An adaptive estimation of MAVE

Qin Wang and Weixin Yao

An Adaptive Estimation of MAVE

Qin Wang *and Weixin Yao †

June 20, 2011

Abstract

Minimum average variance estimation (MAVE, Xia et al. 2002) is an effective dimension reduction method. It requires no strong probabilistic assumptions on the predictors, and can consistently estimate the central mean subspace. It is applicable to a wide range of models, including time series. However, the least squares criterion used in MAVE will lose its efficiency when the error is not normally distributed. In this article, we propose an adaptive MAVE which can be adaptive to different error distributions. We show that the proposed estimate has the same convergence rate as the original MAVE. An EM algorithm is proposed to implement the new adaptive MAVE. Using both simulation studies and a real data analysis, we demonstrate the superior finite sample performance of the proposed approach over the existing least squares based MAVE when the error distribution is non-normal and comparable performance when the error is normal.

*Department of Statistical Sciences and Operations Research, Virginia Commonwealth University. E-mail: qwang3@vcu.edu

†Corresponding author. Department of Statistics, Kansas State University. E-mail: wxyao@ksu.edu.

Key Words: Sufficient dimension reduction, Central mean subspace, MAVE, Adaptive estimation.

1 Introduction

Since the pioneer work of Li (1991), sufficient dimension reduction has received much attention as an efficient tool to tackle the challenging problem of high dimensional data analysis. The basic idea of sufficient dimension reduction in a regression problem is to replace the original high dimensional predictor with its appropriate low dimensional projection while preserving full regression information. Let y and $\mathbf{X}$ be a univariate response and a p-dimensional predictor vector respectively. A d-dimensional ($d \leq p$) subspace $\mathcal{S}$=Span$\{\mathbf{B}_{p\times d} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_d)\}$ is called a *dimension reduction subspace* of $y|\mathbf{X}$ if

$$y \perp\!\!\!\perp \mathbf{X}|P_{\mathcal{S}}\mathbf{X}, \tag{1.1}$$

where $\perp\!\!\!\perp$ indicates independence and $P_{(\cdot)}$ stands for an orthogonal projection operator in the standard inner product. When the intersection of all subspaces satisfying (1.1) also satisfies (1.1), it is called the *central subspace* (*CS*; Cook 1994, 1996, 1998) and is denoted by $\mathcal{S}_{y|\mathbf{X}}$. Its dimension, denoted by D, is then called the structural dimension of $y|\mathbf{X}$. When the conditional mean function is of primary interest, the objective of sufficient dimension reduction is to seek a d-dimensional subspace $\mathcal{S}$ such that

$$y \perp\!\!\!\perp E(y|\mathbf{X})|P_{\mathcal{S}}\mathbf{X}. \tag{1.2}$$

Subspaces satisfying condition (1.2) are called mean dimension reduction subspaces (Cook and Li, 2002). When the intersection of all subspaces satisfying condition

(1.2) also satisfies condition (1.2), it is called the *central mean subspace* (*CMS*) and is denoted by $\mathcal{S}_{E(y|\mathbf{X})}$. Its dimension is called the structural dimension of $E(y|\mathbf{X})$. As shown in Cook (1998) and Yin, Li and Cook (2008), under mild conditions, the *CS* and the *CMS* exist and are unique respectively. We assume the existence of the *CS* and the *CMS* throughout the article.

Knowledge of the *central subspace* or the *central mean subspace* is very useful for parsimoniously characterizing the conditional distribution of $y|\mathbf{X}$ or $E(y|\mathbf{X})$. All the existing dimension reduction methods can be classified into three groups according to the distribution form of interest. Sliced inverse regression (SIR; Li 1991), sliced average variance estimation (SAVE; Cook and Weisberg 1991), principal Hessian directions (PHD; Li 1992), and contour regression (CR; Li, Zha and Chiaromonte 2005) are among the methods to estimate the dimension reduction subspace through the inverse conditional distribution of $\mathbf{X}|y$. They are computationally efficient, but do impose certain probabilistic assumptions on the predictors. Forward regression approach directly targets on the conditional distribution $y|\mathbf{X}$ through the use of kernel smoothing techniques. Xia et al. (2002) proposed the minimum average variance estimation (MAVE) as the first attempt in this category. It is a nice combination of local linear smoothing and projection pursuit regression. It requires no strong assumptions on the probabilistic structure of predictor $\mathbf{X}$, and can be applied to time series models as well. MAVE can estimate the directions in the *central mean subspace* consistently without undersmoothing the link function. With the use of low dimensional kernel, the refined MAVE (rMAVE) can achieve a faster consistency rate and better estimation accuracy. The third group, the correlation approach such as Yin, Li and Cook (2008), investigates the joint information of $(y, \mathbf{X})$. OLS (Li and Duan, 1989) and PHD (Li, 1992) can be flexibly regarded as in this group as well.

Since the introduction of this novel tool, many related studies have been carried

out to improve MAVE in both theory and applications. Antoniadis et al. (2003) applied MAVE to tumor classification using gene expression data. Amato et al. (2006) extended MAVE to functional data analysis. Xia and Härdle (2006) applied MAVE to partially linear single-index models so that no $\sqrt{n}$-consistent pilot estimator is needed and the choice of bandwidth is more flexible. Čížek and Härdle (2006) proposed a robust version by replacing the least squares with local L- or M- estimation so that MAVE is robust to outliers in the dependent variable. Wang and Yin (2008) incorporated shrinkage estimation to MAVE so that variable selection and dimension reduction can be achieved simultaneously. Recently, Wang and Xia (2008) extended MAVE to the whole *central subspace* and proposed a new efficient estimation method called sliced regression (SR).

Despite the nice properties of MAVE as a useful tool in both dimension reduction and semi-parametric modeling, it is not the most efficient in the semi-parametric sense because of the use of least squares. This is also briefly mentioned in Xia and Härdle (2006). In many real applications, the error is very likely to be non-normally distributed and sometimes far from the normal. So it is natural to treat the error density as another unknown parameter similar to the link function.

In this article, we propose an adaptive estimation procedure based on the combination of kernel density estimation and MAVE so that the new estimator can be adaptive to different error distributions and thus improve the estimation efficiency when the error is not normal. We show that the proposed estimate has the same convergence rate as the original MAVE. A stable EM algorithm is proposed to implement the adaptive estimation. Using a Monte Carlo simulation study, we demonstrate that the proposed approach provides more efficient estimate than the existing least squares based MAVE when the error distribution is not normal. In addition, when the error is exactly normal, the new method is comparable to the existing MAVE. We illustrate

the proposed adaptive estimation method with an analysis of a real data set.

The rest of the article is organized as follows. Section 2 introduces the new adaptive estimation approach, including a brief review of the original MAVE, the adaptive estimation procedure, and the investigation of its asymptotic properties. Section 3 evaluates the numerical performance of the proposed approach through both simulation studies and a real data analysis. A short discussion is in section 4. All technical details are deferred to appendix.

2 Adaptive MAVE

2.1 A brief review of MAVE

The regression-type model of interest in MAVE can be written as

$$y = g(\mathbf{B}_0^T \mathbf{X}) + \epsilon, \tag{2.1}$$

where $g(\cdot)$ is an unknown smooth link function, $\mathbf{B}_0 = (\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0D})$ is a $p \times D$ orthogonal matrix ($\mathbf{B}_0^T \mathbf{B}_0 = I_{D\times D}$) with the structural dimension $D < p$ and $\mathrm{E}(\epsilon \mid \mathbf{X}) = 0$.

Given a random sample $\{(\mathbf{X}_i, y_i), i = 1, \ldots, n\}$, the MAVE estimates the *CMS* directions $\mathbf{B}_0$ by solving the following minimization problem

$$\min_{\mathbf{B}, a_j, \mathbf{b}_j, j=1,\ldots,n} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \left[y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T (\mathbf{X}_i - \mathbf{X}_j) \right\} \right]^2 w_{ij} \right), \tag{2.2}$$

with respect to $a_j \in \mathbb{R}^1$, $\mathbf{b}_j \in \mathbb{R}^d$ and $\mathbf{B}_{p\times d}$, where $\mathbf{B}^T\mathbf{B} = I_d$ and d is the working dimension. The kernel weight w_{ij} is a function of the distance between $\mathbf{X}_i$ and $\mathbf{X}_j$ satisfying $\sum_{i=1}^{n} w_{ij} = 1$. The minimization of (2.2) can be solved iteratively with

respect to $\{(a_j, \mathbf{b}_j), j = 1, \cdots, n\}$ and $\mathbf{B}$ separately. The estimation of MAVE is very efficient since only two quadratic programming problems are involved and both have explicit solutions. To improve the estimation accuracy, a lower dimensional kernel weight $\tilde{w}_{ij}$ as a function of $\tilde{\mathbf{B}}^T(\mathbf{X}_i - \mathbf{X}_j)$ can be used after an initial estimate $\tilde{\mathbf{B}}$ was obtained. The use of a smaller bandwidth in the refined procedure can also improve the consistency rate.

Note that in (2.2), the least square criterion is used. It corresponds to the maximum likelihood estimation (MLE) when the error is normally distributed. However, when the error distribution is not normal, the existing least squares based MAVE will lose some efficiency. Therefore, it is desirable to derive an estimator which can be adaptive to different error distributions. In the following, we will propose such an adaptive estimator based on the extension of the MAVE and kernel density estimate. In this article, we focus mainly on the estimation of the *CMS* directions $\mathbf{B}_0$ while the structural dimension D is assumed to be known. To determine the dimension D, the cross-validation approach proposed in Xia et al. (2002) and some other information based criteria can also be applied in our adaptive estimation framework. More details can be found in the previous paper and the references therein.

2.2 Adaptive estimation of the central mean subspace

Let $f_\epsilon(\epsilon)$ be the density function of ϵ. If f_ϵ is known, one would estimate the *CMS* directions by maximizing the following objective function

$$\max_{\mathbf{B}, a_j, \mathbf{b}_j, j=1,\ldots,n} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \log f_\epsilon \left[y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T (\mathbf{X}_i - \mathbf{X}_j) \right\} \right] w_{ij} \right). \tag{2.3}$$

However, in practice, f_ϵ is usually unknown but can be estimated by

$$\tilde{f}_\epsilon(\epsilon) = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(\epsilon - \tilde{\epsilon}_i),$$

where $K_{h_1}(\nu) = h_1^{-1}K(\nu/h_1)$ with $K(\nu)$ being a kernel function and h_1 being the bandwidth, $\tilde{\epsilon}_i = y_i - \tilde{g}(\tilde{\mathbf{B}}^T\mathbf{X}_i)$, and $\tilde{g}(\tilde{\mathbf{B}}^T\mathbf{X}_i) = \tilde{a}_i$ is the initial estimate based on either the traditional MAVE or other dimension reduction methods. Thus, our new adaptive MAVE (aMAVE), based on the initial residuals $\{\tilde{\epsilon}_i, i = 1, \ldots, n\}$, maximizes

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{n}\sum_{i=1}^{n} \log\left(\sum_{l=1}^{n} K_{h_1}\left[y_i - \left\{a_j + \mathbf{b}_j^T\mathbf{B}^T(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]\right) w_{ij}, \tag{2.4}$$

where $\boldsymbol{\theta} = \{\mathbf{B}, (a_j, \mathbf{b}_j), j = 1, \ldots, n\}$, and

$$w_{ij} = \frac{K_h\{(\mathbf{X}_i - \mathbf{X}_j)\}}{\sum_{l=1}^{n} K_h\{(\mathbf{X}_l - \mathbf{X}_j)\}},$$

with $K_h(\boldsymbol{\nu}) = h^{-p}\prod_{k=1}^{p} K(\nu_k/h)$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_p)^T$ being a p-dimensional vector and h being the bandwidth. The refined estimate can be obtained by replacing w_{ij} with

$$\tilde{w}_{ij} = \frac{K_{h_2}\{\tilde{\mathbf{B}}^T(\mathbf{X}_i - \mathbf{X}_j)\}}{\sum_{l=1}^{n} K_{h_2}\{\tilde{\mathbf{B}}^T(\mathbf{X}_l - \mathbf{X}_j)\}},$$

where $\tilde{\mathbf{B}}$ is an initial estimate of $\mathbf{B}_0$.

Theorem 2.1. *Suppose that the Conditions C1-C10 in the Appendix hold and model (2.1) is true. Let* $\mathbf{B}$ *be the CMS direction estimated from the adaptive MAVE. If* $nh^p/\log n \to \infty, h \to 0$, $d \geq D$, *and* $h_1 = h/\log(n)$, *then*

$$||(I - \mathbf{B}\mathbf{B}^T)\mathbf{B}_0|| = O_p(h^3 + h\delta_n + h^{-1}\delta_n^2),$$

where $\delta_n = \{\log n/(nh^p)\}^{1/2}$.

The condition of $h_1 = h/\log(n)$ is used by Linton and Xiao (2007) for the simplicity of adaptiveness proof. They claimed that a wider range of bandwidth for h_1 can be used without changing the convergence rate but with more complicated proof. In addition, it can be seen that our adaptive MAVE achieves the same convergence rate as the traditional MAVE. It would be desirable to compare the asymptotic variances of both estimators. However, similar to the traditional MAVE, it is not easy to provide the asymptotic variance and the asymptotic distribution for the proposed adaptive MAVE. Our simulation study demonstrates that the proposed estimate has better finite sample performance than the existing MAVE for various error distributions.

The idea of adaptiveness is not new. Beran (1974) and Stone (1975) considered adaptive estimation for location models. Bickel (1982), Manski (1984), Steigerwald (1992), Schick (1993), Drost and Klaassen (1997), Hodgson (1998), Yuan and De Gooijer (2007), and Yuan (2010) extended this adaptive idea to regression, time series and some other models. Linton and Xiao (2007) proposed an adaptive nonparametric regression estimator by maximizing the estimated local likelihood function, in which the unknown error density was replaced by a kernel density estimate using some initial regression estimate. Our proposed new estimation procedure uses similar kernel error idea of Stone (1975) and Linton and Xiao (2007) to gain the adaptiveness based on some consistent initial estimate.

2.3 Estimation Algorithm

Note that the maximizer of (2.4) does not have an explicit formula. In this section, we propose an EM algorithm to maximize (2.4) by noticing its mixture log-likelihood structure.

Algorithm 2.1. *Given the initial estimates* $\{\tilde{\epsilon}_i, i = 1, \ldots, n\}$ *and the initial value of* $\boldsymbol{\theta} = \{\mathbf{B}, (a_j, \boldsymbol{b}_j), j = 1, \ldots, n\}$*, denoted by* $\boldsymbol{\theta}^{(0)}$*, the EM algorithm to maximize (2.4) at the* $(k+1)^{st}$ *step is as follows:*

E step: *find the classification probabilities*

$$p_{ijl}^{(k+1)} = \frac{w_{ij} K_{h_1}\left[y_i - \left\{a_j^{(k)} + \boldsymbol{b}_j^{(k)^T} \mathbf{B}^{(k)^T}(\boldsymbol{X}_i - \boldsymbol{X}_j)\right\} - \tilde{\epsilon}_l\right]}{\sum_{m=1}^{n} K_{h_1}\left[y_i - \left\{a_j^{(k)} + \boldsymbol{b}_j^{(k)^T} \mathbf{B}^{(k)^T}(\boldsymbol{X}_i - \boldsymbol{X}_j)\right\} - \tilde{\epsilon}_m\right]}.$$

M step: *update parameter estimates of* $\boldsymbol{\theta}$ *by maximizing*

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\sum_{l=1}^{n} p_{ijl}^{(k+1)} \log K_{h_1}\left[y_i - \left\{a_j + \boldsymbol{b}_j^T \mathbf{B}^T(\boldsymbol{X}_i - \boldsymbol{X}_j)\right\} - \tilde{\epsilon}_l\right],$$

i.e., minimizing

$$\sum_{j=1}^{n}\sum_{i=1}^{n}\sum_{l=1}^{n} p_{ijl}^{(k+1)} \left[y_i - \left\{a_j + \boldsymbol{b}_j^T \mathbf{B}^T(\boldsymbol{X}_i - \boldsymbol{X}_j)\right\} - \tilde{\epsilon}_l\right]^2, \tag{2.5}$$

when $K(\nu)$ *is chosen to be a Gaussian kernel.*

Remark 1. The choice of a Gaussian kernel for $K(\nu)$ gives us a nice quadratic form as in (2.5). However, the kernel function in the calculation of w_{ij} and $\tilde{w}_{ij}$ need not be Gaussian. Other symmetric kernel functions can be used as well. Note that, however, as in most nonparametric regression, the choice of kernel function is not critical in terms of numerical results.

Remark 2. After getting the updated estimate of $\mathbf{B}$, one might also update the refined kernel weight $\tilde{w}_{ij}$ to improve the estimation accuracy but with more computation.

Similar to MAVE, (2.5) can be minimized with respect to $\{(a_j, \mathbf{b}_j), j = 1, \cdots, n\}$ and $\mathbf{B}$ iteratively.

1. Given $\mathbf{B} = \mathbf{B}^{(t)}$, the estimate of $(a_j, \mathbf{b}_j)$ is

$$\begin{pmatrix} a_j^{(t+1)} \\ \mathbf{b}_j^{(t+1)} \end{pmatrix} = \left\{ \sum_{i=1}^{n} \sum_{l=1}^{n} p_{ijl}^{(k+1)} \mathcal{X}_{ij} \mathcal{X}_{ij}^T \right\}^{-1} \left\{ \sum_{i=1}^{n} \sum_{l=1}^{n} p_{ijl}^{(k+1)} \mathcal{X}_{ij} (y_i - \tilde{\epsilon}_l) \right\}, \quad (2.6)$$

   where $\mathbb{X}_{ij} = (\mathbf{X}_i - \mathbf{X}_j), \mathcal{X}_{ij} = (1, \mathbb{X}_{ij}^T \mathbf{B}^{(t)})^T$.

2. Update $\mathbf{B}$ with estimated $(a_j^{(t+1)}, \mathbf{b}_j^{(t+1)})$

$$vec(\tilde{\mathbf{B}}^{(t+1)}) = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} p_{ijl}^{(k+1)} \mathbb{X}_{ijl} \mathbb{X}_{ijl}^T \right\}^{-1} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} p_{ijl}^{(k+1)} \mathbb{X}_{ijl} (y_i - a_j^{(t+1)} - \tilde{\epsilon}_l) \right\}, \quad (2.7)$$

   where $vec(\mathbf{B}) = (\beta_1^T, \cdots, \beta_d^T)^T$, $\mathbb{X}_{ijl} = \mathbf{b}_j^{(t+1)} \otimes \mathbb{X}_{ij}$ and $\otimes$ represents the Kronecker product.

3. Orthonormalize the estimated $\mathbf{B}$ as

$$\mathbf{B}^{(t+1)} = \tilde{\mathbf{B}}^{(t+1)} (\tilde{\mathbf{B}}^{(t+1)^T} \tilde{\mathbf{B}}^{(t+1)})^{-\frac{1}{2}}.$$

4. Repeat step 1 through step 3 until some convergence criterion is met. For example, the matrix norm $||\mathbf{B}^{(t+1)} \mathbf{B}^{(t+1)^T} - \mathbf{B}^{(t)} \mathbf{B}^{(t)^T}||$ can be used to compare with some pre-specified tolerance value.

The above EM algorithm monotonically increases the local log-likelihood (2.4) after each iteration, as shown in the following theorem.

Theorem 2.2. *Each iteration of the above E and M steps will monotonically increase*

the local log-likelihood (2.4), i.e.,

$$\ell(\boldsymbol{\theta}^{(k+1)}) \geq \ell(\boldsymbol{\theta}^{(k)}),$$

for all k, where $\ell(\cdot)$ is defined as in (2.4).

3 Examples

In this section, we first conduct a simulation study to compare our proposed adaptive MAVE (aMAVE) with the traditional least squared based refined MAVE (rMAVE) for different kinds of error densities. Then a baseball hitters' salary data is applied to illustrate the new adaptive MAVE. For aMAVE, we simply use a rule of thumb bandwidth $h_1 = 1.06n^{-1/5}\hat{\sigma}$ for the kernel density estimate of $f_\epsilon(\epsilon)$, where $\hat{\sigma}$ is a robust estimate of σ based on the initial residuals $\{\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n\}$, i.e.,

$$\hat{\sigma} = \min\{(\tilde{\epsilon}_{(0.75)} - \tilde{\epsilon}_{(0.25)})/1.34,\ \sigma(\tilde{\epsilon})\},$$

where $\tilde{\epsilon}_{(p)}$ is the p^{th} sample quantile of $\{\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n\}$ and $\sigma(\tilde{\epsilon})$ is the sample standard deviation of $\{\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n\}$. Better estimates might be obtained if using some more sophisticated bandwidth for kernel density estimation. See, for example, Sheather and Jones (1991) and Raykar and Duraiswami (2006). In addition, one might also use cross validation method to select the bandwidth, which requires more computation. The Gaussian kernel is used in the calculation of w_{ij} and the choice of bandwidth follows Xia et al (2002).

3.1 Simulation studies

The following four error distributions f_ϵ of ϵ (with mean 0 and standard error around 1) are considered in our numerical experiment. The standard normal distribution serves as a baseline in our comparison. The second one is a scaled t-distribution with 3 degrees of freedom. The third density is bimodal and the last one is left skewed.

1. $N(0,1)$;
2. $t_3/\sqrt{3}$;
3. $0.5N(-1,0.5^2)+0.5N(1,0.5^2)$;
4. $0.3N(-1.4,1)+0.7N(0.6,0.4^2)$.

For each of the above error distributions, we consider the following three models:

Model 1: $y=\beta^T\mathbf{X}+\epsilon$, where $\beta=(1,1,0,\cdots,0)^T/\sqrt{2}$.

Model 2:

$$y=\frac{\beta_1^T\mathbf{X}}{0.5+(1.5+\beta_2^T\mathbf{x})^2}+0.5\epsilon,$$

where $\beta_1=(1,0,\cdots,0)^T$ and $\beta_2=(0,1,0,\cdots,0)^T$.

Model 3: $y=\cos(2\beta_1^T\mathbf{X})-\cos(\beta_2^T\mathbf{X})+0.5\epsilon$, where $\beta_1=(1,0,\cdots,0)^T$ and $\beta_2=(0,1,0,\cdots,0)^T$.

Given the generated data $\{(\mathbf{X}_1,y_1),\ldots,(\mathbf{X}_n,y_n)\}$ where $\mathbf{X}=(x_1,\cdots,x_p)$ are independent standard normal random variables, we estimate the *CMS* directions based on our new aMAVE and the traditional rMAVE. In order to compare different estimators, we use the space distance measure m defined as $||(I-\mathbf{B}_0\mathbf{B}_0^T)\mathbf{B}||$ if $d<D$ and $||(I-\mathbf{B}\mathbf{B}^T)\mathbf{B}_0||$ if $d\geq D$ (Xia et al, 2002). The number of data replicates is 500.

Let p and n be the dimension of β and the sample size, respectively. Tables 1, 2, and 3 report the estimation accuracy comparison based on the average m^2 for three models with different combinations of (n, p) and various error distributions f_ϵ, respectively. From the summary of all three models, we can see that the proposed aMAVE is comparable to the rMAVE for normal errors but more efficient than the rMAVE when the error is non-normal and the efficiency gain can be quite substantial even for small sample sizes.

Table 1: Model 1 Estimation Accuracy Comparison m^2

f_ϵ		$n = 50,\ p = 5$	$n = 50,\ p = 10$	$n = 100,\ p = 5$	$n = 100,\ p = 10$
1	rMAVE	0.081	0.213	0.038	0.099
	aMAVE	0.084	0.212	0.040	0.100
2	rMAVE	0.067	0.169	0.036	0.084
	aMAVE	0.055	0.145	0.026	0.067
3	rMAVE	0.106	0.258	0.048	0.123
	aMAVE	0.070	0.231	0.022	0.076
4	rMAVE	0.116	0.261	0.046	0.125
	aMAVE	0.060	0.187	0.017	0.062

Table 2: Model 2 Estimation Accuracy Comparison (m_1^2, m_2^2)

f_ϵ		$n = 50,\ p = 5$	$n = 100,\ p = 5$	$n = 100,\ p = 10$	$n = 200,\ p = 5$	$n = 200,\ p = 10$
1	rMAVE	0.063, 0.196	0.020, 0.043	0.096, 0.215	0.008, 0.013	0.029, 0.060
	aMAVE	0.058, 0.191	0.018, 0.041	0.089, 0.205	0.007, 0.012	0.026, 0.053
2	rMAVE	0.055, 0.138	0.016, 0.034	0.082, 0.181	0.006, 0.011	0.028, 0.048
	aMAVE	0.050, 0.121	0.012, 0.026	0.066, 0.156	0.004, 0.007	0.020, 0.033
3	rMAVE	0.074, 0.215	0.025, 0.056	0.120, 0.275	0.009, 0.016	0.035, 0.073
	aMAVE	0.067, 0.198	0.021, 0.047	0.105, 0.260	0.005, 0.009	0.025, 0.055
4	rMAVE	0.080, 0.223	0.025, 0.058	0.113, 0.273	0.010, 0.017	0.042, 0.077
	aMAVE	0.067, 0.202	0.017, 0.040	0.093, 0.243	0.005, 0.008	0.025, 0.048

Table 3: Model 3 Estimation Accuracy Comparison (m_1^2, m_2^2)

f_ϵ		$n = 50,\ p = 5$	$n = 100,\ p = 5$	$n = 100,\ p = 10$	$n = 200,\ p = 5$	$n = 200,\ p = 10$
1	rMAVE	0.030, 0.082	0.008, 0.025	0.046, 0.111	0.002, 0.010	0.011, 0.031
	aMAVE	0.030, 0.079	0.008, 0.024	0.046, 0.110	0.002, 0.009	0.011, 0.029
2	rMAVE	0.024, 0.070	0.005, 0.022	0.039, 0.070	0.002, 0.009	0.010, 0.028
	aMAVE	0.023, 0.059	0.004, 0.017	0.034, 0.061	0.001, 0.006	0.008, 0.020
3	rMAVE	0.043, 0.101	0.009, 0.031	0.063, 0.139	0.003, 0.013	0.015, 0.039
	aMAVE	0.040, 0.097	0.007, 0.025	0.059, 0.126	0.002, 0.008	0.010, 0.028
4	rMAVE	0.037, 0.092	0.009, 0.031	0.063, 0.138	0.003, 0.013	0.014, 0.036
	aMAVE	0.031, 0.078	0.006, 0.019	0.059, 0.126	0.002, 0.006	0.008, 0.022

3.2 Hitters' salary data

This data concerns the salary of 263 major league baseball hitters in 1987 and their performance. An obvious question of interest is "Are they paid based on their performance?". It has drawn much attention from statisticians. Among others, Chaudhuri et al. (1994) proposed a piece-wise polynomial regression tree (SUPPORT) approach. Li et al. (2000) proposed a dimension-reduction based regression tree, PHDRT, and identified several outliers. Xia et al. (2002) applied MAVE to find the low dimensional projection and chose a partially linear model to fit the data. All previous studies suggested using different models to fit different parts of the data. Along the same line, we split the data into two groups (junior/veteran) based on 'the years in the major leagues' and the cutoff is chosen to be 7 as suggested by Chaudhuri et al. (1994). The response variable is taken to be the logarithm of the annual salary in 1987 as in all previous studies, and the 13 predictors used in our analysis are listed in Table 4.

We apply the adaptive MAVE to both groups, and one significant direction is identified for each group as shown in Table 4. The scatter plot of response vs the direction (Figure 1) shows clear linear patterns in both groups, which is consistent with previous studies. The sign change of the coefficient estimates for the variable

Table 4: Hitters' salary data

Performance			$\hat{\boldsymbol{\beta}}_{junior}$	$\hat{\boldsymbol{\beta}}_{veteran}$
in 1986	x_1	time at bat	-0.242	0.192
	x_2	hits	0.372	-0.035
	x_3	home runs	0.134	-0.016
	x_4	runs	-0.069	0.004
	x_5	runs batted in	-0.115	-0.049
	x_6	walks	0.111	0.058
up to 1986	x_7	years in major leagues	0.305	-0.206
	x_8	time at bat	-0.113	-0.704
	x_9	hits	0.751	0.552
	x_{10}	home runs	-0.048	-0.026
	x_{11}	runs	0.164	0.202
	x_{12}	runs batted in	0.207	0.266
	x_{13}	walks	0.114	0.007

x_7 (years in the major league) between the two groups supports the existence of an 'aging effect' as discovered by Li et al. (2000) and Xia et al. (2002).

Furthermore, to compare the performance of our adaptive approach with the traditional MAVE, 100 bootstrap samples are drawn from the original data. Both aMAVE and rMAVE are applied to each bootstrap sample. The average distance of the direction estimates from bootstrap samples to the direction from the original data is calculated. For the veteran group, the average distance from rMAVE is 0.282, compared to 0.240 from aMAVE. For the junior group, the average distances from rMAVE and aMAVE are 0.571 and 0.564, respectively. A detailed look of the residual Q-Q plots in Figure 1 from the local linear estimation might give some explanations to the improvement of aMAVE over rMAVE, especially for the veteran group. For the junior group, the residual is very close to normal distribution since the Q-Q plot is almost a straight line. Our adaptive MAVE gives comparable result as rMAVE. But

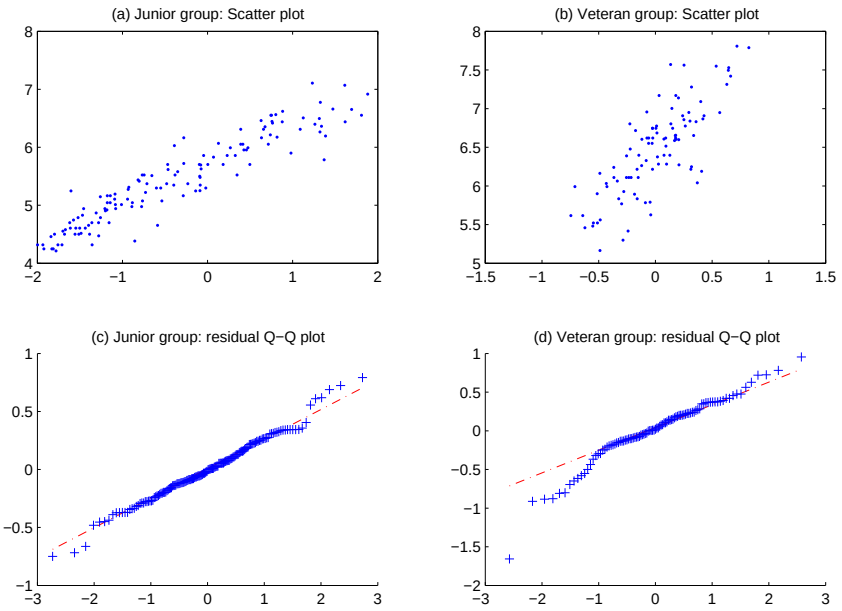

Figure 1: Hitters' salary: the plot of response vs the direction from aMAVE and the residual Q-Q plot

for the veteran group, a clear long tail shows up in the Q-Q plot, which explains the improved estimation efficiency from aMAVE over rMAVE.

4 Discussion

We have developed an adaptive MAVE to estimate the dimension reduction subspace more efficiently. Its estimation can be easily implemented with the proposed EM-type algorithm. Based on our empirical study for various error densities and models, the proposed aMAVE is more efficient than the existing least squares based MAVE, even for small sample sizes, when the error density is not normally distributed. In addition, the aMAVE is comparable to MAVE if the error distribution is exactly normal. It is proved that the adaptive MAVE has the same consistency rate as the MAVE.

In this paper, we focus only on estimating the *CMS* directions through MAVE

formulation. The proposed approach can be easily adapted to other dimension reduction methods. It can also be combined with shrinkage estimation to estimate the *CMS* directions and to select informative variables simultaneously. Such extensions are of great interest in our future research.

Acknowledgments

The authors would like to thank the editor, an associate editor, and two anonymous referees for their valuable comments and suggestions.

Appendix

A. Regularity conditions

C1. $\{(\mathbf{X}_i, y_i), i = 1, \ldots, n\}$ are iid sequence from the joint density $f_{\mathbf{X},y}(\mathbf{x}, y)$.

C2. $\{\varepsilon_i\}$ are i.i.d. with $\mathrm{E}(\epsilon_i) = 0, \mathrm{E}(|\epsilon_i|^3) < \infty$. $\{\mathbf{X}_i\}$and $\{\varepsilon_i\}$ are mutually independent. Additionally, the predictor $\mathbf{X}$ has a bounded support.

C3. The density $f_\epsilon(\cdot)$ of ε has bounded continuous derivatives up to order 4. Let $\ell(\varepsilon) = \log f_\epsilon(\varepsilon)$. Assume $\ell'''(\cdot)$ is bounded and $\mathrm{E}\{\ell'(\varepsilon)^2 + |\ell''(\varepsilon)| + |\ell'''(\varepsilon)|\} < \infty$.

C4. $\mathrm{E}|y|^k < \infty$ for all $k > 0$, $\mathrm{E}||\mathbf{X}||^k < \infty$ for all $k > 0$.

C5. The density function $f_{\mathbf{X}}(\cdot)$ of $\mathbf{X}$ has bounded derivatives up to order 4 and is abounded away from 0 in a neighbor around 0.

C6. The density function $f_y(\cdot)$ of y has bounded derivative and is bounded away from 0 on a compact support.

C7. The conditional densities $f_{\mathbf{X}|y}(\cdot)$ of $\mathbf{X}$ given y and $f_{(\mathbf{X}_0,\mathbf{X}_l)|(y_0,y_l)}(\cdot)$ of $(\mathbf{X}_0, \mathbf{X}_l)$ given (y_0, y_l) are bounded for all $l \geq 1$.

C8. $g(\cdot)$ has bounded, continuous 3^{rd} derivatives.

C9. $\mathrm{E}(\mathbf{X} \mid y)$ and $\mathrm{E}(\mathbf{X}\mathbf{X}^T \mid y)$ have bounded, continuous 3^{rd} derivatives.

C10. $K(\cdot)$ is a spherical symmetric density function with a bounded derivative and support. All the moments of $K(\cdot)$ exist and $\int UU^T K(U)dU = I$.

The above conditions are imposed to facilitate the proof and most of them are similar to Xia et al (2002). They are not the weakest possible conditions. For example, for C1, $\{(\mathbf{X}_i, y_i), i = 1, \ldots, n\}$ can be weakened to have a stationary and absolutely

regular sequence. The independence of $\{\mathbf{X}_i\}$ and $\{\varepsilon_i\}$ can be also relaxed based on the discussion of Section 4 of Linton and Xiao (2007). In addition, C4 can be weakened to have the existence of finite moments.

B. Proof of Theorem 2.1

Note that the estimate $\hat{\boldsymbol{\theta}} = \{\hat{a}_j, \hat{\mathbf{b}}_j, j = 1, \ldots, n, \hat{\mathbf{B}}\}$ is the maximizer of the following objective function

$$\max_{\mathbf{B}, a_j, \mathbf{b}_j, j=1,\ldots,n} \left(\sum_{j=1}^{n} \sum_{i=1}^{n} \log \tilde{f}_\epsilon \left[y_i - \left\{ a_j + \mathbf{b}_j^T \mathbf{B}^T (\mathbf{X}_i - \mathbf{X}_j) \right\} \right] w_{ij} \right), \tag{4.1}$$

where

$$\tilde{f}_\epsilon(\epsilon) = \frac{1}{n} \sum_{i=1}^{n} K_{h_1}(\epsilon - \tilde{\epsilon}_i)$$

is the kernel density estimate of $f_\epsilon(\cdot)$, and $\tilde{\epsilon}_i$ is the residual based on the traditional MAVE estimate. Based on the adaptive nonparametric regression result of Linton and Xiao (2007), the convergence rate of $\hat{\boldsymbol{\theta}}$ in (4.1)is the same as the true density $f_\epsilon(\cdot)$ is used. Therefore, we will mainly prove the convergence rate of $\hat{\boldsymbol{\theta}}$ assuming $f_\epsilon(\cdot)$ is known. Since the basic idea of our proof is very similar to Xia et al. (2002), we adopt the same notations for the ease of readers to follow.

Let $\mathbf{V}$ denote the gradient of $g(\cdot)$ w.r.t its arguments, i.e.,

$$\mathbf{V}(u_1, \ldots, u_D) = \partial g(u_1, \ldots, u_D)/\partial U$$

and $\mathbf{V}_k(u_1, \ldots, u_D) = \partial g(u_1, \ldots, u_D)/\partial u_k$. Similarly we define $\mathbf{V}^2_{k,l}(u_1, \ldots, u_D) = \partial^2 g(u_1, \ldots, u_D)/(\partial u_k \partial u_l)$ and $\mathbf{V}^3_{k,l,m}(u_1, \ldots, u_D) = \partial^3 g(u_1, \ldots, u_D)/(\partial u_k \partial u_l \partial u_m), 1 \leq k, l, m \leq D$.

Based on the Taylor expansion of $g(\mathbf{B}_0^T\mathbf{X}_i)$ for $\mathbf{X}_i$ close to $\mathbf{x}$, we have

$$g(\mathbf{B}_0^T\mathbf{X}_i) = g(\mathbf{B}_0^T\mathbf{x}) + (\mathbf{X}_i - \mathbf{x})^T\mathbf{B}_0\mathbf{V}(\mathbf{B}_0^T\mathbf{x}) + P_{2,i}(\mathbf{x}) + P_{3,i}(\mathbf{x}) + R_i(\mathbf{x}), \qquad (4.2)$$

where $\mathbf{B}_0 = (\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0D})$,

$$P_{2,i}(\mathbf{x}) = h^2 P_{2,h,i}(\mathbf{x}) = \frac{1}{2}\sum_{k,l=1}^{D} \mathbf{V}_{k,l}^2(\mathbf{B}_0^T\mathbf{x})\{\boldsymbol{\beta}_{0k}^T(\mathbf{X}_i - \mathbf{x})\}\{\boldsymbol{\beta}_{0l}^T(\mathbf{X}_i - \mathbf{x})\},$$

$$P_{3,i}(\mathbf{x}) = h^3 P_{3,h,i}(\mathbf{x}) = \frac{1}{6}\sum_{k,l,m=1}^{D} \mathbf{V}_{k,l,m}^3(\mathbf{B}_0^T\mathbf{x})\{\boldsymbol{\beta}_{0k}^T(\mathbf{X}_i - \mathbf{x})\}\{\boldsymbol{\beta}_{0l}^T(\mathbf{X}_i - \mathbf{x})\}\{\boldsymbol{\beta}_{0m}^T(\mathbf{X}_i - \mathbf{x})\},$$

and $R_i(\mathbf{x})$ is defined as the reminder. Let $\mathbf{X}_{h,i}(\mathbf{B}, \mathbf{x}) = (1, (\mathbf{X}_i - \mathbf{x})^T\mathbf{B}/h)^T$ and $K_{h,i}(\mathbf{x}) = K_h(\mathbf{X}_i - \mathbf{x})$. Since $\mathbf{B}_0 = \mathbf{B}\mathbf{B}^T\mathbf{B}_0 + (I - \mathbf{B}\mathbf{B}^T)\mathbf{B}_0$, we have

$$\begin{aligned} y_i = {} & \mathbf{X}_{h,i}^T(\mathbf{B}, \mathbf{x}) \begin{pmatrix} g(\mathbf{B}_0^T\mathbf{x}) \\ \mathbf{B}^T\mathbf{B}_0\mathbf{V}(\mathbf{B}_0^T\mathbf{x})h \end{pmatrix} \\ & + (\mathbf{X}_i - \mathbf{x})^T(I - \mathbf{B}\mathbf{B}^T)\mathbf{B}_0\mathbf{V}(\mathbf{B}_0^T\mathbf{x}) + h^2 P_{2,h,i}(\mathbf{x}) + h^3 P_{3,h,i}(\mathbf{x}) + R_i(\mathbf{x}) + \varepsilon_i. \quad (4.3) \end{aligned}$$

Consider the local likelihood criterion based on local linear kernel smooth

$$T_n(\mathbf{B}, \mathbf{x}) = \sum_{i=1}^{n} \ell\left\{ y_i - \mathbf{X}_{h,i}^T(\mathbf{B}, \mathbf{x}) \begin{pmatrix} a(\mathbf{x}) \\ b(\mathbf{x})h \end{pmatrix} \right\} K_{h,i}(\mathbf{x}),$$

where $\ell(\cdot) = \log f_\epsilon(\cdot)$. Note that for any fixed $\mathbf{x}$ and $\mathbf{B}$, $\{\hat{a}(\mathbf{x}), \hat{b}(\mathbf{x})h\}$ is the maximizer of $T_n(\mathbf{B}, \mathbf{x})$. Based on the Taylor expansion and the order of third derivative of

$T_n(\mathbf{B}, \mathbf{x})$, we have

$$\begin{pmatrix} \hat{a}(\mathbf{x}) \\ \hat{b}(\mathbf{x})h \end{pmatrix} = \begin{pmatrix} g(\mathbf{B}_0^T\mathbf{x}) \\ \mathbf{B}^T\mathbf{B}_0\mathbf{V}(\mathbf{B}_0^T\mathbf{x})h \end{pmatrix} + S_n^{-1}(\mathbf{B}, \mathbf{x})W_n(\mathbf{B}, \mathbf{x})$$

where

$$S_n(\mathbf{B}, \mathbf{x}) = n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B}, \mathbf{x})\mathbf{X}_{h,i}^T(\mathbf{B}, \mathbf{x})\ell''(\varepsilon_i)(1 + O_p(a_n))$$

and

$$W_n(\mathbf{B}, \mathbf{x}) = n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B}, \mathbf{x})\ell'(r_i),$$

with $a_n = h^2 + \delta_n$, $\delta_n = (nh^p)^{-1/2}(\log n)^{1/2}$, and

$$r_i = y_i - \mathbf{X}_{h,i}^T(\mathbf{B}, \mathbf{x}) \begin{pmatrix} g(\mathbf{B}_0^T\mathbf{x}) \\ \mathbf{B}^T\mathbf{B}_0\mathbf{V}(\mathbf{B}_0^T\mathbf{x})h \end{pmatrix}.$$

From Taylor expansion and the bounded $\ell'''(\cdot)$, we have

$$\begin{aligned} \begin{pmatrix} \hat{a}(\mathbf{x}) \\ \hat{b}(\mathbf{x})h \end{pmatrix} = \begin{pmatrix} g(\mathbf{B}_0^T\mathbf{x}) \\ \mathbf{B}^T\mathbf{B}_0\mathbf{V}(\mathbf{B}_0^T\mathbf{x})h \end{pmatrix} &+ S_n^{-1}(\mathbf{B}, \mathbf{x})n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B}, \mathbf{x})\ell'(\varepsilon_i) \\ &+ S_n^{-1}(\mathbf{B}, \mathbf{x})n^{-1}\sum_{i=1}^{n} K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B}, \mathbf{x})\ell''(\varepsilon_i)(r_i - \varepsilon_i) \end{aligned} \tag{4.4}$$

Let

$$L_{n,i}(\mathbf{B},\mathbf{x}) = (\mathbf{X}_i - \mathbf{x})^T - \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})S_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})(\mathbf{X}_i-\mathbf{x})^T,$$

$$Q_{n,i}(\mathbf{B},\mathbf{x}) = \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})S_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\{P_{2,h,i}(\mathbf{x}) + P_{3,h,i}(\mathbf{x})h\},$$

$$\Xi_{n,i}(\mathbf{B},\mathbf{x}) = \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})S_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\ell'(\varepsilon_i),$$

$$R_{n,i}(\mathbf{B},\mathbf{x}) = \mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})S_n^{-1}(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})R_i(\mathbf{x}).$$

Replace the estimator of $\hat{a}(\mathbf{x})$ and $\hat{b}(\mathbf{x})$ in $T_n(\mathbf{B},\mathbf{x})$, we have

$$\sum_{j=1}^{n}\frac{T_n(\mathbf{B},\mathbf{X}_j)}{\varsigma_n(\mathbf{X}_j)} = \sum_{j=1}^{n}\sum_{i=1}^{n} l\left\{L_{n,i}(\mathbf{B},\mathbf{X}_j)(I-\mathbf{B}\mathbf{B}^T)\mathbf{B}_0\mathbf{V}(\mathbf{B}_0^T\mathbf{X}_j) + \Delta_{ij}(\mathbf{B})\right\}\frac{K_{h,i}(\mathbf{X}_j)}{\varsigma(\mathbf{X}_j)}, \tag{4.5}$$

where $\varsigma(\mathbf{x}) = n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})$, and

$$\Delta_{ij}(\mathbf{B}) = \varepsilon_i + \{P_{2,h,i}(\mathbf{X}_j) + P_{3,h,i}(\mathbf{X}_j)h\}h^2 + R_i(\mathbf{X}_j) - Q_{n,i}(\mathbf{B},\mathbf{X}_j)h^2 - \Xi_{n,i}(\mathbf{B},\mathbf{X}_j) - R_{n,i}(\mathbf{B},\mathbf{X}_j).$$

Note that

$$\begin{aligned}\sum_{j=1}^{n}\frac{T_n(\mathbf{B},\mathbf{X}_j)}{\varsigma_n(\mathbf{X}_j)} = \sum_{j=1}^{n}\sum_{i=1}^{n} l\Big\{ & L_{n,i}(\mathbf{B},\mathbf{X}_j)(I-\mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0k}\mathbf{V}_k(\mathbf{B}_0^T\mathbf{X}_j) \\ & + \sum_{l\neq k}L_{n,i}(\mathbf{B},\mathbf{X}_j)(I-\mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0l}\mathbf{V}_l(\mathbf{B}_0^T\mathbf{X}_j) + \Delta_{ij}(\mathbf{B})\Big\}K_{h,i}(\mathbf{X}_j)/\varsigma(\mathbf{X}_j),\end{aligned} \tag{4.6}$$

Following Xia et al (2002), we have

$$(I - \mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0k} = D_{k,k}^{-1} \sum_{j=1}^{n} \sum_{i=1}^{n} \left\{ \ell' \left(\sum_{l \neq k} L_{n,i}(I - \mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0l}\mathbf{V}_l + \Delta_{ij} \right) K_{h,i}(\mathbf{X}_j)\mathbf{V}_k L_{n,i}^T \right\} \varsigma_n^{-1}(\mathbf{X}_j) \tag{4.7}$$

where

$$D_{k,l} = -\sum_{j=1}^{n} \sum_{i=1}^{n} \ell'' \left(\sum_{l \neq k} L_{n,i}(I - \mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0l}\mathbf{V}_l + \Delta_{ij} \right) K_{h,i}\mathbf{V}_k\mathbf{V}_l L_{n,i}^T L_{n,i} / \varsigma_n(\mathbf{X}_j). \tag{4.8}$$

Since

$$\begin{aligned}
&n^{-2}D_{k,l} \\
&= -n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} \ell''(\varepsilon_i) K_{h,i}(\mathbf{X}_j)\mathbf{V}_k(\mathbf{B}_0^T\mathbf{X}_j)\mathbf{V}_l(\mathbf{B}_0^T\mathbf{X}_j) L_{n,i}^T(\mathbf{B}, \mathbf{X}_j) L_{n,i}(\mathbf{B}, \mathbf{X}_j) / \varsigma_n(\mathbf{X}_j)(1 + o_p(1)) \\
&= -h^2 n^{-1} \sum_{j=1}^{n} \mathbf{V}_k(\mathbf{B}_0^T\mathbf{X}_j)\mathbf{V}_l(\mathbf{B}_0^T\mathbf{X}_j)(I - \mathbf{B}\mathbf{B}^T)\mathrm{E}\{\ell''(\varepsilon)\} + O_p(h^3 + h\delta_n)
\end{aligned}$$

and

$$\begin{aligned}
&\ell' \left(\sum_{l \neq k} L_{n,i}(I - \mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0l}\mathbf{V}_l + \Delta_{ij} \right) \\
&= \left\{ \ell'(\varepsilon_i) + \ell''(\varepsilon_i) \sum_{l \neq k} L_{n,i}(I - \mathbf{B}\mathbf{B}^T)\boldsymbol{\beta}_{0l}\mathbf{V}_l + \ell''(\varepsilon_i) \left\{ \Delta_{ij}(\mathbf{B}_0) - \varepsilon_i \right\} \right\} (1 + o_p(1)),
\end{aligned}$$

so, we have

$$
\begin{aligned}
&-(I-\mathbf{B}\mathbf{B}^T)\sum_{l=1}^{D}\boldsymbol{\beta}_{0l}\left\{n^{-1}h^2\sum_{j=1}^{n}\mathbf{V}_k(\mathbf{B}_0^T\mathbf{X}_j)\mathbf{V}_l(\mathbf{B}_0^T\mathbf{X}_j)\mathrm{E}\{\ell''(\varepsilon)\}+O_p(h^3+h\delta_n)\right\}\\
&=n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n}\ell'(\varepsilon_i)K_{h,i}(\mathbf{X}_j)\mathbf{V}_k(\mathbf{B}_0^T\mathbf{X}_j)L_{n,i}^T(\mathbf{B},\mathbf{X}_j)/\varsigma_n(\mathbf{X}_j)\\
&\quad+n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{X}_j)\mathbf{V}_k(\mathbf{B}_0^T\mathbf{X}_j)L_{n,i}^T(\mathbf{B},\mathbf{X}_j)\{\Delta_{ij}(\mathbf{B}_0)-\varepsilon_i\}/\varsigma_n(\mathbf{X}_j)\\
&\triangleq C_1+C_2. \qquad (4.9)
\end{aligned}
$$

Let

$$N_n(\mathbf{B},\mathbf{x})=S_n^{-1}n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})(\mathbf{X}_i-\mathbf{x})^T,$$

then we have

$$S_n(\mathbf{B},\mathbf{x})=\begin{pmatrix} f(\mathbf{x}) & \mathbf{B}^T\nabla f(\mathbf{x})h \\ \mathbf{B}^T\nabla f(\mathbf{x})h & f(\mathbf{x})\mathbf{B}\mathbf{B}^T \end{pmatrix}\mathrm{E}\{\ell''(\varepsilon)\}+O_p(h^2+\delta_n)$$

and

$$N_n(\mathbf{B},\mathbf{x})=\begin{pmatrix} f^{-1}(\mathbf{x})\nabla^T f(\mathbf{x})(I-\mathbf{B}\mathbf{B}^T)h^2 \\ \mathbf{B}^T h \end{pmatrix}+O_p(h^3+h\delta_n).$$

Therefore,

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})\{L_{n,i}(\mathbf{B},\mathbf{x})\}^T\Xi_{n,i}(\mathbf{B},\mathbf{x}) \\
=& n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})S_n^{-1}n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\ell'(\varepsilon_i) \\
& - N_n^T(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})S_n^{-1}n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\ell'(\varepsilon_i) \\
=& N_n^T(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\ell'(\varepsilon_i) - N_n^T(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\ell'(\varepsilon_i)(1+O_p(a_n)) \\
=& O_p(h^3\delta_n + h\delta_n^2),
\end{aligned}
$$

and

$$
n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{X}_j)\mathbf{V}_k(\mathbf{B}_0^T\mathbf{X}_j)L_{n,i}^T(\mathbf{B},\mathbf{X}_j)\Xi_{n,i}(\mathbf{B},\mathbf{x})/\varsigma_n(\mathbf{X}_j) = O_p(h^3\delta_n + h\delta_n^2). \tag{4.10}
$$

Note that

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})\{L_{n,i}(\mathbf{B},\mathbf{x})\}^T[P_{2,h,i}(\mathbf{x}) + P_{3,h,i}(\mathbf{x})h - Q_{n,i}(\mathbf{B},\mathbf{x})] \\
=& n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})[P_{2,h,i}(\mathbf{x}) + P_{3,h,i}(\mathbf{x})h - Q_{n,i}(\mathbf{B},\mathbf{x})] \\
& - N_n^T(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})[P_{2,h,i}(\mathbf{x}) + P_{3,h,i}(\mathbf{x})h - Q_{n,i}(\mathbf{B},\mathbf{x})] \\
=& n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})[P_{2,h,i}(\mathbf{x}) + P_{3,h,i}(\mathbf{x})h] \\
& - N_n^T(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})[P_{2,h,i}(\mathbf{x}) + P_{3,h,i}(\mathbf{x})h],
\end{aligned}
$$

Since the expectation of the right-hand side can be calculated by taking expectation with respect to ε_i at first, which gives $\mathrm{E}\{\ell''(\varepsilon_i)\}$, following the results in Xia et al (2002), we have

$$
\begin{aligned}
& n^{-2}h^2\sum_{j=1}^{n}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{X}_j)\mathbf{V}_k(\mathbf{B}_0^T\mathbf{X}_j)L_{n,i}^T(\mathbf{B},\mathbf{X}_j)[P_{2,h,i}(\mathbf{x})+P_{3,h,i}(\mathbf{x})h-Q_{n,i}(\mathbf{B},\mathbf{x})]/\varsigma_n(\mathbf{X}_j) \\
=&(I-\mathbf{B}\mathbf{B}^T)\mathbf{B}_0 n^{-1}\sum_{j=1}^{n}f^{-1}(\mathbf{X}_j)\left\{\bar{\mathbf{V}}(\mathbf{X}_j)-\frac{1}{2}\sum_{l=1}^{D}\mathbf{V}_{l,l}^2(\mathbf{B}^T\mathbf{X}_j)\mathbf{B}_0\nabla f(\mathbf{x})\right\}\mathrm{E}\{\ell''(\varepsilon)\}h^4 \\
&+O_p(h^5+h^3\delta_n), \qquad (4.11)
\end{aligned}
$$

where $\bar{\mathbf{V}}(\mathbf{B}_0^T\mathbf{x}) = \tilde{\mathbf{V}}^2(\mathbf{B}_0^T\mathbf{x})\mathbf{B}_0^T\nabla f(\mathbf{x}) + \tilde{\mathbf{V}}^3(\mathbf{B}_0^T\mathbf{x})$ with $\tilde{\mathbf{V}}^2(\mathbf{B}_0^T\mathbf{x})$ being a $D\times D$ matrix of the upper left part of $\sum_{m,l=1}^{D}\left\{\mathbf{V}_{m,l}^2(\mathbf{B}_0^T\mathbf{x})\times\int K(U)UU^Tu_lu_mdU\right\}$, $\kappa_4 = \int u^4K(u)du$ and

$$
\mathbf{B}_0\tilde{\mathbf{V}}^3(\mathbf{B}_0^T\mathbf{x}) = \frac{1}{6}\left\{\sum_{l=1}^{D}\mathbf{V}_{l,l,l}^3(\mathbf{B}_0^T\mathbf{x})\kappa_4\boldsymbol{\beta}_{0l}+\sum_{m\neq l}\mathbf{V}_{m,m,l}^3(\mathbf{B}_0^T\mathbf{x})\boldsymbol{\beta}_{0l}\right\}.
$$

Similarly we have

$$
\begin{aligned}
& n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})L_{n,i}^T(\mathbf{B},\mathbf{x})[R_i(\mathbf{x})-R_{n,i}(\mathbf{B},\mathbf{x})] \\
=&n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})R_i(\mathbf{x})-N_n^T(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})R_i(\mathbf{x}) \\
=&O_p(h^5),
\end{aligned}
$$

and

$$n^{-2}\sum_{j=1}^{n}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{X}_j)\mathbf{V}_k(\mathbf{B}_0^T\mathbf{X}_j)L_{n,i}^T(\mathbf{B},\mathbf{X}_j)[R_i(\mathbf{x})-R_{n,i}(\mathbf{B},\mathbf{x})]/\varsigma_n(\mathbf{X}_j)$$

$$=O_p(h^5). \tag{4.12}$$

Therefore, from (4.10), (4.11) and (4.12) we have

$$C_2 =(I-\mathbf{B}\mathbf{B}^T)\mathbf{B}_0 n^{-1}\sum_{j=1}^{n} f^{-1}(\mathbf{X}_j)\left\{\bar{\mathbf{V}}(\mathbf{X}_j)-\frac{1}{2}\sum_{l=1}^{D}\mathbf{V}_{l,l}^2(\mathbf{B}^T\mathbf{X}_j)\mathbf{B}_0\nabla f(\mathbf{x})\right\}\mathrm{E}\{\ell''(\varepsilon)\}h^4$$

$$+O_p(h^5+h^3\delta_n+h\delta_n^2). \tag{4.13}$$

Next we will check the order of C_1. Note that,

$$\begin{aligned}
&n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})L_{n,i}^T(\mathbf{B},\mathbf{x})\ell'(\varepsilon_i)\\
=&n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\ell'(\varepsilon_i)\\
&-n^{-1}\sum_{i=1}^{n}\ell''(\varepsilon_i)K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\mathbf{X}_{h,i}^T(\mathbf{B},\mathbf{x})S_n^{-1}n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})\ell'(\varepsilon_i)\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\\
=&n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\ell'(\varepsilon_i)-N_n^T(\mathbf{B},\mathbf{x})n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})\mathbf{X}_{h,i}(\mathbf{B},\mathbf{x})\ell'(\varepsilon_i).\\
=&n^{-1}\sum_{i=1}^{n}K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\ell'(\varepsilon_i)-n^{-1}f^{-1}(\mathbf{x})(I-\mathbf{B}\mathbf{B}^T)h^2\nabla f(\mathbf{x})\sum_{i=1}^{n}K_{h,i}(\mathbf{x})\ell'(\varepsilon_i)\\
&-n^{-1}\mathbf{B}\mathbf{B}^T\sum_{i=1}^{n}K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\ell'(\varepsilon_i)+O_p(h^3\delta_n+h\delta_n^2)\\
=&(I-\mathbf{B}\mathbf{B}^T)n^{-1}\left\{\sum_{i=1}^{n}K_{h,i}(\mathbf{x})(\mathbf{X}_i-\mathbf{x})\ell'(\varepsilon_i)-\frac{h^2}{f(\mathbf{x})}\Delta f(\mathbf{x})\sum_{i=1}^{n}K_{h,i}(\mathbf{x})\ell'(\varepsilon_i)\right\}+O_p(h^3\delta_n+h\delta_n^2),
\end{aligned}$$

since $\mathrm{E}\{\ell'(\varepsilon)\} = 0$, we have

$$C_1 = n^{-2} \sum_{j=1}^{n} \sum_{i=1}^{n} K_{h,i}(\mathbf{X}_j) \mathbf{V}_k(\mathbf{B}_0^T \mathbf{X}_j) L_{n,i}^T(\mathbf{B}, \mathbf{X}_j) \ell'(\varepsilon_i) / \varsigma_n(\mathbf{X}_j) = O_p(\delta_n h^3 + h\delta_n^2). \tag{4.14}$$

Therefore, from (4.13) and (4.14),

$$\begin{aligned}
&- (I - \mathbf{B}\mathbf{B}^T) \sum_{l=1}^{D} \beta_{0l} \left\{ n^{-1} h^2 \sum_{j=1}^{n} \mathbf{V}_k(\mathbf{B}_0^T \mathbf{X}_j) \mathbf{V}_l(\mathbf{B}_0^T \mathbf{X}_j) \mathrm{E}\{\ell''(\varepsilon)\} + O_p(h^3 + h\delta_n) \right\} \\
=&(I - \mathbf{B}\mathbf{B}^T) \mathbf{B}_0 n^{-1} \sum_{j=1}^{n} f^{-1}(\mathbf{X}_j) \left\{ \bar{\mathbf{V}}(\mathbf{X}_j) - \frac{1}{2} \sum_{l=1}^{D} \mathbf{V}_{l,l}^2(\mathbf{B}^T \mathbf{X}_j) \mathbf{B}_0 \Delta f(\mathbf{x}) \right\} \mathrm{E}\{\ell''(\varepsilon)\} h^4 \\
&+ O_p(h^5 + h^3 \delta_n + h\delta_n^2),
\end{aligned}$$

for $k = 1, \ldots, D$. Hence,

$$(I - \mathbf{B}\mathbf{B}^T) \mathbf{B}_0 n^{-1} \sum_{j=1}^{n} \mathbf{V}(\mathbf{B}_0^T \mathbf{X}_j) \mathbf{V}^T(\mathbf{B}_0^T \mathbf{X}_j) \mathrm{E}\{\ell''(\varepsilon)\} = O_p(h^3 + h\delta + h^{-1}\delta_n^2).$$

Since $n^{-1} \sum_{j=1}^{n} \mathbf{V}(\mathbf{B}_0^T \mathbf{X}_j) \mathbf{V}^T(\mathbf{B}_0^T \mathbf{X}_j) = O_p(1)$, we have

$$||(I - \mathbf{B}\mathbf{B}^T) \mathbf{B}_0|| = O_p(h^3 + h\delta + h^{-1}\delta_n^2).$$

C. Proof of Theorem 2.2

Note that

$$
\begin{aligned}
&\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)}) \\
=&\sum_{j=1}^{n}\sum_{i=1}^{n} \log\left\{ \frac{\sum_{l=1}^{n} K_{h_1}\left[y_i - \left\{a_j^{(k+1)} + \mathbf{b}_j^{(k+1)^T}\mathbf{B}^{(k+1)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]}{\sum_{l=1}^{n} K_{h_1}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]} \right\} w_{ij} \\
=&\sum_{j=1}^{n}\sum_{i=1}^{n} \log\left\{ \sum_{l=1}^{n} \left(\frac{K_{h_1}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]}{\sum_{l=1}^{n} K_{h_1}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]} \right.\right. \\
&\left.\left. \times \frac{K_{h_1}\left[y_i - \left\{a_j^{(k+1)} + \mathbf{b}_j^{(k+1)^T}\mathbf{B}^{(k+1)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]}{K_{h_1}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]} \right) \right\} w_{ij} \\
=&\sum_{j=1}^{n}\sum_{i=1}^{n} \log\left\{ \sum_{l=1}^{n} p_{ijl}^{(k+1)} \frac{K_{h_1}\left[y_i - \left\{a_j^{(k+1)} + \mathbf{b}_j^{(k+1)^T}\mathbf{B}^{(k+1)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]}{K_{h_1}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]} \right\} w_{ij},
\end{aligned}
$$

where

$$
p_{ijl}^{(k+1)} = \frac{K_{h_1}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]}{\sum_{l=1}^{n} K_{h_1}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]}.
$$

From the Jensen's inequality, we have

$$
\begin{aligned}
&\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)}) \\
\geq&\sum_{j=1}^{n}\sum_{i=1}^{n} \left[\sum_{l=1}^{n} p_{ijl}^{(k+1)} \log\left\{ \frac{K_{h_1}\left[y_i - \left\{a_j^{(k+1)} + \mathbf{b}_j^{(k+1)^T}\mathbf{B}^{(k+1)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]}{K_{h_1}\left[y_i - \left\{a_j^{(k)} + \mathbf{b}_j^{(k)^T}\mathbf{B}^{(k)^T}(\mathbf{X}_i - \mathbf{X}_j)\right\} - \tilde{\epsilon}_l\right]} \right\} w_{ij} \right]
\end{aligned}
$$

Based on the property of M-step of (2.5), we have $\ell(\boldsymbol{\theta}^{(k+1)}) - \ell(\boldsymbol{\theta}^{(k)}) \geq 0$.

References

Amato, U., Antoniadis, A. and De Feis, I. (2006). Dimension reduction in functional regression with applications. *Computational Statistics and Data Analysis*, 50, 2422–2446.

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, 19, 563–570.

Beran, R. (1974). Asymptotically efficient adaptive rank estimates in location models. *Annals of Statistics*, 2, 63–74.

Bickel, P. J. (1982). On adaptive estimation. *Annals of Statistics*, 10, 647–671.

Chaudhuri, P., Hang, M. C., Loh, W. Y. and Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica*, 4, 143–167.

Čížek, P. and Härdle, W. (2006). Robust estimation of dimension reduction space. *Computational Statistics and Data Analysis*, 51, 545–555.

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

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91, 983–992.

Cook, R. D. (1998). *Regression Graphics: Ideas for studying regressions through graphics*. New York: Wiley.

Cook, R. D. and Li, B. (2002). Dimension reduction for the conditional mean in regression. *Annals of Statistics*, 30, 455–474.

Cook, R. D. and Weisberg, S. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, 86, 328–332.

Drost, F. and Klaassen, C. A. J. (1997). Efficient estimation in semiparametric GRACH models. *Journal of Econometrics*, 81, 193–221.

Hodgson, D. (1998). Adaptive estimation of cointegrating regression with ARMA errors. *Journal of Econometrics*, 85, 231–268.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Annals of Statistics*, 33, 1580–1616.

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

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.

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

Li, K. C., Lue, H. H. and Chen, C. H. (2000). Interactive tree-truncated regression via principal Hessian directions. *Journal of the American Statistical Association*, 95, 547–560.

Linton, O. and Xiao, Z. (2007). A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory*, 23, 371–413.

Manski, C. F. (1984). Adaptive estimation of nonlinear regression models. *Econometric Review*, 3, 187–208.

Raykar, V. C. and Duraiswami, R. (2006). Fast optimal bandwidth selection for kernel density estimation. In proceedings of the sixth *SIAM International Conference on Data Mining*, Bethesda, April 2006, 524–528.

Schick, A. (1993). On efficient estimation in regression models. *Annals of Statistics*, 21, 1486–1521.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of Royal Statistical Society*, B, 53, 683–690.

Steigerwald, D. (1992). Adaptive estimation in time series regression models. *Journal of Econometrics*, 54, 251–276.

Stone, C. (1975). Adaptive maximum likelihood estimation of a location parameters. *Annals of Statistics*, 3, 267–284.

Wang, H. and Xia, Y. (2008). Sliced Regression for Dimension Reduction. *Journal of the American Statistical Association*, 103, 811–821.

Wang, Q. and Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics and Data Analysis*, 52, 4512–4520.

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *Journal of Multivariate Analysis*, 97, 1162-1184.

Xia, Y., Tong, H., Li, W. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *Journal of Royal Statistical Society*, B, 64, 363–410.

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.

Yuan, A. and De Gooijer, J. G. (2007). Semiparametric regression with kernel error model. *Scandinavian Journal of Statistics*, 34, 841–869.

Yuan, A. (2010). Semiparametric inference with kernel likelihood. *Journal of Nonparametric Statistics*, 21, 207–228.