Applied Mathematics, 2013, 4, 1189-1198
http://dx.doi.org/10.4236/am.2013.48159 Published Online August 2013 (http://www.scirp.org/journal/am)

Estimation of Regression Model Using a Two Stage Nonparametric Approach

Desale Habtzghi[1], Jin-Hong Park[2*]
[1]Department of Statistics, University of Akron, Akron, USA
[2]Department of Mathematics, College of Charleston, Charleston, USA
Email: *dh52@uakron.edu, *parkj@cofc.edu

Received April 2, 2013; revised May 2, 2013; accepted May 9, 2013

ABSTRACT

Based on the empirical or theoretical qualitative information about the relationship between response variable and covariates, we propose a new approach to model polynomial regression using a shape restricted regression after estimating the direction by sufficient dimension reduction. The purpose of this paper is to illustrate that in the absence of prior information other than the shape constraints, our approach provides a flexible fit to the data and improves regression predictions. We use central subspace to estimate the directions and fit a final model by shape restricted regression, when the shape is known or is stipulated from empirical inspection. Comparisons with an alternative nonparametric regression are included. Simulated and real data analyses are conducted to illustrate the performance of our approach.

Keywords: Dimension Reduction; Central Subspace; Polynomial Regression; Shape Restricted

1. Introduction

Even if the assumption of monotonicity, convexity or concavity is common, shape restricted regression has not been extensively applied in real applications for two main reasons. As the number of observations (n), the data dimensionality (d), and the number of constraints (m) increases, computational and statistical difficulties (*i.e.* overfitting) are encountered, refer to [1,2] for detailed discussion. These and other authors proposed different methods to overcome the computational difficulties but there is no optimal solution.

To tackle these limitations, we estimate the direction by the sufficient dimension reduction and fit a final model by the shape restricted regression based on the theoretical shape or stipulated shape of the empirical results. The recent literature for the sufficient dimension reduction proposed practical methods which provide adequate information about the regression with many predictors. Reference [3] considered a general method for estimating the direction in regressions that can be described fully by linear combinations of the predictors without assuming a model for the conditional distribution of $Y|X$, where Y and X are response and explanatory variables, respectively. They also introduced a method to estimate the direction in a single-index regression and [4] extended it to multiple index regression by successive direction extraction.

More specifically, the main goal of this research is to show that the polynomial regression modeling by Central Subspace (CS) and Shape Restriction (SR) methods works well in practice, especially if the scatter plot shows a pattern. As is known that the curve fitting is finding a curve which matches a series of data points and possibly other constraints. This approach is commonly used by scientists and engineers to visualize and plot the curve that best describes the shape and behavior of their data. When more than two dimensions are used, we do not have the luxury of graphical representation any more but have theoretical information about the relationship of the response variable and predictors. Shape restricted regression is a non-parametric approach for building models whose fits are monotone, convex or concave in their covariates. Thses assumptions are commonly applied in biology [5], ranking [6], medicine [7], statistics [8] and psychology [9].

In general, one fits a straight line when the relationship between the response variable and the linear combination of the predictors is linear. Otherwise, one applies poly-

*Corresponding author.

nomial, logarithmic or exponential regression to fit the data. These regressions are practical methodologies when the mean function with predictors is smooth. It is well-known that the estimation approaches from regression theory are useful in building linear or nonlinear relationships between the values of the predictors and the corresponding conditional mean of the response variable. See [10] for a detailed exposition of widely studied regression methods, particularly polynomial regression. However, the straightforward and efficient analysis may not be generally possible with many predictors. In many situations when the underlying regression function or scatter plot has a particular shape or form, the fitted model can be characterized by certain order or shape restrictions. In this case, the shape restricted classes of regression function are preferred. This nonparametric regression method provides a flexible fit to the data and improves regression predictions.

In addition, when the empirical results between the response and predictors appear to have a particular shape that has certain order or shape restrictions, the shape restricted regression functions may best explain the relationships. Taking shape restrictions into account, one can reduce the model root mean square error or increase the power of the test. This improves the efficiency of a statistical analysis, provided that the hypothesized shape restriction actually holds [11].

In order to contextualize the goal of this article, it is necessary to review the concept of CS and SR. In Section 2, we summarize the notion of CS and an estimation method of CS when the dimension d is assumed to be known. Also, we suggest a data dependent approach to detect the unknown dimensions. In Section 3, we review the shape restricted regression and the constraint cone, over which we minimize the sum of squared errors of our approach for one dimension case. We apply our new approach to the simulated and a real data in Section 4. There are a few comments and concluding remarks in Section 5.

2. Estimation Method by Central Subspace

Let Y be a scalar response variable and X be a $p\times 1$ covariate vector. Suppose the goal is to make an inference about how the conditional distribution $Y|X$ varies with the values of X. Then, the sufficient dimension reduction method is to find the number of linear combinations, $\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{X},\cdots,\boldsymbol{\beta}_q^{\mathrm{T}}\boldsymbol{X}$, for $q\le p$ such that the conditional distribution of $Y|\boldsymbol{X}$ is the same as the conditional distribution of $Y|\left(\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{X},\cdots,\boldsymbol{\beta}_q^{\mathrm{T}}\boldsymbol{X}\right)$. In other words, there would be no loss of information of predictors if X were replaced by the $q(\le p)$ linear combinations. This is equivalent to finding a $p\times q$ matrix $\boldsymbol{B}=\left(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_q\right)$ such that

$$Y \perp\!\!\!\perp \boldsymbol{X}\,|\,\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X}, \tag{1}$$

where $\perp\!\!\!\perp$ indicates independence, (1) holds when $\boldsymbol{B}$ is a matrix whose columns form a basis for the subspace of $\mathbb{R}^p$. Therefore, a Dimension Reduction Subspace (DRS) for Y on $\boldsymbol{X}$ is defined as any subspace $\mathcal{S}(\boldsymbol{B})$ of $\mathbb{R}^p$, for which (1) holds. Here $\mathcal{S}(\boldsymbol{B})$ is defined as the space spanned by the columns of $\boldsymbol{B}$. That is, (1) represents that Y is independent of $\boldsymbol{X}$ given $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X}$ and $p\times 1$ vector $\boldsymbol{X}$ can be replaced by the $q\times 1$ vector $\boldsymbol{B}^{\mathrm{T}}\boldsymbol{X}$. This indicates a useful reduction in the dimension of $\boldsymbol{X}$, where all the information in $\boldsymbol{X}$ about Y is included in the q-linear combinations. Here, (1) holds trivially for $\boldsymbol{B}=I_p$ and a dimension reduction subspace always exists. Hence, if the intersection of DRSs is itself a DRS, the Central Subspace (CS) is defined as the intersection of all DRSs, which is written as $\mathcal{S}_{Y|\boldsymbol{X}}(\boldsymbol{B}_d)$ for dimension d and $\boldsymbol{B}_d=(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_d)$. That is, CS is the minimum DRS that preserves the original information relating to the data.

In this article, we use a method for estimating the CS, $\mathcal{S}_{Y|\boldsymbol{X}}(\boldsymbol{B}_d)$, which does not require a pre-specified model for $Y|\boldsymbol{X}$. If dimension $d(\le p)$ of CS are known, we need to estimate only the set of vectors $(\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_d)$. The ultimate goal of this paper is to use these estimated vectors to fit a final model using SR, discussed in Section 3. While [12] considered multivariate kernel estimation of the predictor density, the method introduced in this paper uses one predictor at a time. As a result, it can reduce the computational complexity and avoid the sparsity caused by high-dimensional kernel smoothing.

Suppose a matrix $\boldsymbol{h}\in\mathbb{R}^{p\times q}$ with $q\le p$ and define an information index $\Upsilon(\boldsymbol{h})$ by

$$\Upsilon(\boldsymbol{h})=\mathrm{E}\left[\log\frac{p\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{X},Y\right)}{p(Y)\,p\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{X}\right)}\right]=\mathrm{E}\left[\log\frac{p\left(Y|\boldsymbol{h}^{\mathrm{T}}\boldsymbol{X}\right)}{p(Y)}\right]. \tag{2}$$

The two forms in (2) are the informational correlation and the expected conditional log-likelihood, respectively. The idea behind this setting is to maximize the information index Υ over all $p\times d$ matrices $\boldsymbol{h}$ when $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h}=I$. Because $p(Y)$ does not involve $\boldsymbol{h}$, maximizing $\Upsilon(\boldsymbol{h})$ is equivalent to maximizing the expected conditional log-likelihood. This information index is similar to the Kullback-Leibler information between the joint density $p\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{X},Y\right)$ and the product of the marginal densities $p(Y)\,p\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{X}\right)$, quantifying the dependence of Y on $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{X}$. The important properties of the above information index $\Upsilon(\boldsymbol{h})$ is supported by Proposition 1 of [4].

The computation starts to maximize the sample version of Υ to estimate a basis for the CS. If all the densities were known, a sample version $\Upsilon_n(\boldsymbol{h})$ of $\Upsilon(\boldsymbol{h})$ is

$$\Upsilon_n(\boldsymbol{h}) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{X}_i, Y_i\right)}{p\left(Y_i\right) p\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{X}_i\right)}.$$

$\Upsilon_n(\boldsymbol{h})$ is maximized over all $p \times d$ matrices $\boldsymbol{h}$. Because the densities in $\Upsilon_n(\boldsymbol{h})$ are practically unknown, we use the nonparametric approach to estimate one-dimensional and multi-dimensional density estimates. Here, for the choice of kernels and selection of bandwidths, we follow the general guideline proposed by [13]. Since the Gaussian kernel performed well for the simulated and real data sets, we use density estimates based on a Gaussian kernel for the one-dimensional density and a product of Gaussian kernels for the multi-dimensional densities. Let G be the univariate Gaussian kernel, $(w_1, \cdots, w_a)^T$ be the $a \times 1$ vector, and $\left(w_{1i}, \cdots, w_{ai}\right)^{\mathrm{T}}$ be the i^{th} observation. Then the a-dimensional density estimate has the following form:

$$p_n\left(w_1, \cdots, w_a\right) = \left(n\prod_{j=1}^{a} b_{nj}\right)^{-1} \sum_{i=1}^{n}\prod_{j1}^{a} G\left(\frac{w_j - w_{ji}}{b_{nj}}\right), \qquad (3)$$

where $b_{nj} = k_a s_j n^{-1/(4+a)}$ for $j = 1, \cdots, a$. s_j is the corresponding sample standard deviation of w_j and the constant k_a $\left(=\left[4/a+2\right]^{1/a+4}\right)$ is the optimal bandwidth in the sense of minimizing the mean integrated square error from [13]. The density in $\Upsilon_n(\boldsymbol{h})$ is replaced by the estimates defined in (3) and maximize (4) for all $p \times d$ matrices $\boldsymbol{h}$ such that $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h} = I_d$.

$$\hat{\Upsilon}_n(\boldsymbol{h}) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_n\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{X}_i, Y_i\right)}{p_n\left(Y_i\right) p_n\left(\boldsymbol{h}^{\mathrm{T}}\boldsymbol{X}_i\right)} \qquad (4)$$

This method incorporates $\boldsymbol{h}^{\mathrm{T}}\boldsymbol{h} = I_d$; it is the sequential quadratic programming procedure of [14].

Since prior information about d may not be available in practice, it will be useful to find a simpler way to determine d using the data. The sufficient dimension reduction methods have been proposed for the determination of the minimal dimension d of the CS. See [15-17] for details. In this paper, using the estimating function $\hat{\Upsilon}_n(\boldsymbol{h})$ defined in (4), we suggest the following Akaike Information Criterion (AIC) to determine d.

$$AIC: \hat{d} = \arg\min_{d}\left\{-2n\hat{\Upsilon}_n\left(\boldsymbol{h}_{d,p}\right) + 2dp\right\} \qquad (5)$$

3. Fitting Model with Shape Restricted Regression

In this section, we review some fundamental concepts that can help us to lay the groundwork for the construction of the shape restricted method. More details about the properties of the constraint cone and polar cones can be found in [11,18-21].

Suppose we have the following model

$$Y = f\left(\boldsymbol{X}^*\right) + \epsilon$$

where $\boldsymbol{X}^* = \left(\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{X}, \cdots, \boldsymbol{\beta}_q^{\mathrm{T}}\boldsymbol{X}\right)$ and $f: \mathbb{R}^q \to \mathbb{R}$.

In this model the errors ε_i's are independent and have standard normal distribution. $f(\cdot)$ can be monotone, convex or concave based on the qualitative information about the relationship between response variable and predictors or empirical results.

For simplicity let $q = 1$ and $\theta_i = f\left(x_i^*\right)$. For $q > 1$ see [20,22] for detailed discussion. The constrained set over which we minimize the sum of squared errors is constructed as follows: the monotone nondecreasing constraints can be written as

$$\theta_1 \le \theta_2 \le \cdots \le \theta_n$$

The restriction of f to the set of convex functions is accomplished by the inequalities

$$\frac{\theta_2 - \theta_1}{x_2^* - x_1^*} \le \frac{\theta_3 - \theta_2}{x_3^* - x_2^*} \le \cdots \le \frac{\theta_n - \theta_{n-1}}{x_n^* - x_{n-1}^*}.$$

In our case, x_i^* is a realized value of the linear combination of the predictors; $\boldsymbol{X}^*$ is estimated using CS. Any of these sets of inequalities defines m half spaces in R^n, and their intersection forms a closed polyhedral convex cone in R^n. The cone is designated by $\boldsymbol{C} = \left\{\boldsymbol{\theta} : \boldsymbol{A}\boldsymbol{\theta} \ge 0\right\}$ for $m \times n$ constraint matrix $\boldsymbol{A}$. Here, $m = n-1$ for monotone, nondecreasing convex and $m = n-2$ for convex.

For monotone constraints, the nonzero elements of the $m \times n$ dimensional constraint matrix $\boldsymbol{A}$ are $A_{i,i} = -1$ and $A_{i,i+1} = 1$ for $1 \le i \le n-1$. For convex constraints the nonzero elements of A are $A_{i,i} = x_{i+2}^* - x_{i+1}^*$, $A_{i,i+1} = x_i^* - x_{i+2}^*$ and $A_{i,i+2} = x_{i+1}^* - x_i^*$ for $1 \le i \le n-2$.

For example, if $n = 5$, the monotone constraint matrix $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

If $n = 5$ and the x-coordinates are equally spaced, the nondecreasing concave and convex constraints are given by the following constraint matrices, respectively:

$$\boldsymbol{A} = \begin{pmatrix} -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix},$$

and

$$\boldsymbol{A} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{pmatrix}.$$

Some computational details: The ordinary least-squares regression estimator is the projection of the data vector $\boldsymbol{y}$ on to a lower-dimensional linear subspace of $\mathbb{R}^n$. In contrast the shape restricted estimator can be obtained through the projection of $\boldsymbol{y}$ on to an m dimensional polyhedral convex cone in $\mathbb{R}^n$ [23]. We have the following useful proposition which shows the existence and uniqueness of the projection of the vector $\boldsymbol{y}$ on a closed convex set (see [11]).

Proposition 1 *Let* $\boldsymbol{C}$ *be a closed convex subset of* $\mathbb{R}^n$.

1) For $\boldsymbol{y}\in\mathbb{R}^n$ and $\boldsymbol{\theta}\in\boldsymbol{C}$, the following properties are equivalent:

a) $\left\|\boldsymbol{y}-\hat{\boldsymbol{\theta}}\right\| = \min_{\boldsymbol{\theta}\in\boldsymbol{C}}\left\|\boldsymbol{y}-\boldsymbol{\theta}\right\|$

b) $\left\langle \boldsymbol{y}-\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right\rangle \le 0$ for all $\boldsymbol{\theta}\in\boldsymbol{C}$

2) For every $\boldsymbol{y}\in\mathbb{R}^n$, there exists a unique point where $\hat{\boldsymbol{\theta}}\in\boldsymbol{C}$ satisfies (a) and (b). $\hat{\boldsymbol{\theta}}$ is said to be the projection of $\boldsymbol{y}$ onto $\boldsymbol{C}$,

where the notation $\langle \boldsymbol{a},\boldsymbol{b}\rangle = \sum a_i b_i$ refers to the vector inner product of $\boldsymbol{a}$ and $\boldsymbol{b}$. It is easy to see that (1*b*) of Proposition 1 becomes

$$\left\langle \boldsymbol{y}-\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}\right\rangle = 0 \quad \text{and} \quad \left\langle \boldsymbol{y}-\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}\right\rangle \le 0,\ \forall \boldsymbol{\theta}\in\boldsymbol{C}, \tag{6}$$

which are the necessary and sufficient conditions for $\boldsymbol{\theta}$ to minimize $\sum_{i=1}^{n}\left(y_i-\theta_i\right)^2$ over $\boldsymbol{C}$.

Let $\boldsymbol{V}$ be the linear space spanned by $\mathbf{1}=(1,\cdots,1)^{\mathrm{T}}$ for a monotone, nondecreasing convex, and nondecreasing concave, and let $\boldsymbol{V}$ be the linear space spanned by $\mathbf{1}=(1,\cdots,1)^{\mathrm{T}}$ and $\boldsymbol{x}^*=(x_1^*,\cdots,x_n^*)^{\mathrm{T}}$ for convex regression. Note that $\boldsymbol{V}\in\boldsymbol{C}$ in both cases. The constraint cone can be specified by a set of linearly independent vectors $\boldsymbol{\delta}^1,\cdots,\boldsymbol{\delta}^m$ as

$\boldsymbol{\Omega}=\left\{\boldsymbol{\theta}:\boldsymbol{\theta}=\sum_{j=1}^{m} b_j\boldsymbol{\delta}^j,\ b_j\ge 0\right\}$ and the constraint set as

$\boldsymbol{C}=\left\{\boldsymbol{\theta}:\boldsymbol{\theta}=\sum_{j=1}^{m} b_j\boldsymbol{\delta}^j+\boldsymbol{\nu}: b_1,\cdots,b_m,b_j\ge 0 \text{ and } \boldsymbol{\nu}\in\boldsymbol{V}\right\}$,

where $m=n-1$ for monotone, nondecreasing concave, nondecreasing convex and $m=n-2$ for convex. The vectors $\boldsymbol{\delta}^j$ can be obtained from the formula $\left[\boldsymbol{\delta}^1,\cdots,\boldsymbol{\delta}^m\right]' = \left(\boldsymbol{AA}'\right)^{-1}\boldsymbol{A}$. For example, any convex vector $\boldsymbol{\theta}\in\boldsymbol{C}$ is a nonnegative linear combination of the $\boldsymbol{\delta}^j$ vectors plus a linear combination of $\mathbf{1}$ and $\boldsymbol{x}^*$.

If $\boldsymbol{C}$ is the set of all convex vectors in $\mathbb{R}^n$, we can also define the vectors $\boldsymbol{\delta}^j$ to be the rows of the following matrix:

$$\begin{pmatrix} 0 & 0 & \frac{x_3^*-x_2^*}{x_n^*-x_2^*} & 0 & \cdots & \frac{x_{n-1}^*-x_2^*}{x_n^*-x_2^*} & 1 \\ 0 & 0 & 0 & \frac{x_4^*-x_3^*}{x_n^*-x_3^*} & \cdots & \frac{x_{n-1}^*-x_3^*}{x_n^*-x_3^*} & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

For a large data set, it is better to use the above vectors $\boldsymbol{\delta}^j$ because the previous method of obtaining the edges is computationally intensive. Another advantage is that the computations of the inner products with the second approach are faster because of all the zero entries in the vectors.

The polar cone of the constraint cone $\boldsymbol{\Omega}$ is ([19], p. 121)

$$\boldsymbol{\Omega}_0=\left\{\boldsymbol{\rho}:\left\langle\boldsymbol{\rho},\boldsymbol{\theta}\right\rangle\le 0,\forall\boldsymbol{\theta}\in\boldsymbol{\Omega}\right\}.$$

Geometrically, the polar cone is the set of points in $\mathbb{R}^n$ which make an obtuse angle with all points in $\boldsymbol{\Omega}$. Let us note some straightforward properties of $\boldsymbol{\Omega}_0$:

1) $\boldsymbol{\Omega}_0$ is a closed convex cone,

2) The only possible element in $\boldsymbol{\Omega}\cap\boldsymbol{\Omega}_0$ is 0,

3) $\boldsymbol{\gamma}^1,\cdots,\boldsymbol{\gamma}^m\in\boldsymbol{\Omega}_0$.

where $\boldsymbol{\gamma}^j$ is negative rows of $\boldsymbol{A}$, *i.e.*, $\left[\boldsymbol{\gamma}^1,\cdots,\boldsymbol{\gamma}^m\right]=-\boldsymbol{A}'$. The following proposition is a useful tool for finding the constrained least squares estimator. Its proof was discussed in detail by [23].

Proposition 2 *Given* $\boldsymbol{y}\in\mathbb{R}^n$ *such that* $\boldsymbol{y}=\sum_{j\in\boldsymbol{J}} b_j\boldsymbol{\delta}^j+\sum_{j\notin\boldsymbol{J}} b_j\boldsymbol{\gamma}^j+\nu$, *the projection of* $\boldsymbol{y}$ *onto the constraint set* $\boldsymbol{\Omega}$ *is*

$$\hat{\boldsymbol{\theta}}=\sum_{j\in\boldsymbol{J}} b_j\boldsymbol{\delta}^j+\nu. \tag{7}$$

and the residual vector $\hat{\boldsymbol{\rho}}=\boldsymbol{y}-\hat{\boldsymbol{\theta}}=\sum b_j\boldsymbol{\gamma}^j$ is the projection of $\boldsymbol{y}$ onto the polar cone $\boldsymbol{\Omega}_0$.$^{j\notin\boldsymbol{J}}$

If the set $\boldsymbol{J}$ is determined, using Proposition 2, the constrained least squares estimate, $\hat{\boldsymbol{\theta}}$, can be found through ordinary least-squares regression (OLS), using $\boldsymbol{\nu}\in\boldsymbol{V}$ and $\boldsymbol{\delta}^j$ for $j\in\boldsymbol{J}$ as regressors. It is clear that the number of iterations is finite, as there is only a finite number of faces of the cone. To find the set $\boldsymbol{J}$ and $\hat{\boldsymbol{\theta}}$, the mixed primal-dual bases algorithm of [18] or the hinge algorithm of [23] can be used. For monotone regression the pooled adjacent violators algorithm, known as PAVA [11] may be used. A code of the algorithm was written in **R**. The code can be obtained from the authors upon request.

4. Numerical Illustration

We examined the performance of the proposed methodology using a real and four simulated data sets. For each of the simulated data set, we carried out the computational algorithms as described in Sections 2 and 3 for sample size n = 100. Recently, [24] investigated a sufficient dimension reduction for different sample sizes and dimensions in the time series context. They found that the performance to detect a true dimension is improved as sample size increases and the computation is more intensive for higher dimensions such as $d=2$ and 3. For the first three examples, we simulated four data

sets which have the worst scenario, $n=100$, and computationally less intensive dimensions, $d=1$ and 2 only, to illustrate clearly how the shape restricted method works well in these directions. For the nonparametric alternative we used a kernel. Although optimal bandwidth selection is essential, we used data adaptive fixed bandwidth $b=\left(x_{\max}-x_{\min}\right)\Big/\left(8n^{0.2}\right)$ that was recommended by [25]. Here $x_{\max}$ and $x_{\min}$ are the maximum and minimum values used in estimation, respectively, and *n* is the sample size.

We considered quadratic regression models of a linear combination of six predictors for the first example and ten predictors for the second example. In the third example, we simulated data from cubic regression of a linear combination with ten predictors. The fourth example is more complicated data simulated from two dimensional model including ten predictors. In all the simulated data sets, ε are independent standard normal random variables. Finally, we applied our new approach to a real data set, *Highway Accident Data*.

Example 1: Model 1

$$Y=\left\{\left(-2x_1+x_2-4x_3+3x_4+x_5+2x_6\right)\Big/\sqrt{35}\right\}^2+0.5\varepsilon.$$

We simulated data from the above model where the mean function is a quadratic function of a linear combination with six predictors. First, we estimated the dimension and direction by CS combined with AIC (5) as described in Section 2. As shown in **Table 1**, we detected a true dimension $d=1$. We estimated a vector, $\hat{\boldsymbol{\beta}}_1=\left(-0.401,0.206,-0.668,0.525,0.130,0.241\right)$, by the algorithms described in Section 2. The scatter plot, $\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\boldsymbol{X}$ vs Y, shows a quadratic relation, see **Figure 1**. Then, based on the scatter plot, we fitted the data using convex regression (SR). The decision to use the convex regression was based on a visual examination of the scatter plot.

Example 2: Model 2

$$Y=\left\{1/2+\left(-2x_1+x_2+x_4\right)\Big/\sqrt{6}\right\}^2+0.5\varepsilon.$$

In this model, we considered another quadratic mean function of one linear combination with ten predictors. Using the same procedures as the previous example, we estimated dimension and direction by CS. Based on AIC, a true dimension $d=1$ was detected, see **Table 1** for details. The estimated vector is

$$\hat{\boldsymbol{\beta}}_1=\left(0.815,-0.378,-0.004,-0.437,0.000,0.003,0.013,0.013,0.046,0.029\right).$$

The scatter plot of **Figure 2**, $\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\boldsymbol{X}$ vs Y, shows quadratic relation. The scatter plot suggested that a reasonable choice of the relationship between the $\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\boldsymbol{X}$ and Y is a convex curve. Hence, we fitted a model using convex regression. **Figure 2** shows that the shape restricted regression fits the data better than kernel regression, which leads to a smaller error sum of squares.

Example 3: Model 3

$$Y=1/2+\left\{\left(x_1+x_3+x_5-x_{10}\right)\Big/\sqrt{4}\right\}^2+\left\{\left(x_1+x_3+x_5-x_{10}\right)\Big/\sqrt{4}\right\}^3+0.5\varepsilon.$$

In this example, we simulated data from a cubic polynomial model of one linear combination with ten predictors. In the first step, we estimated dimension and direction by CS. **Table 1** indicated that a true dimension $d=1$ was detected by AIC. The estimated vector is

$$\hat{\boldsymbol{\beta}}_1=\left(0.497,0.011,0.489,0.056,0.476,0.019,-0.055,-0.071,0.030,-0.525\right).$$

The scatter plot, $\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\boldsymbol{X}$ vs Y, shows a cubic curve. In the second step, we fitted a model by concave-convex regression based on the shape of the scatter plot. Our estimator was computed by minimizing the sum of squared errors over the concave-convex set. The inflection point was found by minimizing the sum of squared errors until the conditions in (6) were satisfied. Here, the shape restricted regression does fairly well in all ranges of the data set. The SR and kernel regression are almost on top of each other in the entire range of the data. See **Figure 3** and **Table 1** for details.

Example 4: Model 4

$$Y=\left(x_2+x_4+x_6+x_8\right)\Big/\sqrt{4}+\exp\left\{\left(x_1+x_3+x_5+x_{10}\right)\Big/\sqrt{4}\right\}+0.1\varepsilon,$$

where $d=2$ and $p=10$. Here, we consider a nonlinear function including two dimensions. **Table 1** showed that we estimated true dimensions $d=2$ by AIC. The two directions are

$$\hat{\boldsymbol{\beta}}_1=\left(0.2468,0.4704,0.2715,0.4662,0.2357,0.3895,0.0020,0.4492,0.0031,0.1330\right)$$

and

$$\hat{\boldsymbol{\beta}}_2=\left(0.3122,-0.2625,0.2117,-0.2094,0.3894,-0.1028,-0.1432,-0.1400,0.0238,0.7363\right).$$

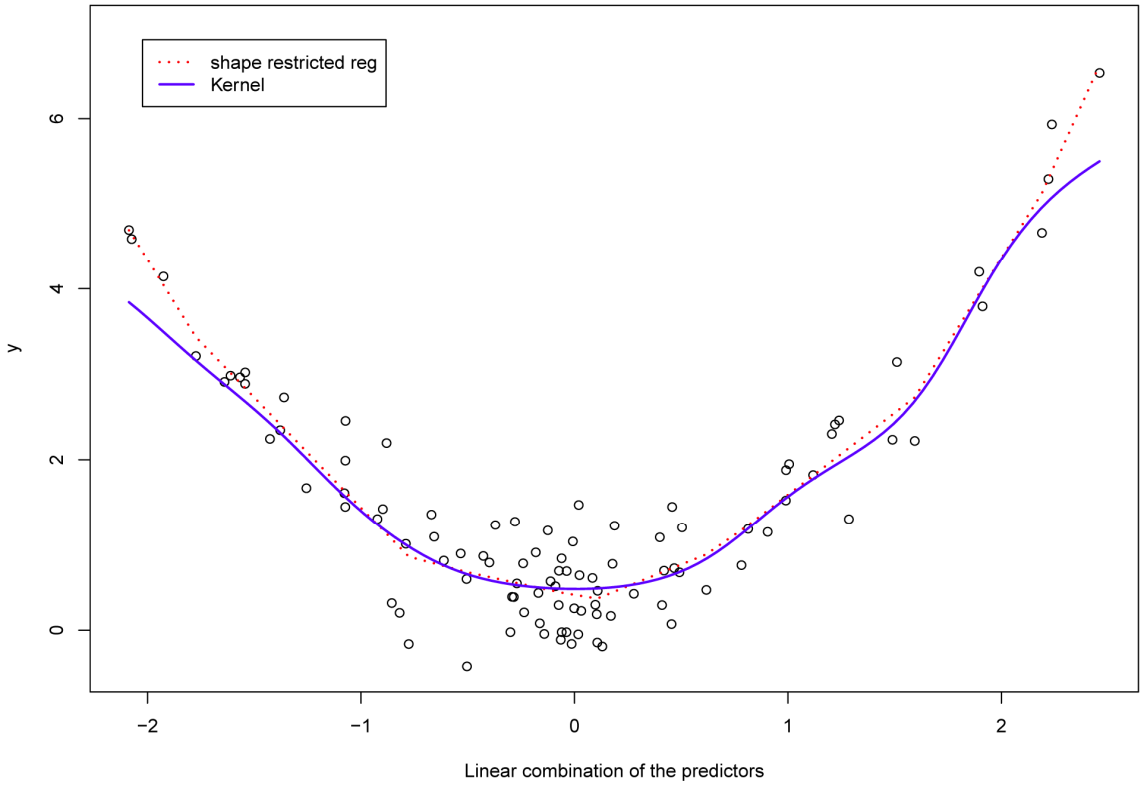

Figure 1. (Model 1) Data are generated from quadratic function of a linear combination with six predictors. The solid curve is quadratic fit and the dotted curve is the shape restricted fit.

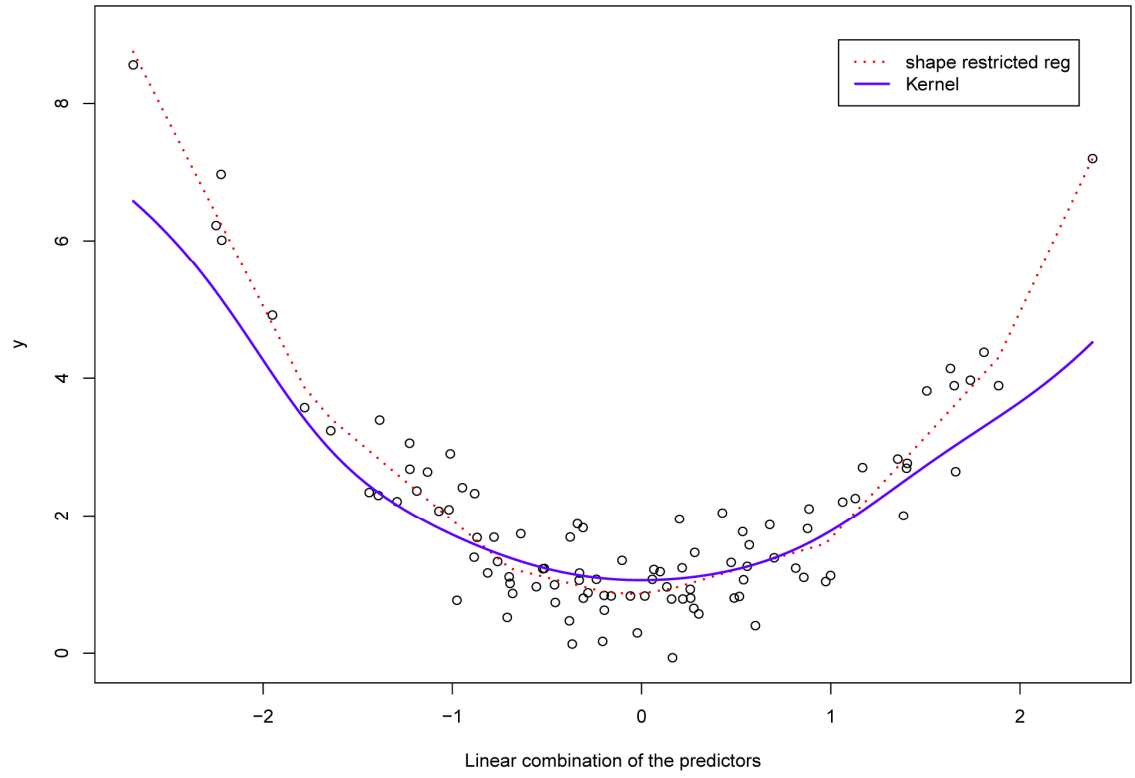

Figure 2. (Model 2) Data are generated from quadratic function of a linear combination with ten predictors. The solid curve is quadratic fit and the dotted curve is the shape restricted fit.

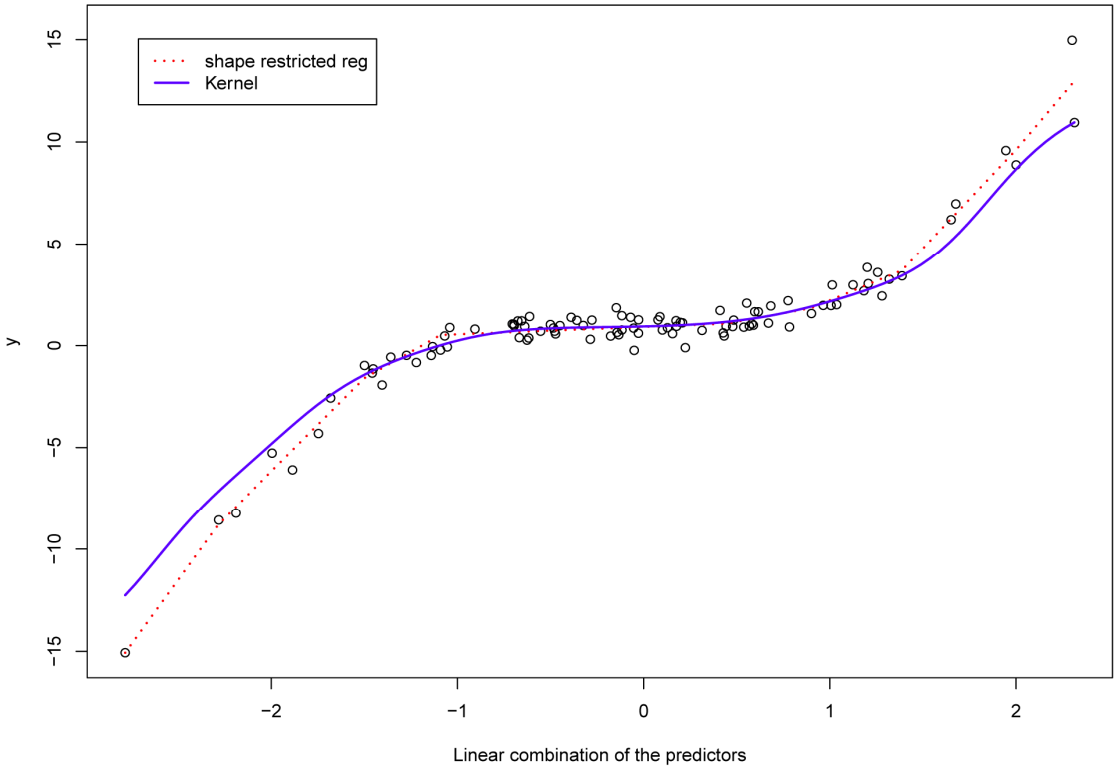

Figure 3. (Model 3) Data are generated from cubic function of a linear combination with six predictors. The solid curve is cubic fit and the dotted curve is the shape restricted fit.

Table 1. AIC values for the simulated and a real data sets using Equations (5).

d	1	2	3	4
Model 1	−78.94*	−74.38	−61.78	−50.78
Model 2	−79.78*	−73.26	−61.10	−43.96
Model 3	−42.78*	−18.50	12.22	47.24
Model 4	−116.46	−117.90*	−97.88	71.96
Highway Data	−27.20*	−26.08	−22.77	−11.31

The response variable Y is non-decreasing in both predictors $\left(\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\boldsymbol{X}, \hat{\boldsymbol{\beta}}_2^{\mathrm{T}}\boldsymbol{X}\right)$. In addition, the marginal scatter plots, $\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\boldsymbol{X}$ vs Y and $\hat{\boldsymbol{\beta}}_2^{\mathrm{T}}\boldsymbol{X}$ vs Y, display an increasing trends. Next, we fitted a model by a multiple isotonic regression. The isotonic fit is shown in **Figure 4**. This shows that our approach may be a better choice than parametric or nonparametric models that do not use the constraints and works well even for two dimensional model.

For the purpose of comparison, we computed Average Squared Error Loss (ASEL) of our models and alternative kernel regressions as the following:

$$\mathrm{ASEL} = \frac{1}{N}\sum_{i=0}^{N}\left(\hat{Y}_i - Y_i\right)^2$$

The entries of the **Table 2** are the square roots of the ASEL. The results from **Table 2** demonstrate that our method performed fairly well in all cases. It is better than kernel regression, in particular, when the data is generated from quadratic and cubic regressions. In general, we can see that our method provides better or comparable fits for the simulated examples, which is also supported by the results of **Figures 1-3**.

Example 4: *Highway Accident Data*

For illustration, we applied our method to a real data set, *Highway Accident Data*. See Weisberg (2005) for a detailed description about this data. The data include 39 sections of large highways in the state of Minnesota in 1973 and the variables relate the automobile accident rate in accidents per million vehicle miles to several potential terms. We use *log*(*Rate*) as a response variable and eleven terms as explanatory variables. The definition of terms of this data is described in Table 10.5 of Weisberg (2005).

First, we estimated the direction of the predictor variables without losing any information using CS. As shown in **Table 2**, the dimension $d = 1$ is detected by AIC. The estimated direction is

$$\left(0.469, 0.215, 0.452, -0.167, 0.025, 0.069, -0.052, 0.046, -0.558, 0.398, 0.150\right).$$

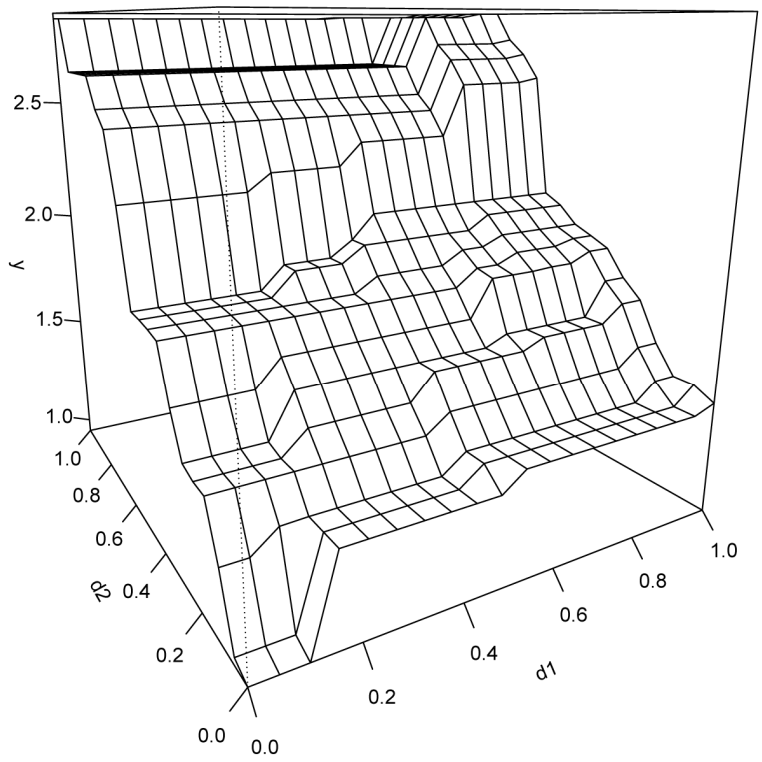

Figure 4. (Model 4) Data are generated from nonlinear mean function which has two dimensional model including ten predictors.

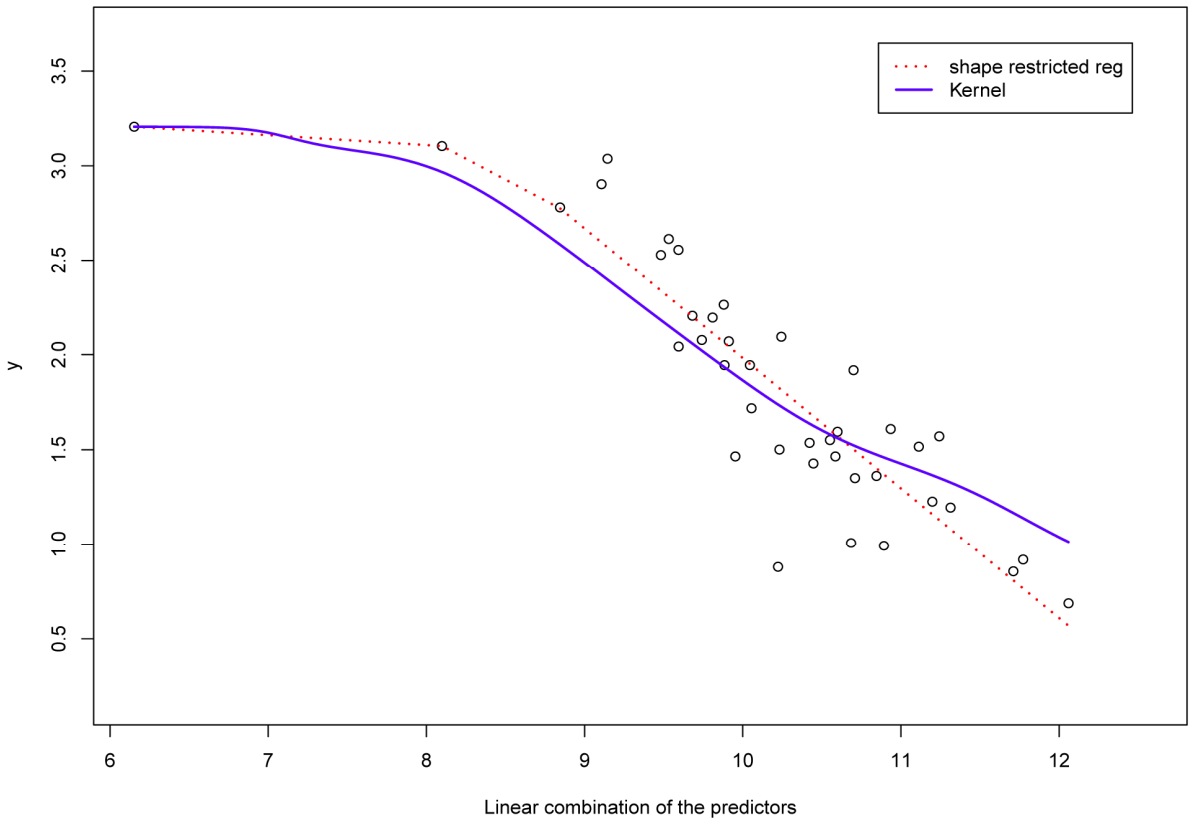

Figure 5. *Highway Accident Data:* The solid line is cubic fit and and the dotted line is the shape restricted fit.

Table 2. Average Squared Error Loss (ASEL) of the shape restricted regression and kernel regression.

Method	Model 1	Model 2	Model 3	Highway Accident
SR	0.184	0.212	0.302	0.057
Kernel	1.165	1.339	4.953	1.206

From the scatter plot of **Figure 4**, there is some curvature in the relationship between $\hat{\boldsymbol{\beta}}_1^{\mathrm{T}}\boldsymbol{X}$ vs Y. Hence, a concave curve may be a good choice to reflect the relationship between the response variable and the linear combination of the predictors. **Figure 4** shows the concave and cubic regression fits. The plot of **Figure 5** and the ASEL in **Table 2** suggests that our method gives reasonable fit to this data set.

5. Comments

The polynomial regression is one of handy methods in regression analysis. However, this straightforward analysis is not generally possible with many predictors. Hence, the major message that we would like to deliver in this paper is that the estimation of direction by CS and fitting the model by SR is advantageous for high dimensional data that has many predictors. After estimating the direction by sufficient dimension reduction, it is not easy to choose the appropriate polynomial regression model from the pattern of the scatter plot without any theoretical basis. Furthermore, the parametric or nonparametric models that do not use the constraints are not capable of giving different shapes of actual fit such as bimodal function and concave-convex. For such conditions, when the only available information is the shape (decreasing, increasing, concave, convex or bathtub) of the underlying regression function, our approach provides more acceptable fits/estimates.

6. Acknowledgements

Jin-Hong Park is supported in part by the faculty research and development at the Mathematics Department and the College of Charleston.

REFERENCES

[1] R. Luss, S. Rosset and M. Shahar, "Decomposing Isotonic Regression for Efficiently Solving Large Problems," *Proceedings of the Neural Information Processing Systems Conference*, Vancouver, 6-9 December 2010, pp. 1513-1521.

[2] W. Maxwell and J. Muckstadt, "Establishing Consistent and Realistic Reorder Intervals in Production-Distribution Systems," *Operations Research*, Vol. 33, No. 6, 1985, pp. 1316-1341. doi:10.1287/opre.33.6.1316

[3] X. Yin and R. D. Cook, "Direction Estimation in Single-Index Regressions," *Biometrika*, Vol. 92, No. 2, 2005, pp. 371-384. doi:10.1093/biomet/92.2.371

[4] X. Yin, B. Li and R. D. Cook, "Successive Direction Extraction for Estimating the Central Subspace in a Multiple-Index Regression," *Journal of Multivariate Analysis*, Vol. 99, No. 8, 2008, pp. 1733-1757. doi:10.1016/j.jmva.2008.01.006

[5] G. Obozinski, G. Lanckriet, C. Grant, M. Jordan and W. Noble, "Consistent Probabilistic Outputs for Protein Function Prediction," *Genome Biology*, Vol. 9, No. 1, 2008, pp. 247-254. doi:10.1186/gb-2008-9-s1-s6

[6] Z. Zheng, H. Zha and G. Sun, "Query-Level Learning to Rank Using Isotonic Regression," 46*th Annual Allerton Conference on Communication*, *Control*, *and Computing*, Allerton House, 24-26 September 2008, pp. 1108-1115.

[7] M. J. Schell and B. Singh, "The Reduced Monotonic Regression Method," *Journal of the American Statistical Association*, Vol. 92, No. 437, 1997, pp. 128-35. doi:10.1080/01621459.1997.10473609

[8] R. Barlow and H. Brunk, "The Isotonic Regression Problem and Its Dual," *Journal of the American Statistical Association*, Vol. 49, No. 5, 1972, pp. 784-789.

[9] J. Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, Vol. 29, No. 1, 1964, pp. 1-27. doi:10.1007/BF02289565

[10] S. Weisberg, "Applied Linear Regression," Wiley, Hoboken, 2005.

[11] T. Robertson, F. T. Wright and R. L. Dykstra, "Order Restricted Statistical Inference," John Wiley & Sons, New York, 1988.

[12] Y. Xia, "A Constructive Approach to the Estimation of Dimension Reduction Directions," *The Annals of Statistics*, Vol. 35, No. 6, 2007, pp. 2654-2690. doi:10.1214/009053607000000352

[13] D. W. Scott, "Multivariate Density Estimation: Theory, Practice, and Visualization," John Wiley & Sons, New York, 1992. doi:10.1002/9780470316849

[14] P. Gill, W. Murray and M. H. Wright, "Practical Optimization," Academic Press, New York, 1981.

[15] K. C. Li, "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," *Annals of Statistics*, Vol. 87, No. 420, 1992, pp. 1025-1039.

[16] J. R. Schott, "Determining the Dimensionality in Sliced Inverse Regression," *Journal of the American Statistical Association*, Vol. 89, No. 425, 1994, pp. 141-148. doi:10.1080/01621459.1994.10476455

[17] Y. Xia, H. Tong, W. K. Li and L. X. Zhu, "An Adaptive Estimation of Dimension Reduction," *Journal of the Royal Statistical Society*, *Ser. B*, Vol. 64, No. 3, 2002, pp. 363-410. doi:10.1111/1467-9868.03411

[18] D. A. S. Fraser and H. Massam, "A Mixed Primal-Dual Bases Algorithm for Regression under Inequality Constraints. Application to Convex Regression," *Scandinavian Journal of Statistics*, Vol. 16, 1989, pp. 65-74.

[19] R. T. Rockafellar, "Convex Analysis," Princeton University Press, 1970.

[20] E. Seijo and B. Sen, "Nonparametric Least Squares Estimation of a Multivariate Convex Regression Function," *Annals of Statistics*, Vol. 39, No. 2, 2011, pp. 1633-1657. doi:10.1214/10-AOS852

[21] M. J. Silvapulle and P. K. Sen, "Constrained Statistical Inference, Inequality, Order, and Shape Restrictions," Wiley, New York, 2005.

[22] M. C. Meyer, "Inference for Multiple Isotonic Regression," Technical Report, Colorado State Uinversity, 2010.

[23] M. C. Meyer, "An Extension of the Mixed Primal-Dual Bases Algorithm to the Case of More Constraints than Dimensions," *Journal of Statistical Planning and Inference*, Vol. 81, No. 1, 1999, pp. 13-31. doi:10.1016/S0378-3758(99)00025-7

[24] J.-H. Park, T. Sriram and X. Yin, "Dimension Reduction in Time Series," *Statistica Sinica*, Vol. 20, 2010, pp. 747-770.

[25] H.-G. Muller and J.-L. Wang, "Hazard Rate Estimation under Random Censoring with Varying Kernels and Bandwidths," *Biometrics*, Vol. 50, No. 1, 1994, pp. 61-76. doi:10.2307/2533197

