c. \end{cases},$$

where $c > 0$ is the cutoff constant. If c is too large, then ρ_c approaches $\rho(u) = u^2$. Note that M-type robust regression is similar to the mean regression. On the other hand, when c is very small, M-type robust regression approximately equals the median regression, which is the quantile regression with $\tau = 0.5$. Thus the estimator from the M-type robust regression lies between the mean curve and the median curve. In fact, M-type robust regression is regarded as mean regression mitigated by the influence of outliers. In Section 5, we confirm the performance of the proposed method for mean, quantile and composite quantile regression but not M-type robust regression. However as mentioned above, the performance of M-type robust regression is similar to or between the performances of the mean regression and the median regression. The computation is detailed by Lee and Oh [25].

5. Numerical study

In this section, the dimension of the nonparametric component is fixed as $q = 2$. Hence the surface regression is explored numerically. We use $\phi_k(\boldsymbol{z}) = ||\boldsymbol{z} - \boldsymbol{\kappa}_k||^2 \log ||\boldsymbol{z} - \boldsymbol{\kappa}_k||^2$ as the radial basis function.

5.1. Simulation

We demonstrate the performance of our approach *via* a Monte Carlo simulation. In this simulation, we consider mean regression, quantile regression and CQR. We use $\boldsymbol{z} = (z_1, z_2)^T$ in the nonparametric function g as follows:

$$\begin{aligned} g(\boldsymbol{z}) = & \frac{0.75}{\pi\sigma_1\sigma_2} \exp[-(z_1 - 0.2)^2/\sigma_1^2 - (z_2 - 0.3)^2/\sigma_2^2] \\ & + \frac{0.45}{\pi\sigma_1\sigma_2} \exp[-(z_1 - 0.7)^2/\sigma_1^2 - (z_2 - 0.8)^2/\sigma_2^2], \end{aligned}$$

where $\sigma_1 = 0.3$ and $\sigma_2 = 0.4$. Note that this surface function g was used by Wood [36]. The responses y_i are generated from the regression model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + g(\boldsymbol{z}_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\boldsymbol{\beta} = (1, 2, -3)$ and z_{i1}, z_{i2} are independently generated from the uniform distribution on the interval $[0, 1]$. For the predictor of the parametric components, we consider (i) $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3})^T$ independently generated from $N(\mathbf{0}, \text{diag}[1, 2^2, 0.5^2])$, (ii) $x_{i1} \sim B(1, 1/2)$, $x_{i2} \sim B(1, 1/3)$ and $x_{i3} \sim B(1, 1/4)$, and (iii) $x_{i1} \sim N(0, 1)$, $x_{i2} \sim B(1, 1/2)$ and $x_{i3} \sim B(1, 1/3)$. For the distribution of the error, we consider (i) the standard normal distribution, (ii) the t-distribution with 3 degrees of freedom (df) and (iii) the so-called slash distribution $N(0, 1)/U(0, 1)$. In the mean regression, the predictors (i), (ii), (iii) and the error (i) are simulated. In the quantile regression and the CQR, the predictor (i) and the errors (i), (ii) and (iii) are applied. Two sample sizes $n = 200$ and $n = 1000$, and $R = 1000$ replications were used.

We estimate $\boldsymbol{\beta}$ and g *via* the proposed method. For the number and location of knots included in the thin plate spline function model, 9 equidistant knots were used for the z_1-axis and z_2-axis, leading to a total number

Table 1. MSE of the estimator of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ and the MISE of $\hat{g}(z_1, z_2)$ in each setting for mean regression. All entries for MSE and MISE are 10^2 times their actual values for ease of presentation.

	d	BIC		$d=1$		$d=10$		$d=K$		PLAM		SIM		BKS	
Design	n	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
Normal	β_1	0.601	0.092	0.651	0.107	0.716	0.108	0.801	0.105	0.642	0.121	0.623	0.094	0.652	0.124
	β_2	0.131	0.021	0.134	0.028	0.138	0.027	0.330	0.124	0.151	0.041	0.154	0.052	0.161	0.042
	β_3	0.212	0.038	0.210	0.038	0.243	0.058	0.249	0.177	0.251	0.044	0.213	0.051	0.304	0.062
	$g(\boldsymbol{z})$	4.125	0.837	4.290	0.915	5.028	0.805	9.275	5.371	4.521	0.934	4.631	1.035	4.821	1.241
Binary	β_1	2.475	0.328	2.535	0.368	2.723	0.373	2.795	0.884	2.534	0.372	2.571	0.412	2.715	0.531
	β_2	2.127	0.327	2.337	0.455	2.493	0.458	2.486	1.049	2.342	0.516	2.421	0.503	2.631	0.652
	β_3	2.781	0.327	2.938	0.494	2.996	0.495	3.067	0.754	3.151	0.618	3.212	0.601	3.313	0.526
	$g(\boldsymbol{z})$	34.46	3.845	39.74	4.172	74.86	9.586	74.67	24.12	37.24	4.232	39.41	4.622	43.24	4.631
Hybrid	β_1	0.434	0.083	0.714	0.106	0.736	0.106	0.720	0.124	0.721	0.124	0.731	0.141	0.772	0.184
	β_2	2.359	0.436	3.179	0.448	2.625	0.458	2.466	0.456	3.199	0.447	3.192	0.460	3.214	0.561
	β_3	3.152	0.146	3.423	0.498	3.125	0.496	3.185	0.492	3.455	0.502	3.472	0.523	3.642	0.634
	$g(\boldsymbol{z})$	27.62	2.491	30.28	4.111	53.55	13.59	73.90	51.13	28.18	4.311	30.51	4.032	33.18	4.721

of knots $K = 81$. For the SDR method, we used SIR proposed by Li [26]. The dimension d of the SIR was fixed at $d = 1$, $d = 10$ and $d = K = 81$, and selected *via* BIC. The BIC for SIR was studied by Li and Yin [28]. We note that when $d = K$ is used, the estimator was reduced to that obtained *via* the ordinary method (2.3). The performance of $\hat{\boldsymbol{\beta}}$ and $\hat{g}$ were evaluated using the mean squares error (MSE) and the mean integrated squares error (MISE) with 1000 replications, respectively. The MISE of $\hat{g}(z_1, z_2)$ was estimated by

$$\text{MISE} = \frac{1}{1000}\sum_{r=1}^{1000}\frac{1}{N}\sum_{i=1}^{N}\{\hat{g}_r(z_{i1}^*, z_{i2}^*) - g(z_{i1}^*, z_{i2}^*)\}^2,$$

where $\hat{g}_r$ is the estimator of g for the r-th repetition and $\{(z_{i1}^*, z_{i2}^*) : i = 1, \ldots, N\}$ is the $N = 50 \times 50$ regular grid on the unit square. For comparison, we calculate the MSE and the MISE of the estimators with following three existing approaches: (i) partial linear additive model (see [30]), (ii) single index model (see [20]) and (iii) the bivariate kernel smoothing method (see [40]). In the partial linear additive model (PLAM), we use the cubic B-spline method with 5 equidistant knots to estimate the univariate regression function. For the single index model (SIM) and the bivariate kernel smoothing (BKS), we use the Epanechnikov kernel and the bandwidth selected by the generalized cross-validation.

In Table 1, the simulation result for mean regression is illustrated. In each setting, the performance of $\hat{\boldsymbol{\beta}}$ and $\hat{g}$ improved when the sample size increased. This indicates that the estimator has consistency. In the nonlinear component, the proposed estimators with $d = 1$ performed better rather than those with large d in all settings of the design. Thus the efficiency of the proposed method is observed. On the other hand, we see that the behavior of the estimator in the parametric component is almost similar with regard to the dimension d.

The results for the quantile regression with $\tau = 0.5, 0.75$ and 0.9 are reported in Table 2. First for all τ and error, the asymptotic consistency of the estimator can be observed. Overall, the estimator with small d behaved better than that with the ordinary method ($d = K$). It appears from the results that the BIC is able to select the appropriate d. The performance of the estimator with normal error is better than that with t_3 or the error of the slash distribution although it is obvious. However even if the error is distributed on a slash distribution that has a heavy tail, the MSE and the MISE are not large when $n = 1000$. This indicates that the robustness of the quantile estimation is maintained for the proposed method.

Table 3 reports the performance of the CQR estimates. For quantile points, we use $\tau_\ell = \ell/(1+L)(\ell = 1, \ldots, L)$ with $L = 19$. However the estimations of $\boldsymbol{\beta}$ and $\boldsymbol{w}$ are not sensitive to the number of quantiles L. In all errors, the proposed estimator has good behavior and the MSE of $\hat{\boldsymbol{\beta}}$ is the same as that in quantile regression. In this

Table 2. MSE of the estimator of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ and the MISE of $\hat{g}(z_1, z_2)$ in each setting for quantile regression for $\tau = 0.5, 0.75$ and 0.9. All entries for MSE and MISE are 10^2 times their actual values for ease of presentation.

50% quantile ($\tau = 0.5$)															
	d	BIC		$d = 1$		$d = 10$		$d = K$		PLAM		SIM		BKS	
Error	n	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
Normal	β_1	0.823	0.168	0.888	0.189	0.922	0.184	0.899	0.187	0.912	0.191	0.931	0.202	1.002	0.252
	β_2	0.202	0.043	0.219	0.041	0.201	0.040	0.229	0.040	0.220	0.043	0.231	0.051	0.272	0.082
	β_3	1.426	0.425	1.775	0.144	1.603	0.295	3.908	0.676	3.213	0.315	3.521	0.414	4.266	0.471
	$g(z)$	5.237	0.313	5.948	0.353	8.690	0.743	10.49	7.024	6.012	0.413	6.371	0.512	7.032	0.421
t_3	β_1	1.236	0.202	1.376	0.221	1.457	0.226	1.180	0.803	1.412	0.214	1.468	0.242	1.641	0.622
	β_2	0.221	0.072	0.230	0.056	0.358	0.050	0.248	0.142	0.232	0.055	0.241	0.059	0.251	0.073
	β_3	4.324	0.846	5.064	0.986	5.116	0.913	4.883	0.997	5.021	0.821	5.142	0.933	5.624	1.256
	$g(z)$	12.46	5.237	12.77	4.087	13.33	6.149	43.57	10.48	12.51	4.125	13.25	4.512	14.14	5.882
Slash	β_1	3.462	0.641	3.830	0.720	5.814	0.691	2.951	0.844	3.931	0.731	4.124	0.851	4.316	0.931
	β_2	1.245	0.151	1.347	0.187	1.726	0.207	2.902	2.171	1.521	0.312	1.631	0.212	1.522	0.221
	β_3	5.731	3.042	6.118	3.148	7.896	2.740	16.08	4.012	6.421	3.422	6.755	3.432	7.121	4.243
	$g(z)$	54.26	12.33	52.36	15.73	70.08	22.01	121.7	76.26	51.26	14.92	55.21	16.23	61.26	17.23
75% quantile ($\tau = 0.75$)															
	d	BIC		$d = 1$		$d = 10$		$d = K$		PLAM		SIM		BKS	
Error	n	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
Normal	β_1	0.433	0.125	0.833	0.177	0.985	0.162	1.096	0.194	0.842	0.181	0.862	0.211	0.943	0.234
	β_2	0.215	0.043	0.259	0.045	0.502	0.045	0.891	0.048	0.262	0.051	0.272	0.064	0.291	0.072
	β_3	3.416	0.921	4.689	0.810	4.156	0.845	6.607	2.903	4.839	0.851	4.719	0.872	4.812	0.893
	$g(z)$	11.25	3.315	12.45	3.745	23.67	9.254	33.74	27.64	12.12	3.124	12.65	3.535	13.62	3.826
t_3	β_1	1.426	0.261	1.571	0.339	1.591	0.319	1.821	0.346	1.633	0.461	1.722	0.372	1.928	0.511
	β_2	0.328	0.048	0.418	0.054	0.401	0.048	0.444	0.078	0.472	0.052	0.491	0.062	0.511	0.093
	β_3	7.127	1.128	7.333	1.263	8.073	1.092	7.576	1.319	7.812	1.621	7.925	1.521	7.762	1.427
	$g(z)$	26.35	5.856	22.95	5.864	45.52	9.452	89.84	40.64	23.21	6.114	26.21	6.241	31.25	8.844
Slash	β_1	4.352	1.572	4.847	1.950	11.32	1.762	10.99	1.847	4.731	1.724	4.812	1.911	4.931	2.021
	β_2	1.524	0.225	1.822	0.386	2.356	0.378	2.064	0.329	1.732	0.321	1.841	0.341	2.132	0.501
	β_3	7.272	6.355	9.431	6.019	48.41	5.960	42.50	6.324	8.622	6.429	9.421	6.739	10.33	6.831
	$g(z)$	54.52	9.456	62.72	10.56	71.61	10.29	118.1	42.61	56.22	10.14	59.32	11.01	61.24	12.31
90% quantile ($\tau = 0.9$)															
	d	BIC		$d = 1$		$d = 10$		$d = K$		PLAM		SIM		BKS	
Error	n	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
Normal	β_1	1.124	0.226	1.264	0.254	1.452	0.263	1.978	0.375	1.421	0.323	1.531	0.272	1.835	0.315
	β_2	0.420	0.072	0.410	0.074	0.406	0.070	0.463	0.067	0.461	0.081	0.512	0.092	0.511	0.102
	β_3	7.124	1.124	7.366	1.084	5.308	1.016	6.632	1.310	7.521	1.126	7.941	1.721	8.216	1.831
	$g(z)$	22.15	7.835	22.32	7.365	35.01	9.649	74.63	26.10	22.21	7.215	23.57	7.622	27.47	7.821
t_3	β_1	4.413	0.410	4.270	0.380	3.477	0.415	4.598	1.036	4.312	0.512	4.824	0.413	4.952	0.481
	β_2	0.837	0.126	0.957	0.188	1.110	0.191	1.209	0.198	0.962	0.184	0.972	0.202	1.127	0.341
	β_3	4.372	0.716	6.822	0.956	9.045	3.205	10.28	4.127	6.214	0.834	6.321	0.912	7.102	1.026
	$g(z)$	53.29	7.312	59.09	7.732	74.31	12.32	86.26	47.69	54.21	7.632	59.23	7.763	60.33	8.252
Slash	β_1	26.29	5.137	28.09	5.414	21.66	3.887	30.33	6.855	26.32	5.244	27.21	5.354	28.12	5.834
	β_2	11.24	3.231	12.84	3.302	5.702	2.740	7.048	3.705	12.52	3.235	12.86	3.438	13.23	3.464
	β_3	20.32	4.316	22.72	5.177	24.49	4.097	60.73	8.380	21.24	5.037	22.34	5.312	23.42	5.325
	$g(z)$	112.3	16.34	122.2	18.04	190.5	46.43	547.2	101.04	114.3	17.34	121.4	19.24	129.3	21.24

TABLE 3. MSE of the estimator of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ and the MISE of $\hat{g}(z_1, z_2)$ in each setting for CQR. All entries for MSE and MISE are 10^2 times their actual values for ease of presentation.

	d	BIC		$d = 1$		$d = 10$		$d = K$		PLAM		SIM		BKS	
Error	n	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000	200	1000
Normal	β_1	0.868	0.135	1.128	0.115	1.653	0.071	1.675	7.719	1.012	0.121	1.112	0.141	1.522	0.151
	β_2	0.053	0.014	0.077	0.010	0.219	0.011	0.226	0.998	0.062	0.011	0.071	0.012	0.082	0.041
	β_3	1.035	0.243	1.190	0.233	1.772	0.247	1.327	0.240	1.120	0.242	1.314	0.312	1.621	0.416
	$g(\boldsymbol{z})$	3.156	0.217	4.412	0.368	9.942	0.551	16.56	3.816	4.123	0.371	4.236	0.331	4.824	0.421
t_3	β_1	1.373	0.164	1.653	0.234	2.083	0.314	1.238	0.319	1.427	0.312	1.726	0.371	1.682	0.421
	β_2	0.336	0.013	0.367	0.068	0.849	0.056	0.516	0.066	0.363	0.062	0.366	0.072	0.384	0.082
	β_3	2.924	0.525	2.854	0.604	3.782	0.795	4.001	0.762	3.234	0.621	3.123	0.671	3.314	0.701
	$g(\boldsymbol{z})$	5.626	0.669	6.746	0.762	12.54	1.424	33.46	13.39	6.126	0.741	6.464	0.785	7.216	0.837
Slash	β_1	23.42	5.813	20.22	5.749	17.78	6.854	4.425	7.101	20.12	5.712	22.16	5.813	24.31	5.931
	β_2	1.328	0.324	1.994	0.391	2.911	0.463	10.07	0.583	2.013	0.413	2.192	0.492	2.424	0.512
	β_3	14.27	3.451	15.87	3.607	16.62	3.754	82.49	4.546	15.23	3.712	16.25	3.712	16.81	3.821
	$g(\boldsymbol{z})$	42.46	13.26	58.76	20.20	97.67	28.26	165.0	81.52	48.72	14.53	53.21	19.34	59.32	20.64

case, we also confirmed the consistency of the estimator. Overall, the performance of CQR is better than that of mean regression. Thus, we confirmed the efficiency of CQR as Zou and Yuan [42] had reported.

For the mean regression,the quantile regression and the CQR case, we see that the proposed estimator with d selected by BIC have good performance compared with PLAM, SIM and BKS. In the PLAM, the interaction structure of the function of covariate can not be captured. SIM is useful method to dimension reduction in practice. However in generally, the loss of information may be occurred since q-dimensional predictor is compulsory transformed to 1 dimension. The bivariate kernel method is traditional nonparametric smoothing technique but the computational cost of the bandwidth selection grows when $q \geq 2$. Thus, the proposed method covers the disadvantages of the above methods. Although we can not compare the proposed method with the above and other existing nonparametric/nonlinear methods directly, we could found that the proposed method is one of efficient methods in some settings.

When a BIC is used to select d, the MISE of $\hat{g}$ improved in comparison with the fixed d. Thus the selection of d is important. On the other hand the performance with a BIC is quite similar to that with $d = 1$. Thus, it appears that BIC selected a small d, indicating that the SDR method controls the fitness and smoothness.

The mean regression with error t_3 and a slash distribution are not effective. Thus the simulations in these case have not been performed. Quantile regression and CQR with designs (ii) and (iii) were however explored. Although the results are not reported owing to lack of space, the estimators behaved well.

5.2. Data example: Boston housing data

We now apply the proposed method to Boston housing data, which was originally analyzed by Harrison and Rubinfeld [15]. The data consists of 14 variables (including a binary feature) and 506 samples. The purpose was to evaluate the effect of various predictor variables on housing price. For the predictors, continuous and binary variables are included. The binary predictor is an indicator of whether the census tract bordered the Charles Rivers. This predictor was also used by Harrison and Rubinfeld [15], Wang *et al.* [35] and others. In our model, the response is taken to be the median value of owner occupied homes and the partial linear model is assumed. The predictors $\boldsymbol{x} = (x_1, \ldots, x_{11})$ in the parametric component are x_1: the crime rate by town; x_2: the percentage of the town' residential land zoned for lots greater than 25 000 square feet; x_3: the percentage of nonmetal business acres per town; x_4: the indicator of whether the census tract borders the Charles Rivers; x_5: nitrogen oxide concentration in parts per hundred million (pphm); x_6: the weighted distance to five Boston employment centers from houses; x_7: the percentage of owner units built prior to 1940; x_8: an index of accessibility to radial highways; x_9: the full tax rate of the property; x_{10}: the pupil-teacher ratio by town school

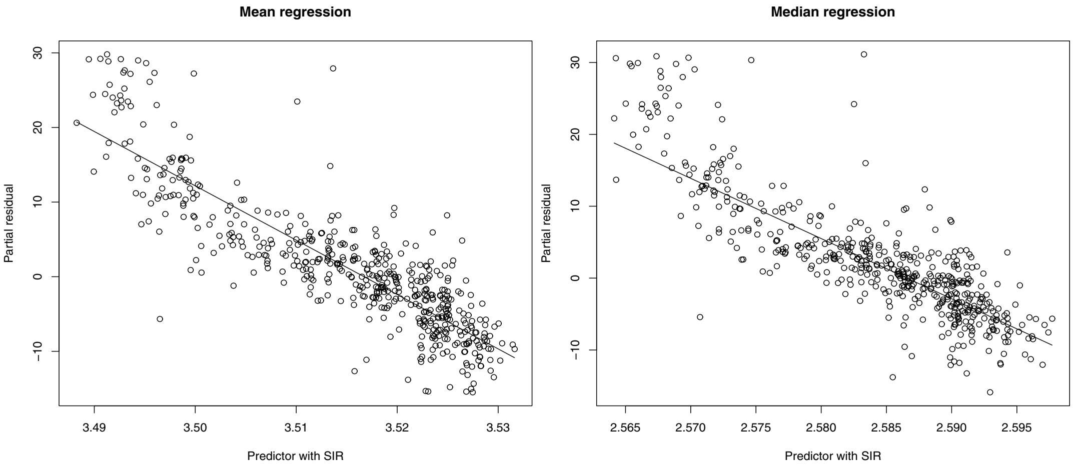

Figure 1. Linear regression with partial residual $y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and transformed predictor $\hat{\boldsymbol{\theta}}_1^T \boldsymbol{\phi}(\boldsymbol{z}_i)$. The solid lines in the left and right panels indicate the mean estimator and the median estimator, respectively.

Table 4. Estimators of $\boldsymbol{\beta}$ in the Boston Housing data. MR is the mean regression. QR is the quantile regression and CQR is the composite quantile regression.

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
MR	–0.132	0.026	–0.015	4.412	11.665	0.026	–0.145	0.084	–0.005	0.293	0.014
QR	–0.156	0.021	0.002	6.325	13.701	–0.015	-0.101	0.021	–0.004	0.227	0.012
CQR	–0.283	0.088	0.055	2.090	14.983	–0.060	–0.747	0.027	–0.020	0.280	0.012

district and x_{11}: the proportion of the population that is African-American. For the nonlinear component, we consider the surface of $\boldsymbol{z} = (z_1, z_2)$, where z_1 is the average number of rooms in owner units and z_2 is the percentage of the population in the area having low economic status. The same $\boldsymbol{z}$ was used by Doksum and Koo [9] for nonparametric smoothing.

We aim to investigate the relationship between the response and the various predictors using mean regression and median regression. Although we also applied CQR to the data, the efficiency of the estimator of g is not guaranteed since the error (the residual) does not appear to have symmetric density. Therefore we only discuss the estimator of $\boldsymbol{\beta}$ in CQR. We approximate the unknown function in the nonparametric component by the radial basis function model with the thin plate splines. The number of knots is $K = 64$ and the location is an 8×8 regular grid on the range of (z_1, z_2). For the SDR method, we use SIR with 20 data points per slice. However it was found by Chen and Li [4] that SIR is not sensitive to the number of slices. The dimension of the SIR is fixed at $d = 1$ to observe the linear regression with the projected predictor $\hat{\boldsymbol{\theta}}_1^T \boldsymbol{\phi}(\boldsymbol{z}_i)$ and the partial residual $y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$. Although we used a BIC to select d, the results are quite similar. Table 4 shows that the estimator of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{11})^T$ of the parametric component in the mean regression, median regression (quantile regression with $\tau = 0.5$) and CQR. The results are quite similar, but the estimators $\hat{\beta}_3$ and $\hat{\beta}_6$ in mean regression and those in median and CQR have a different sign although the size is very small. It appears that the robustness of quantile regression and CQR develops at difference of these sign.

We focus on the nonparametric component. In Figure 1, the dataset with partial residual $y_i - \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and transformed predictor $\hat{\boldsymbol{\theta}}_1^T \boldsymbol{\phi}(\boldsymbol{z}_i)$ and its linear estimators are illustrated. In each panel, the estimator can capture the structure of the dataset. Thus, the SDR method gives a good transformation for the radial basis functions

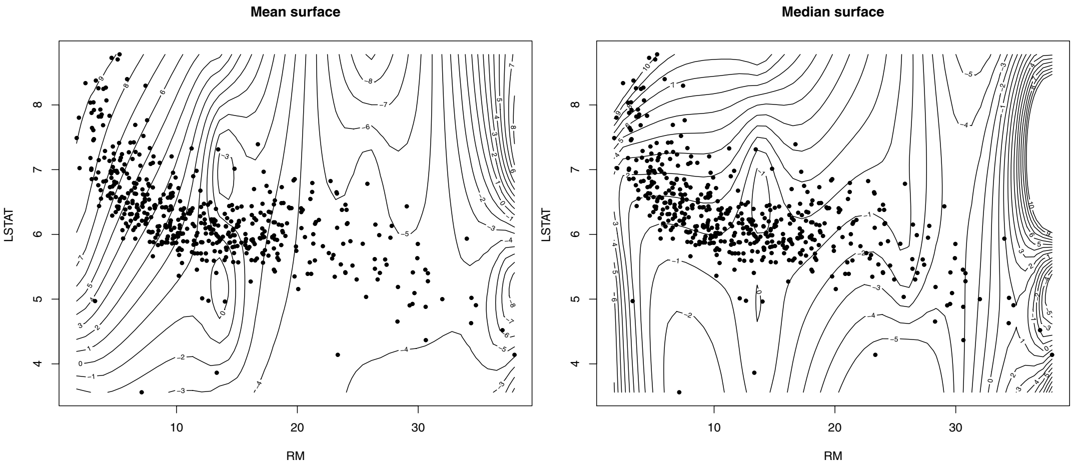

Figure 2. The estimator $\hat{g}(z_1, z_2)$ of the surface $g(z_1, z_2)$. The left and the right panels show the mean regression and median regression, respectively. The dots are the location of (z_{i1}, z_{i2}).

even with $d = 1$. Compared with the mean and median estimators, the mean line is to be pulled toward the data of the left upper area and some upper outlier points whereas the median line captures the data in a central area.

Figure 2 illustrates the estimator of surface $g(z_1, z_2)$. The left panel and the right panel are the mean and median surface estimator, respectively. Roughly speaking, the surfaces are drawn like a plane from the contour line. Note that both estimators have a gentle curve. Both surfaces show a rapid change at the bottom-left and upper-right areas, where there is no data. However in an area where the data is distributed, it appears the surface can capture the structure of the conditional mean/median. When we look at the contour line carefully in both panels, the median surface is slightly greater than the mean surface in the same location (z_1, z_2). This may indicate that the mean surface is affected by the data deviating from the population unlike the median surface.

6. Discussion

In this work, we proposed a new estimation with the sufficient dimension reduction (SDR) method for partial linear models. We assumed that the nonparametric component had the radial basis function models and aimed for the surface regression. Using the SDR method for the nonparametric component, a computationally stable and smooth curve was obtained, maintaining the performance of the estimates of the parametric component. The proposed method is considered in robust regression. In terms of the theoretical and numerical results, the proposed method performs well not only for mean regression but also for quantile and composite quantile regression. Although we used sliced inverse regression (SIR) as the SDR method, other methods such as sliced average variance estimation, minimum average variance estimation and directional regression can also be used. However to satisfy the $\sqrt{n}$-consistency and its condition, SIR is easy to apply.

There are several extensions for future studies. First if the dimension of $\boldsymbol{x}$ is large, then variable selection should be considered. To do this, the lasso [33], SCAD [10] and other methods have been developed. Direct use of such methods leads to the practice of variable selection on $\boldsymbol{x}$. However the smoothing parameter included in the above method should be selected appropriately. Together with the alternating algorithm and the optimization of quantile regression, the reduction of the computational cost should be studied. We consider this extension the most important issue for our next step. Second, the large dimension of $\boldsymbol{z}$ is considered. In this paper, the dimension q of $\boldsymbol{z}$ is 2 and we focused on the estimation of the nonlinear surface. For $q > 3$, the radial basis function method might not work well as the result of the curse of dimensionality. The regression with $q > 3$ would be interesting to explore. Finally, in our approach, the ordinary SDR method is used. In contrast, Wang *et al.* [35] and Feng *et al.* [12] studied the partial SDR method. The partial SDR method finds a matrix Θ such that

$$Y \perp\!\!\!\perp \boldsymbol{Z} | (\boldsymbol{X}, \Theta \boldsymbol{Z})$$

is satisfied. Thus, comparing their method and our method would be interesting. Although it is beyond the scope of this paper, we believe that the above three and other extensions are promising topics for further research.

Appendix

The proof of Theorem 3.1 is very simple. However we briefly mention this in order to describe where the assumptions are used.

Proof of Theorem 3.1. For simplicity, we write $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_{(t-1)}$ and all new symbols are defined without using the iteration step t. Let $\Lambda = \mathrm{Cov}[\boldsymbol{\phi}(\boldsymbol{Z})] - \mathrm{Cov}[E[\boldsymbol{\phi}(\boldsymbol{Z})|Y - \boldsymbol{X}^T\boldsymbol{\beta}]]$ and let $\tilde{\Lambda} = \mathrm{Cov}[\boldsymbol{\phi}(\boldsymbol{Z})] - \mathrm{Cov}[E[\boldsymbol{\phi}(\boldsymbol{Z})|Y - \boldsymbol{X}^T\tilde{\boldsymbol{\beta}}]]$. From Assumption A2 and B, we have

$$\begin{aligned} m(y - \boldsymbol{x}^T\tilde{\boldsymbol{\beta}}) &= m(y - \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{x}^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})) \\ &= m(y - \boldsymbol{x}^T\boldsymbol{\beta}) + O_P(n^{-1/2}). \end{aligned}$$

Therefore $\tilde{\Lambda} = \Lambda + O_P(n^{-1/2})$ and $\hat{\Lambda}$ converge to Λ at $n^{-1/2}$ rate. From the perturbation theory (for example; Tyler [34]), $\hat{\theta}_j (j = 1, \ldots, d)$ which is the eigenvectors of $\hat{\Lambda}$ converge to the corresponding eigenvectors of Λ. By Assumption A1, Θ fall in the dimension reduction subspace and hence $\hat{\Theta}$ converge to an sufficient dimension reduction subspace at the order $n^{-1/2}$. □

Proof of Theorem 3.2. First we show that the asymptotic property of $\hat{\mu} = \mu_{(t)}$ and $\hat{\boldsymbol{w}} = \boldsymbol{w}_{(t)}$. In this proof, for simplicity we write $\tilde{\Theta} = \Theta_{(t)}$ and $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_{(t-1)}$. By the same argument, the asymptotic consistency of $\boldsymbol{\beta}_{(t)}$ can be derived. Let $U_i = Y_i - \boldsymbol{x}^T\boldsymbol{\beta} - \mu - \boldsymbol{w}^T\Theta^T\boldsymbol{\phi}(\boldsymbol{z}_i)$ and $\tilde{U}_i = Y_i - \boldsymbol{x}^T\tilde{\boldsymbol{\beta}} - \mu - \boldsymbol{w}^T\tilde{\Theta}^T\boldsymbol{\phi}(\boldsymbol{z}_i)$ for $i = 1, \ldots, n$, and

$$Q_n(v, \boldsymbol{\delta}) = \sum_{i=1}^{n} \rho\left(\tilde{U}_i - n^{-1/2}v - n^{-1/2}\boldsymbol{\delta}^T\tilde{\Theta}^T\boldsymbol{\phi}(\boldsymbol{z}_i)\right) - \rho(\tilde{U}_i).$$

Then the minimizer of Q is obtained as

$$\begin{bmatrix} \hat{v} \\ \hat{\boldsymbol{\delta}} \end{bmatrix} = \sqrt{n}\begin{bmatrix} \hat{\mu} - \mu \\ \hat{\boldsymbol{w}} - \boldsymbol{w} \end{bmatrix}.$$

For simplicity, we define for $i = 1, \ldots, n$,

$$\alpha_i(v, \boldsymbol{\delta}) = n^{-1/2}v - n^{-1/2}\boldsymbol{\delta}^T\tilde{\Theta}^T\boldsymbol{\phi}(\boldsymbol{z}_i).$$

Define

$$R_n(v, \boldsymbol{\delta}, \tilde{U}_1, \ldots, \tilde{U}_n) = Q_n(v, \boldsymbol{\delta}) - E[Q_n(v, \boldsymbol{\delta})] - \sum_{i=1}^{n}\left\{\rho'(\tilde{U}_i) - E[\rho(\tilde{U}_i)]\right\}\alpha_i(\boldsymbol{\delta}).$$

It is easy to show that $E[R_n(v, \boldsymbol{\delta}, \tilde{U}_1, \ldots, \tilde{U}_n))] = 0$. From

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + o_P(1), \quad \tilde{\Theta} = \Theta + o_P(1) \tag{A.1}$$

and $V[R_n(v,\boldsymbol{\delta},U_1,\dots,U_n)] = o(1)$ by the tedious but easy calculation, we also obtain

$$V[R_n(v,\boldsymbol{\delta},\tilde{U}_1,\dots,\tilde{U}_n)] = o(1)$$

under the Assumption C2 and C4. Consequently, $R_n(v,\boldsymbol{\delta},\tilde{U}_1,\dots,\tilde{U}_n) = o_P(1)$. Therefore Q_n can be written as

$$Q_n(v,\boldsymbol{\delta}) = E[Q_n(v,\boldsymbol{\delta})] + \sum_{i=1}^n \{\rho'(U_i) - E[\rho(U_i)]\}\,\alpha_i(v,\boldsymbol{\delta}) + o_P(1).$$

Next we use the Taylor expansion of $\Psi(\alpha_i(v,\boldsymbol{\delta}))$ and $\Psi'(\alpha_i(v,\boldsymbol{\delta}))$ around $\alpha_i(v,\boldsymbol{\delta}) = 0$ and

$$\Psi_i(\alpha_i(v,\boldsymbol{\delta})) = \Psi_i(0) + \Psi_i'(0)\alpha_i(v,\boldsymbol{\delta}) + \frac{1}{2}\Psi_i''(0)\{\alpha_i(v,\boldsymbol{\delta})\}^2 + o_P(n^{-1}) \tag{A.2}$$

can be obtained. Since we have from (A.1) that

$$E[Q_n(v,\boldsymbol{\delta})] = \frac{1}{n}\sum_{i=1}^n \Psi_i(\alpha_i(v,\boldsymbol{\delta})) - \Psi_i(0),$$

and $\rho'(\tilde{U}_i) = \rho'(U_i) + o_P(1)$, (A.2) yields

$$Q_n(v,\boldsymbol{\delta}) = \frac{1}{2}\sum_{i=1}^n \Psi_i''(0)\{\alpha_i(v,\boldsymbol{\delta})\}^2 + \sum_{i=1}^n \rho'(U_i)\alpha_i(v,\boldsymbol{\delta}) + o_P(1). \tag{A.3}$$

Let

$$\boldsymbol{W}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \rho'(U_i)\begin{bmatrix} 1 \\ \Theta^T\boldsymbol{\phi}(\boldsymbol{z}_i)\end{bmatrix}.$$

By Assumption C3 and Lyapunov's theorem, $\boldsymbol{W}_n$ converges to the normal with mean 0 and covariance matrix $\Sigma_\phi(\{\rho'\}^2)$. The matrix $\Sigma_\phi((\{\rho'\}^2)$ is given in Section 3. Therefore

$$\sum_{i=1}^n \rho'(U_i)\alpha_i(v,\boldsymbol{\delta}) = \boldsymbol{W}_n^T\begin{bmatrix} v \\ \boldsymbol{\delta}\end{bmatrix} \xrightarrow{D} \boldsymbol{W}^T\begin{bmatrix} v \\ \boldsymbol{\delta}\end{bmatrix},$$

where $\boldsymbol{W} \sim N(\boldsymbol{0}, \Sigma_\phi((\{\rho'\}^2))$. Similarly the first term of the right hand side of (A.3) converges to

$$[v\ \boldsymbol{\delta}]^T\Sigma_\phi((\rho'')\begin{bmatrix} v \\ \boldsymbol{\delta}\end{bmatrix},$$

where $\Sigma_\phi((\rho'')$ is defined in Section 3. Consequently

$$Q_n(\boldsymbol{\delta}) \xrightarrow{D} Q_0(\boldsymbol{\delta}) \equiv \frac{1}{2}[\boldsymbol{v}^T\ \boldsymbol{\delta}^T]C(\Psi'')\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\delta}\end{bmatrix} + \boldsymbol{W}\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{\delta}\end{bmatrix}.$$

From Pollard [31] and Knight [21], the minimizer of Q_n converges to the minimizer of Q_0 since Q_0 is convex function with respect to $[v,\boldsymbol{\delta}^T]^T$. Therefore $[\hat{v},\hat{\boldsymbol{\delta}}^T]^T$ is asymptotically distributed on normal with mean $\boldsymbol{0}$ and covariance matrix $\Sigma_\phi(\rho'')^{-1}\Sigma_\phi(\{\rho'\}^2)\Sigma_\phi(\rho'')^{-1})$. As the same manner, asymptotic normality of $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_{(t)}$ can be shown. □

Acknowledgements. The authors are grateful to Editor, Associate Editor and two anonymous referees for their valuable comments and suggestions, which led to improvements of the paper. The research of the author was partially supported by KAKENHI 26730019.

References

[1] P.K. Bhattacharya and P.L. Zhao, Semiparametric inference in a partial linear model. *Ann. Statist.* **25** (1997) 244–262.

[2] R.J. Carroll, J. Fan, I. Gijbels and M.P. Wand, Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** (1997) 477–489.

[3] H. Chen, Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16** (1988) 136–141.

[4] C.H. Chen and K.C. Li Can, SIR be as popular as multiple linear regression. *Statist. Sinica* **8** (1998) 289–316.

[5] F. Chiaromonte, R.D. Cook and B. Li, Sufficient Dimension Reduction in Regressions With Categorical Predictors. *Ann. Statist.* **30** (2002) 475–497.

[6] R.D. Cook and S. Weisberg, Discussion of "Sliced inverse regression for dimension reduction" by K.C.Li. *J. Amer. Statist. Assoc.* **86** (1991) 328–332.

[7] P. Diaconis and D. Freedman, Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** (1984) 793–815.

[8] X. Ding, X.H. Zhou and Q. Wang, A partially linear single-index transformation model and its nonparametric estimation. *The Canadian J. Statistics* **43** (2015) 97–117.

[9] K. Doksum and J.Y. Koo, On spline estimators and prediction intervals in nonparametric smoothing. *Comput. Statist. Data Anal.* **35** (2000) 67–82.

[10] J. Fan and R. Li Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** (2001) 1348–1360.

[11] J. Fan, T.C. Hu and Y.K. Truong, Robust Nonparametric Function Estimation. *Scand. J. Statist.* **21** (1994) 433–446.

[12] Z. Feng, X.M. Wen, Z. Yu and L. Zhu, On partial sufficient dimension reduction with applications to partially linear multi-index models. *J. Amer. Statist. Assoc.* **108** (2013) 237–246.

[13] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw and W.A. Stahel, Robust Statistics: The approach based on influence functions, *Wiley Ser. Probab. Stat.* Wiley-Interscience (2005).

[14] W. Härdle, Partially Linear Models. Springer, New York (2000).

[15] D. Harrison and D. Rubinfeld, Hedonic housing pries and the demand for clean air. *J. Environ. Econ. Manage.* **5** (1978) 81–102.

[16] T. Hastie and R. Tibshirani, Generalized additive models. Chapman & Hall, London (1990)

[17] X. He and B. Shi, Bivariate tensor-product B-splines in a partially linear regression. *J. Multivariate Anal.* **58** (1996) 162–181.

[18] N. Heckman, Spline smoothing in a partly linear model. *J. Roy. Statist. Soc. Ser. A* **48** (1986) 244–248.

[19] D.R. Hunter and K. Lange, quantile regression via an MM algorithm. *J. Comp. Graph. Statist.* **9** (2000) 60–77.

[20] R. Jiang, Z.G. Zhou, W.M. Qian and W.Q. Shao, Single-index composite quantile regression. *J. Korean Statist. Soc.* **41** (2012) 323–332.

[21] K. Knight, Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist.* **26** (1998) 755–770.

[22] R. Koenker, Quantile regression. Cambridge Univ. Press, Cambridge (2005)

[23] R. Koenker and G. Bassett, Regression quantiles. *Econometrica* **46** (1978) 33–50.

[24] S. Lee, Efficient semiparametric estimation of a partially linear quantile regression model. *Econ. Theory* **19** (2003) 1–31.

[25] T.C.M. Lee and H. Oh, Robust penalized regression spline fitting with application to additive mixed modeling. *Comput. Statist.* **22** (2007) 159–171.

[26] K.C. Li, Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** (1991) 997–1008.

[27] B. Li and S. Wang, On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **33** (2007) 1580–1616.

[28] L. Li, and X. Yin, Sliced inverse regression with regularizations. *Biometrics.* **64** (2008) 124–131.

[29] Y. Li and L.X. Zhu, Asymptotics for sliced average variance estimation. *Ann. Statist.* **35** (2007) 41–69.

[30] X. Liu, L. Wang and H. Liang, Estimation and variable selection for semiparametric additive partial linear models. *Statist. Sinica* **21** (2011) 1225–1248.

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

[32] Y. Sun, Semiparametric efficient estimation of partially linear quantile regression models. *Ann. Econ. Finance* **6** (2005) 105–127.

[33] R. Tibshirani, Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** (1996) 267–288.

[34] D. Tyler, Asymptotic inference for eigenvectors. *Ann. Statist.* **9** (1981) 725–736.

[35] J.L. Wang L. Xue, L. Zhu and Y.S. Xhong, Estimation for a partial-linear single-index model. *Ann. Statist.* **38** (2010) 246–274.

[36] S.N. Wood, Thin plate regression splines. *J. R. Statist. Soc.* B. **65** (2003) 95–114.

[37] Y. Xia and W. Härdle, Semi-parametric estimation of partially linear single-index models. *J. Multivariate Anal.* **97** (2006) 1162–1184.

[38] Y. Xia, H. Tong, W.K. Li and L.X. Zhu, An adaptive estimation of dimension reduction space. *J. Roy. Statist. Soc. Ser. B.* **64** (2002) 363–410.

[39] Y. Yu and D. Ruppert, Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* **97** (2002) 1042–1054.

[40] L.P. Zhu, R. Li and H. Cui, Robust estimation for partially linear models with large-dimensional covariates. *Science China Mathematics.* **56** (2013) 2069–2088.

[41] L.X. Zhu, B.Q. Miao and H. Peng, Sliced inverse regression with large dimensional covariates. *J. Amer. Statist. Assoc.* **101** (2006) 630–643.

[42] H. Zou and M. Yuan, Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36** (2008) 1108–1126.