DIMENSION REDUCTION WITHOUT A MODEL VIA MODELLING METHODS

Xiangrong Yin

Department of Statistics
204 Statistics Building
University of Georgia
Athens, GA 30602

05-18-2005

1 Introduction

Some notation:

The notation $U \perp\!\!\!\perp V|\mathbf{Z}$ means that the random vectors U and V are independent given any value for the random vector $\mathbf{Z}$.

Subspaces will be denoted by $\mathcal{S}$, and $\mathcal{S}(\mathbf{B})$ means the subspace of $\mathcal{R}^t$ spanned by the columns of $t \times u$ matrix $\mathbf{B}$.

$P_{\mathbf{B}}$ denotes the projection operator for $\mathcal{S}(\mathbf{B})$ with respect to the usual inner product and $Q_{\mathbf{B}} = I - P_{\mathbf{B}}$.

The most common goal of a regression is to infer about the conditional mean of $Y|\mathbf{X}$. That is, traditionally $Y = E(Y|\mathbf{X}) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is independent of $\mathbf{X}$.

In dimension reduction, we mainly consider regressions in which $E(Y|\mathbf{X})$ depends on $\mathbf{X}$ through $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ of the predictors. Simply,

$$E(Y|\mathbf{X}) = E(Y|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) = g(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$$

More precisely, we want to identify the central mean subspace defined by Cook and Li (2002). That is, we assume that for some $p \times q$ matrix $\boldsymbol{\beta}$

$$Y \perp\!\!\!\perp E(Y|\mathbf{X})|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}. \tag{1}$$

The subspace spanned by the columns of $\boldsymbol{\beta}$ is a mean dimension reduction subspace. If the intersection of all the mean dimension reduction subspace is itself a mean dimension reduction subspace, then it is a central mean subspace (Cook and Li 2002), denoted by $\mathcal{S}_{E(Y|\mathbf{X})}$.

We assume that the central mean subspace exists, and our goal is to identify it.

Some short review:

Single-index conditional mean: $q = 1$. Härdle & Stoker (1989) developed the average derivative estimation (ADE). Others: McCullagh & Nelder, (1989). (Friedman & Stuetzle, 1981; Hall, 1989; Härdle et al., 1993; Hristache et al., 2001).

Xia et al. (2002) considered local polynomials for estimating $\boldsymbol{\beta}$ with $q > 1$.

Ordinary least squares (OLS) (LI and Duan 1989, Cook and Li 2002), under the condition that $E(\mathbf{X}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ is a linear function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$.

Principal Hessian directions (pHd,Li, 1992), under constant variance of var$(\mathbf{X}|\boldsymbol{\beta}^{T}\mathbf{X})$, and Fourth moments methods (FM, Yin and Cook, 2004), under symmetric condition of $\mathbf{X}|\boldsymbol{\beta}^{T}\mathrm{T}\mathbf{X}$.

We shall consider two scenarios on how the data $(Y_i, \mathbf{X}_i^T)$, $i = 1, \ldots, n$, are generated.

One is *random design model*, that is, the data $(Y_i, \mathbf{X}_i^T)$, $i = 1, \ldots, n$, are iid observations on $(Y, \mathbf{X}^T)$, which has a joint distribution.

The other is *fixed design model*, that is, the $\mathbf{X}$ variables are nonstochastic, often assumed to be equidistributed on a bounded interval. And without loss of generality, it can be assumed on $[0, 1]^p$. See Härdle (1990, p. 21) and Eubank (1999).

However, the basic ideas for them are the same. We mostly deal with the random design, sometimes we present the idea in fixed design to simplify the discussion.

2 Motivation

To simplify our discussion, let $q = 1$ so that $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ is a $p \times 1$ vector, and $\boldsymbol{\beta}_0^T \boldsymbol{\beta}_0 = 1$ for its identifiability.

And we also assume $\mathbf{X}$ has fixed design. Suppose that $E(Y|\mathbf{X}) = g(\boldsymbol{\beta}_0^T \mathbf{X})$ with known $\boldsymbol{\beta}_0$, and that $B_j(X)$ for $j = 1, ..., J$ are basis functions which together with J will be chosen later. Suppose that we can write the model as the following

$$y_i = g(\boldsymbol{\beta}_0^T \mathbf{X}_i) + \boldsymbol{\varepsilon} = \sum_{j=1}^{\infty} \theta_j B_j(\boldsymbol{\beta}_0^T \mathbf{X}_i) + \boldsymbol{\varepsilon}. \qquad (2)$$

Where g is an unknown function, and $E(\epsilon|\mathbf{X}) = 0$. The last condition allows ϵ to be dependent on $\mathbf{X}$.

Model (2) hold under many circumstances (Eubank 1999; Fan and Gijbels 1996). For example, if $g \in C[0, 1]$, a continuous function, then one can use orthogonal polynomial function for B_j; if $g \in L^2[0, 1]$, a square integrable function, then one can use Fourier functions for B_j.

The direction $\boldsymbol{\beta}_0$ is then the solution of

$$\min_{\boldsymbol{\beta}}\{E[y - E(y|\boldsymbol{\beta}^T\mathbf{X})]^2\}.$$

In modeling the mean function, typically let $S_J(\boldsymbol{\beta}_0^T\mathbf{X}) = \sum_{j=1}^J \theta_j B_j(\boldsymbol{\beta}_0^T\mathbf{X})$, then in practice we find an estimator of g by minimizing the following term:

$$RSS_J(\boldsymbol{\beta}_0, \boldsymbol{\theta}) = \sum_{i=1}^{n}[y_i - \sum_{j=1}^{J}\theta_j B_j(\boldsymbol{\beta}_0^T\mathbf{X}_i)]^2 \qquad (3)$$

The minimization in equation (3) is actually over J and $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_J)^T$. For a fixed J, we solve the normal equation, and find the estimates of $\boldsymbol{\theta}$. Once found, we put them back in $S_J(\boldsymbol{\beta}_0^T\mathbf{X})$ for an estimate. If J is not known, there are many ways to choose the best J in estimating g (see Eubank 1999).

More specifically, suppose that $\boldsymbol{\beta}$ and J are temporarily fixed, then define a $n \times J$ matrix

$$\mathbf{B}_J(\boldsymbol{\beta}) = (B_j(\boldsymbol{\beta}^T\mathbf{X}_i))_{i=1,...,n;j=1,...,J}$$

Further suppose that $\mathbf{B}_J(\boldsymbol{\beta})$ has full rank, then

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_J)^T = (\mathbf{B}_J(\boldsymbol{\beta})^T\mathbf{B}_J(\boldsymbol{\beta}))^{-1}\mathbf{B}_J(\boldsymbol{\beta})^T\mathbf{y},$$

where $\mathbf{y} = (y_1, ..., y_n)^T$ is the vector of response values. Thus putting this $\boldsymbol{\theta}$ back into $RSS_J(\boldsymbol{\beta}, \boldsymbol{\theta})$, that is,

$$RSS_J(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i - g_J(\boldsymbol{\beta}^T\mathbf{x}_i)]^2, \qquad (4)$$

where $g_J(\boldsymbol{\beta}^T\mathbf{x}) = (B_1(\boldsymbol{\beta}^T\mathbf{x}), ..., B_J(\boldsymbol{\beta}^T\mathbf{x}))(\mathbf{B}_J(\boldsymbol{\beta})^T\mathbf{B}_J(\boldsymbol{\beta}))^{-1}\mathbf{B}_J(\boldsymbol{\beta})^T\mathbf{y}$.

1. To adapt this in dimension reduction, we find $\boldsymbol{\beta}$ by minimizing $RSS_J(\boldsymbol{\beta})$ in equation (4) over $\boldsymbol{\beta}$ if j is fixed, or both.

2. The estimated mean function $g_J(\boldsymbol{\beta}^T)$ serves as only intermediate step, and choosing a fixed J is not so critical.

3. Note that this g_J is a special case of a *linear estimator* (Eubank, 1999, p. 12). Therefore, any linear estimator can be used this way for dimension reduction.

3 Methods

A linear estimator is defined as

$$\hat{S}_J(\boldsymbol{\beta}^T\mathbf{x}) = n^{-1}\sum_{i=1}^{n} W_{Ji}(\boldsymbol{\beta}^T\mathbf{x})Y_i \tag{5}$$

where $W_{Ji}(\mathbf{x})$ is a weight function. For random design, usually

$$W_{Ji}(\mathbf{x}) = \frac{K_J(\mathbf{x}, \mathbf{x}_i)}{n^{-1}\sum_{i=1}^{n} K_J(\mathbf{x}, \mathbf{x}_i)}, \tag{6}$$

while for fixed design (eg. H$\ddot{a}$rdle 1990),

$$W_{Ji}(\mathbf{x}) = K_J(\mathbf{x}, \mathbf{x}_i). \tag{7}$$

From now on we present schemes for random design unless otherwise stated. Similar ideas for fixed design can be developed. Once a weight function is chosen, and having set $\hat{S}_J(\boldsymbol{\beta}^T\mathbf{x})$, assume that J is fixed, we then find an estimate of $\boldsymbol{\beta}_0$ by minimizing the following over $\boldsymbol{\beta}$:

$$RSS(J, \boldsymbol{\beta}) = \sum_{i=1}^{n}\{y_i - \hat{S}_J(\boldsymbol{\beta}^T\mathbf{x}_i)\}^2 \tag{8}$$

Note that $\hat{S}_J(\boldsymbol{\beta}^T\mathbf{x}_i) = a_i^T\mathbf{y}$ for $i = 1, ..., n$, and define $n \times n$ matrix, $\mathcal{S}_J^T = (a_1, ..., a_n)$. If J is not fixed, we then can simultaneously choose J and $\boldsymbol{\beta}$ by GCV method to minimize the following:

$$GCV(J, \boldsymbol{\beta}) = \frac{n^{-1}RSS(J, \boldsymbol{\beta})}{(n^{-1}\text{trace}[I - \mathcal{S}_J])^2}$$

Choices of Basis Functions:

Polynomial basis. Polynomial regression has at least two motivations: Taylor's theorem (on Sobolev space $W_2^m[0,1]$, Eubank 1999) and Weierstrass approximation (on continuous function $C[0,1]$). Here we use H$\ddot{a}$rdle's (1984) set up:

$$K_J(\mathbf{x}, \mathbf{x}_i) = \sum_{j=0}^{J} B_j(\boldsymbol{\beta}^T \mathbf{x}) B_j(\boldsymbol{\beta}^T \mathbf{x}_i),$$

where $B_j(\boldsymbol{\beta}^T \mathbf{X})$ is Legendre polynomials (H$\ddot{a}$rdle 1990, p. 52).

Links to some existing methods. If $J = 1$, then it is OLS method. If $J = 2$, then it is q-based pHd (Li 1992).

Fourier series. Any other orthogonal series function, particularly *complete orthonormal series* (CONS) (Eubank 1999). Sine Cosine and Fourier series.

Kernel method.

Wavelet basis. One can follow from Antoniadis et al (1990).

Spline method.

4 Kernel representation and Weighted Least Squares

All previous methods linked to one method: kernel method. Generally, the estimated mean function can be written as linear estimator: $\hat{g}(\boldsymbol{\beta}^T\mathbf{X}) = \sum_{j=1}^n W_{Ji}(\boldsymbol{\beta}^T\mathbf{x})y_j$, with $\sum_{i=1}^n W_{Ji}(\boldsymbol{\beta}^T\mathbf{x}) = 1$. However, this is the the solution for weighted least squares. That is,

$$\min_{\eta} n^{-1}\sum_{i=1}^n W_{Ji}(\boldsymbol{\beta}^T\mathbf{x})(Y_i - \eta)^2 = n^{-1}\sum_{i=1}^n W_{Ji}(\boldsymbol{\beta}^T\mathbf{x})(Y_i - \hat{g}(\boldsymbol{\beta}^T\mathbf{x}))^2.$$

This observation is very interesting. For example OLS is a dimension in the CMS if the OLS model is right or the predictors are elliptically distributed. If not, one can use reweighting method (Cook and Nachtsheim 1994) to achieve elliptical distribution.

Xia et al. (2002) developed method without requiring particular distribution on predictors by using local linear approximation.

Here we used (special chosen) weight functions to rewight without requiring any particular distribution on predictors via weight least squares to recover CMS.

5 Multiple dimensional search

Having seen the general procedure for single index case, we now extend it to multiple-index. We simply replace $p \times 1$ vector $\boldsymbol{\beta}$ by $p \times q$ matrix $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_q)$ under $\boldsymbol{\beta}^T \boldsymbol{\beta} = I_q$. And multi-dimensional kernels then can be used in place of univariate kernel.

In our application, we use product kernel for multiple dimensional search for its simplicity. That is, we have

$$K_J(\boldsymbol{\beta}^T \mathbf{x}, \boldsymbol{\beta}_i^{\mathbf{x}}) = \prod_{j=1}^{q} K_J(\boldsymbol{\beta}_j^T \mathbf{x}, \boldsymbol{\beta}_j^T \mathbf{x}_i)$$

6 Algorithm and practical issues

1. Epanechnikov kernel: bandwidth $h_n = A(q)s_{\boldsymbol{\beta}}/n^{1/(q+4)}$, where $A(q)$ is based on Silverman (1986, p.87) and $s_{\boldsymbol{\beta}}$ is the standard deviation of $t = \boldsymbol{\beta}^T\mathbf{X}$.
2. Gaussian kernel: bandwidth, $h_n = A(q)s_{\boldsymbol{\beta}}/n^{1/(q+4)}$ where $A(q)$ is based on Scott (1992, p. 152) and $s_{\boldsymbol{\beta}}$ is the standard deviation of $t = \boldsymbol{\beta}^T\mathbf{X}$.
3. Polynomial basis: $J = 15$, that is, $14th$ order polynomial of Legendre system. Data has been transformed so that $\boldsymbol{\beta}^T\mathbf{X} = t \in [-1, 1]$, we use the following choice: $B_0(t) = 1/\sqrt{2}, B_1(t) = t/\sqrt{2/3}$, and moreover, $(m + 1)B_{m+1}(t) = (2m + 1)tB_m(t) - mB_{m-1}(t)$.
4. Cosine series: $J = 12$. Data has been transformed so that $\boldsymbol{\beta}^T\mathbf{X} = t \in [-1, 1]$, we use the following choice: $B_0(t) = 1/\sqrt{2}, B_j(t) = \cos(j\pi t)$, for $j = 1, ..., J$.
5. Fourier basis: $J = 12$. Data has been transformed so that $\boldsymbol{\beta}^T\mathbf{X} = t \in [-1, 1]$, we use the following choice: $B_0(t) = 1/\sqrt{2}, B_{cj}(t) = \cos(j\pi t), B_{sj} = \sin(j\pi t)$ for $j = 1, ..., J$.

We use *matlab* in our code.

7 Consistency and Asymptotics

We assume that $\mathbf{X} \in [0, 1/\sqrt{p}]^p$ without loss of generality. Also for the ease of our exposition and in common with most investigation of this type (See Hall 1989, Härdle et al. 1993), we confine our attention to one-dimensional case: $d = 1$.

Due to the constraint of $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$, $t = \boldsymbol{\beta}^T\mathbf{x} \in [0, 1]$. Write $g_{nh}(t)$ (instead of $\hat{S}_J(t)$ in equation 5)) to be the estimated mean function of $g(t)$, where h is the bandwidth relating to n. Note that under the assumption of the existence of the central mean subspace, $\boldsymbol{\beta}_0$ is unique. We then have the following result.

Theorem 1 *Suppose that $E|Y| < \infty$ and $E(|g(\boldsymbol{\beta}^T\mathbf{X})| < \infty$ for any $\boldsymbol{\beta}$. In addition, $g(t)$ is uniformly continuous in $t \in [0, 1]$ and $\boldsymbol{\beta}$, where $t = \boldsymbol{\beta}^T\mathbf{x}$. If with probability 1, as $n \to \infty$, $\sup_{\mathbf{x}\in[0,1]^p,\boldsymbol{\beta}} |g_{nh}(\boldsymbol{\beta}^T\mathbf{x}) - g(\boldsymbol{\beta}^T\mathbf{x})| \to 0$, then with probability 1, $\boldsymbol{\beta}_{nh} \to \boldsymbol{\beta}_0$ as $n \to \infty$.*

Lemma 1 *Under some regularity conditions, with probability one, $\sup_{\mathbf{x}\in[0,1]^p,\boldsymbol{\beta}} |g_{nh}(\boldsymbol{\beta}^T\mathbf{x}) - g(\boldsymbol{\beta}^T\mathbf{x})| \to 0$.*

Theorem 2 *Under certain regularity conditions, then there is a constant C such that for any positive ϵ and for n sufficiently large, $P[|\boldsymbol{\beta}_{nh} - \boldsymbol{\beta}_0| > \epsilon] \leq C/(nh^4\epsilon^2)$.*

8 Estimating the dimensionality of CMS

There are are several ways to estimate the dimension d of the CMS, such as bootstrap method used by Ye and Weiss (2003) or Cross-validation method used by Xia et al. (2002).

For bootstrap method we can follow exactly the idea by Ye and Weiss (2003) to calculate the distances between directions or subspaces.

Cross-validation method: Let

$$\hat{a}_{d,j} = \sum_{i=1,i\neq j}^{n} K_{h_d}(i,j)y_i / \sum_{i=1,i\neq j}^{n} K_{h_d}(i,j),$$

where $K_{h_d}(i,j) = K_{h_d}\{\hat{\boldsymbol{\beta}}_1^T(\mathbf{X}_i - \mathbf{X}_j), ..., \hat{\boldsymbol{\beta}}_d^T(\mathbf{X}_i - \mathbf{X}_j)\}$. Let

$$CV(d) = n^{-1}\sum_{j=1}^{n}(y_i - \hat{a}_{d,j})^2, d = 1, ..., p.$$

And define $CV(0) = n^{-1}\sum_{j=1}^{n}(y_i - \bar{y})^2$.

We choose d_0 to be the first valley of $CV(d)$ among all $CV(d)$.

9 Projection pursuit regression type

Our method is similar to projection pursuit regression (Friedman and Stuetzle, 1981; Huber 1985), though our motivation is quite different. The main goal of PPR is to estimate the regression function. While our goal is to reduce the dimension, the regression function serves an intermediate step. This difference may put advantages in our method as we shall see in our examples later.

When kernel is used, our approach is very much like the ones used by Hall (1989) and Härdle et al. (1993) for first projection pursuit approximation. Here we may think that our method is a repeatedly one step projection pursuit regression with multiple dimensions.

Related to the results from Hall (1989) and Härdle et al. (1993), better rates of convergence for the estimated direction can be obtained for our scheme by using leave one out procedure.

To further explore the difference between the usual projection pursuit regression and our approach, we next develop corresponding methods for other subspaces.

10 Other subspaces and Inverse method links

Central kth moment subspace and CS:

Our method also can be used to find CKMS (Yin and Cook 2002) and CS (Cook, 1994a,b, 1996). Assuming that conditional moment generating function exists, then

$$E(e^{ty}|\mathbf{X}) = \int e^{ty} p(y|\mathbf{x})dy = \sum_{k=0}^{\infty} \frac{t^k}{k!} E(Y^k|\mathbf{X})$$

Thus with $g_J^k(\boldsymbol{\beta}^T\mathbf{x}) = n^{-1}\sum_{j=1}^{n} W_{nj}(\boldsymbol{\beta}^T\mathbf{x})y_j^k$, and under $\boldsymbol{\beta}\boldsymbol{\beta} = I_q$, we find each moment function by minimizing (for fixed k)

$$RSS(k, \boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i^k - g_J(\boldsymbol{\beta}^T\mathbf{x}_i))^2 \qquad (9)$$

Using CV method to obtain the estimate $\boldsymbol{\beta}_k$ for each $k = 1, ...K$, and finally using SVD method to the following matrix

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \boldsymbol{\beta}_k \boldsymbol{\beta}_k^T$$

to find the non zero eigenvalues and their corresponding eigenvectors. Practically, we choose $K = 2$ for the most important moments in regressions are the first twos. If all the $\boldsymbol{\beta}_k$'s are the same, then a final estimate can be obtained by minimizing $\sum_{k=1}^{K} RSS(k, \boldsymbol{\beta})$.

Partial CMS and Partial CS:

When there is a categorical variable W for $W = 1, ..., C$, a more appropriate subspace perhaps is *Partial CMS* (Li, Cook and Chiaromonte, 2002). Our method also can be further developed to aim at this subspace. In such a case, simply we modify equation (4) to be the following:

$$RSS(w, \boldsymbol{\beta}) = \sum_{i=1}^{n_w} (y_i - g_J(\boldsymbol{\beta}^T \mathbf{x}_i))^2 \tag{10}$$

For each w we find the best $\boldsymbol{\beta}_w$ by CV method, and finally using SVD method to the following matrix

$$\boldsymbol{\Sigma} = \sum_{w=1}^{C} \boldsymbol{\beta}_w \boldsymbol{\beta}_w^T$$

to find the non zero eigenvalues and their corresponding eigenvectors. If all the $\boldsymbol{\beta}_w$'s are the same, then a final estimate can be obtained by minimizing $\sum_{w=1}^{C} RSS(w, \boldsymbol{\beta})$.

Combine this ideas with the procedure in Section 10, we also can find dimensions in the partial CKMS and partial CS.

Inverse method:

A dual of our method is the inverse method. Under the linearity conditions, suppose $\mathbf{X}$ is already standardized with 0 mean and variance I. With $g_J^{\boldsymbol{\beta}}(y) = n^{-1} \sum_{j=1}^{n} W_{nj}(y) \boldsymbol{\beta}^T \mathbf{X}_j$, and we switch the role of y and $\boldsymbol{\beta}^T \mathbf{x}$ in equation 4. Which is equivalent to the kernel method by Zhu and Fang (1996).

11 Examples

Let r_e, r_g, $r_p(J)$, $r_c(J)$ and $r_f(J)$ be the correlation coefficients between the true variable and the estimated variable, respectively by the corresponding methods. We use vector of elements with all 1's as initial vector unless otherwise stated with sample size $n = 200$.

Model 1: Quadratic model. Let $p = 10$, $\mathbf{X} \sim N_p(0, I)$, and $\epsilon \sim N(0, 1)$. The model is $Y = \boldsymbol{\beta}^T\mathbf{X} + (\boldsymbol{\beta}^T\mathbf{X})^2 + .5\epsilon$, where $\boldsymbol{\beta}^T = (1, 1, 0, ..., 0)$.

Model 2: Exponential model. This simulation is similar to the model used by Fan and Gijbels (1995). Let $p = 10$, $X_1 \sim U(-2, 2)$, and $X_2, ..., X_{10}, \epsilon$ are iid $N(0, 1)$. The model is $Y = \sin(2\boldsymbol{\beta}^T\mathbf{X}) + 2e^{-16(\boldsymbol{\beta}^T\mathbf{X})^2} + .3\epsilon$, where $\boldsymbol{\beta}^T = (1, 0, 0, ..., 0)$.

Model 3: Nonlinear regression. Let $p = 10$, $X_1, ..., X_{10}, \epsilon$ are iid $N(0, 1)$. The model is $Y = \frac{\boldsymbol{\beta}^T\mathbf{X}}{(.5+(\boldsymbol{\beta}^T\mathbf{X}+1.5)^2)} + .5\epsilon$, where $\boldsymbol{\beta}^T = (1, 0, 0, ..., 0)$.

Model 4: Non differentiable regression. Let $p = 10$, $X_1, ..., X_{10}$ are iid $U(0,1)$, and $\epsilon \sim N(0,1)$. With $\boldsymbol{\beta}^T = (1,0,0,...,0)$, the model is

$$\begin{aligned} Y = \quad & 8\boldsymbol{\beta}^T\mathbf{X} + .2\epsilon, \text{ if } \boldsymbol{\beta}^T\mathbf{X} \in (0,.25); \\ & 4 - 8\boldsymbol{\beta}^T\mathbf{X} + .2\epsilon, \text{ if } \boldsymbol{\beta}^T\mathbf{X} \in [.25,.5); \\ & .2\epsilon, \text{ if } \boldsymbol{\beta}^T\mathbf{X} \in [.5,.75); \\ & 4\boldsymbol{\beta}^T\mathbf{X} - 3 + .2\epsilon, \text{ if } \boldsymbol{\beta}^T\mathbf{X} \in (.75,1). \end{aligned}$$

Model 5: Discontinuous regression. Let $p = 10$, $X_2, ..., X_{10}, \epsilon$ are iid $N(0,1)$, and $\epsilon \sim U(0,1)$. With $\boldsymbol{\beta}^T = (1,0,0,...,0)$, the model is

$$\begin{aligned} Y = \quad & 8\boldsymbol{\beta}^T\mathbf{X} + .2\epsilon, \text{ if } \boldsymbol{\beta}^T\mathbf{X} \in (0,.25); \\ & .2\epsilon, \text{ if } \boldsymbol{\beta}^T\mathbf{X} \in [.75,1); \\ & 8 - 8\boldsymbol{\beta}^T\mathbf{X} + .2\epsilon, \text{ if } \boldsymbol{\beta}^T\mathbf{X} \in (.75,1). \end{aligned}$$

Table 1: Correlation coefficients between the true variable and estimated variable.

CCs	model 1	model 2	model 3	model 4	model 5
r_e	.9988	.9962	.9953	.9977*	.9871*
r_g	.9995	.9991	.9939	.9985	.9920*
$r_p(6)$	.9476	.9832	.9318	.9542	.9772
$r_p(10)$	.9795	.9943	.9481	.9896	.9854
$r_p(15)$	.9926	.9968	.9613	.9962	.9921
$r_c(8)$	.9068	.9258*	.9036*	.7690**	.9906
$r_c(10)$	.8922	.9702	.8906	.7596**	.9974*
$r_c(12)$	.8989	.9952*	.9246*	.9903**	.9982
$r_f(8)$	.9973	.9979	.9855	.9652	.9913
$r_f(10)$	.9996	.9989	.9755	.9922*	.9994*
$r_f(12)$	.9960	.9983*	.9928	.9991*	.9980*

*The numbers marked by * mean that initial vector of all 1's is not good for this method to have a proper solution, and it needs to be changed to random normal or random uniform initial vectors; The numbers marked by ** mean initial vector needs to be changed even closer to the true vector.*

Table 2: Bootstrap method with Legendre polynomial.

m_q, m_r	model 1	model 2	model 3	model 4	model 5
$d = 1$	.0109,.0109	.0105,.0105	.0236,.0236	.0060,.0060	.0198,.0198
$d = 2$	.4270,.1663	.3111,.1325	.4543,.1839	.4925,.1906	.5494,.2065

$m_q(d)$ and $m_r(d)$ are the means of $1 - q$ and $1 - r$ for $B = 200$ bootstrap samples using d-dimensional search, respectively.

Table 3: Cross validation with Legendre polynomial.

CV(d)	model 1	model 2	model 3	model 4	model 5
CV(0)	10.9591	.7803	.9971	.4333	.4186
CV(1)	1.0140	.2507	.3946	.2605	.0816
CV(2)	1.2923	.3005	.3944	.2921	.0977

We now simulated 100 datasets for model 4 and model 5 to see how accuracy our method using Legendre system will be together with the effects due to initials values. We choose model 4 and model 5, because these models are much "rough" comparing with other three models. The numbers in Table 11 are the mean value of $m^2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ (as in Xia et al. 2002, section 2.1) for 100 datasets. The four initials are I1= $(1, ..., 1)^T$; I2= $(1, 1, 1, 1, 1, 0, ..., 0)^T$; I3= $(1, 1, 1, 0..., 0)^T$; I4= $(1, 1, ..., 0)^T$.

Table 4: Mean squared distance with Legendre polynomial.

	I1	I2	I3	I4
model 4 $n = 200$	.0196	.0108	.0109	.0102
model 4 $n = 400$	.0051	.0053	.0046	.0045
model 5 $n = 200$	.2653	.1257	.0431	.0358
model 5 $n = 400$	.1132	.0462	.0173	.0170

Model 6: mean function model. This is a two-dimensional model. Let $p = 10$, $X_1, ..., X_{10}, \epsilon$ are iid $N(0,1)$. The model is $Y = (\boldsymbol{\beta}_1^T\mathbf{X})^2 + (\boldsymbol{\beta}_2^T\mathbf{X})^2 + .5\epsilon$, where $\boldsymbol{\beta}_1^T = (1, 0, ..., 0)$ and $\boldsymbol{\beta}_2^T = (0, 1, 0, ..., 0)$. Sample size $n = 200$. The bootstrap method with $B = 200$ conclude $d = 2$ with $m_q(1) = .3256, m_q(2) =, 1028, m_q(3) = .4229$ and $m_r(1) = .3256, m_r(2) = .0435, m_r(3) = .1116$. This agrees with the cross-validation method since $CV(0) = 6.7907$, $CV(1) = 2.7216$, $CV(2) = 1.7813$, and $CV(3) = 2.0794$. And the correlation coefficients between the estimated variables and the two true variables are .9686 and .9351, respectively.

Model 7: mean and variance function model. Let $p = 10$, $X_1, ..., X_{10}, \epsilon$ are iid $N(0,1)$. The model is $Y = \boldsymbol{\beta}_1^T\mathbf{X} + 0.25e^{\boldsymbol{\beta}_2^T\mathbf{X}}\epsilon$, where $\boldsymbol{\beta}_1^T = (1, 1, 0, 0, 0, ..., 0)$ and $\boldsymbol{\beta}_2^T = (0, 0, 1, -1, 0, ..., 0)$. Sample size $n = 200$. The cross-validation method estimate that $d = 1$ for CMS, since $CV(0) = 5.5823$, $CV(1) = 3.3194$, and $CV(2) = 3.3958$. And the correlation coefficient between the estimated variable and the true variable is .9738. Apply our method with $K = 2$, our method identify two-dimensional for central second moment subspace. We finally identify CKMS has dimension 2 with the corresponding correlation coefficients being .9462 and .8491 respectively.

Model 8: partial mean and variance function model. This model is similar to the model studied by Carroll and Li (1996). Let $p = 10$, $X_1, ..., X_{10}, \epsilon$ are iid $N(0, 1)$. The model is $Y = (5+\boldsymbol{\beta}^T\mathbf{X}+W+.5\epsilon)^2$, where $\boldsymbol{\beta}^T = (1, 1, -1, -1, 0, ..., 0)$ and $W \sim Bin(1, .5)$. We simulated a dataset with sample size $n = 400$. Based on our procedure, for each separated group, and combined groups, we again identify 1 dimensional structure with correlation coefficients ranging from .9638 to .9944.

Model 9: Comparison with MAVE (Xia et al. (2002). We adopt the example in Li (1991), which is also used by Xia et al. (2002, model (4.2)) for MAVE method. Let $p = 10$, $X_1, ..., X_{10}, \epsilon$ are iid $N(0, 1)$. The model is $Y = \frac{\boldsymbol{\beta}_1^T\mathbf{X}}{(.5+(\boldsymbol{\beta}_2^T\mathbf{X}+1.5)^2)} + .5\epsilon$, where $\boldsymbol{\beta}_1^T = (1, 0, 0, ..., 0)$ and $\boldsymbol{\beta}_2^T = (0, 1, 0, ..., 0)$. With sample size $n = 200$ and the same distances $m^2(\hat{\boldsymbol{\beta}}_1, B_0)$ and $m^2(\hat{\boldsymbol{\beta}}_2, B_0)$ calculated in Xia et al. (2002. section 2.1.1) for 100 replicated datasets, the mean of them are .1225 and .1623, respectively. With sample size $n = 400$, the corresponding statistics are .057 and .0731 respectively. These results are comparable with MAVE method as we expected.

Real Data Sets.

Motorcycle data. We modified the motorcycle data that was used by H$\ddot{a}$rdle (1990), Eubank (1999) and Fan and Gijbels (1996) for nonparametric estimation methods. The original motorcycle data contains two variables: the response variable (Y) of head acceleration (in g) of a PTMO (post mortern test object) and the predictor variable (X_1) of time (in milliseconds) after a simulated impact with motorcycles.

In our study here we add 9 more independent variables $X_i \sim N(0,1)$ for $i = 2, ..., 10$. We simulated 10 datasets, all of them have the similar results as the following: The correlation coefficients range from .9201 to .9928. With a typical data, the bootstrap method with $B = 200$ shows that $m_q(1) = .0518, m_q(2) = .5032, m_r(1) = .0518, m_r(2) = .2041$ and thus clearly $d = 1$. However, CV method shows that $CV(0) = 2317.5, CV(1) = 1281.3, CV(2) = 1277.9, CV(3) = 1348.7$. Concluding $d = 1$ seems reasonable but not as firm as bootstrap method. In such a case, graphical plot may help. Figure 3a shows the response vs the estimated variable recoving the original variable while Figure 3b shows the response vs the second estimated variable which has no significant structure.

Ozone data. We take a data set from Breiman and Friedman (1985), the data for studying the atmospheric ozone concentration in the Los Angeles basin. The response (Y) is the daily measurement of ozone concentration in Upland. The eight predictors are Sandburg air force base temperature (X_1), inversion base height (X_2), Dagget pressure gradient (X_3), visibility (X_4), Vandenburg 500 millibar height (X_5), humidity (X_6), inversion base temperature (X_7) and wind speed (X_8). This data was also studied by Li (1992) for pHd method.

Table 11 below shows our numerical results with bootstrap method and CV method. Clearly the m_q criterion concludes $d = 2$, m_r criterion may conclude $d = 3$ (note that m_r always tends to have small increases comparing with m_q) while CV criterion conclude $d = 2$. Thus it is reasonable to infer $d = 2$.

Table 5: Cross validation with Legendre polynomial.

$m_q(1)$	$m_q(2)$	$m_q(3)$	$m_q(4)$	$m_q(5)$
.0056	.0252	.3250	.5595	.5041
$m_r(1)$	$m_r(2)$	$m_r(3)$	$m_r(4)$	$m_r(5)$
.0056	.0113	.0884	.1158	.0812
$CV(0)$	$CV(1)$	$CV(2)$	$CV(3)$	$CV(4)$
63.9861	17.4642	17.1435	17.3684	19.0169

12 Further research

Other linear estimators: Many other CONS system can be used in a similar fashion.

Possible improvements: for the methods used in this paper, there are still many possible improvements. For instance, in building a model by Fourier series, one can improve the efficiency by shrinkage idea (Stein, 1956), and further discussions by Efromovich and Pinsker (1982), Efromovich (1985, 1996) and Nussbaum (1985).

Outliers: Outliers may affect our method since we used moments. Particularly with CKMS for y^k when big k is used. Therefore robust estimators such as M-estimator may be used.

Fast computation. Some algorithm can be improved computationally via Fast Fourier Transform (Silverman, 1982) and H*ä*rdle, 1987.

Theoretical interests. Consistency and asymptotic results for methods other than kernel and multiple dimensions.

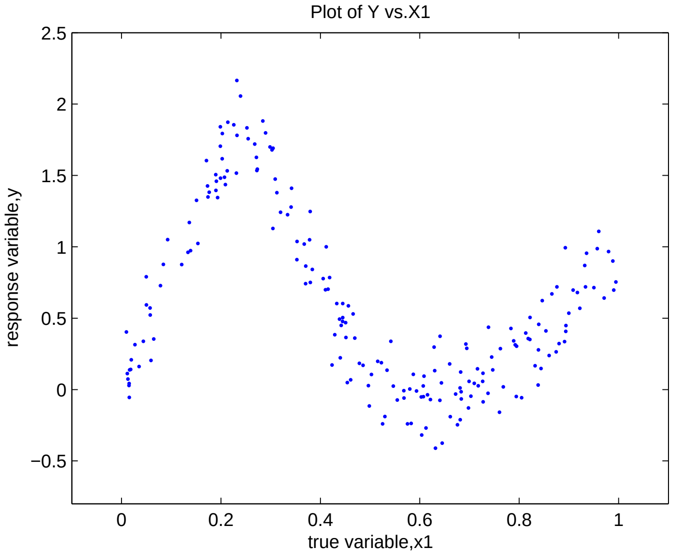


Figure 1: Response vs true variable for model 4: Continuous but not differentiable mean function

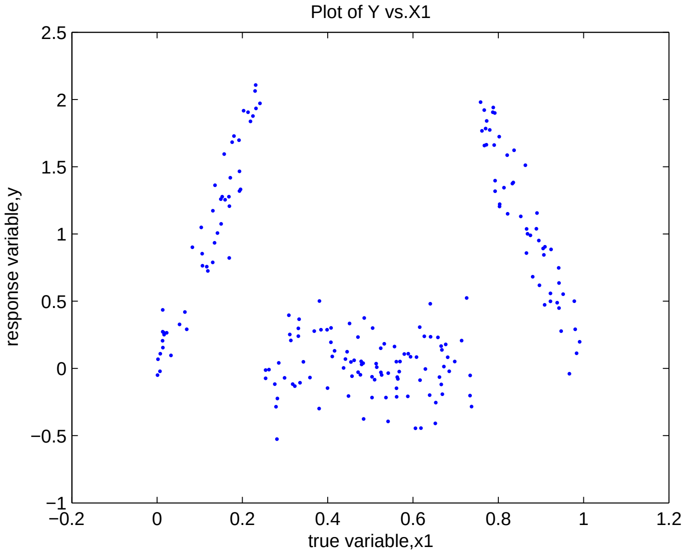
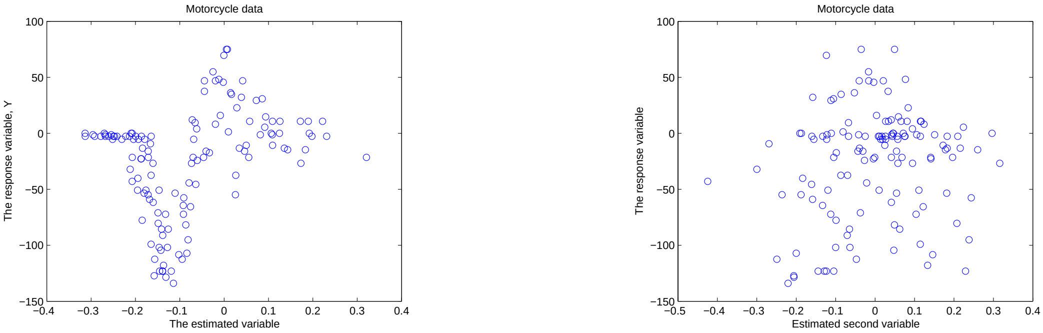


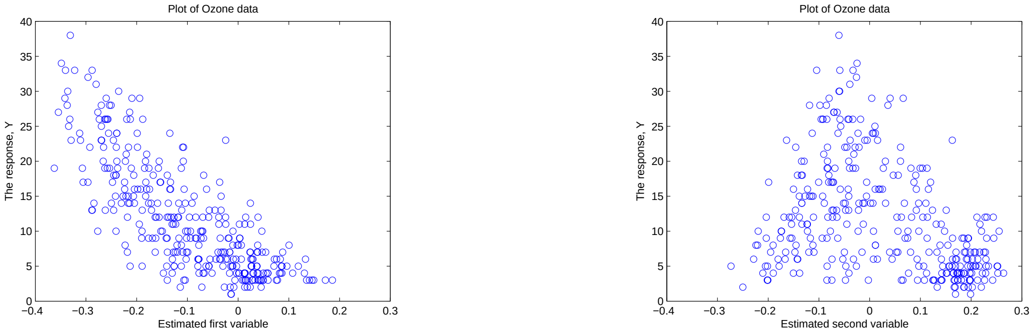
Figure 2: Response vs true variable for model 4: Discontinuous mean function



a. Estimated variable based one direction search

b. Estimated second variable based on two direction search

Figure 3: Motorcycle data (motor1)



a. Y vs first estimated variable, strong linear relation

b. Y vs second estimated variable, nonlinear structure

Figure 4: Ozone data (ozone1)