Iterative Optimal Sufficient Dimension Reduction for Conditional Mean in Multivariate Regression

Jae Keun Yoo
University of Louisville

Abstract: Recently, Yoo and Cook (2007) developed an optimal version of Cook and Setodji (2003). When predictors are not highly skewed, the Yoo-Cook approach can be improved, especially with small samples, by iteratively estimating the inner product matrix used in their method without changing their asymptotic results. Since highly skewed predictors are often transformed for normality in sufficient dimension reduction literature, the proposed method can have more useful application in practice than Yoo and Cook (2007).

Key words: Conditional mean, iterative approach, multivariate regression, sufficient dimension reduction.

1. Introduction

High-dimensional data has become common and inevitable in this current world. In most situations useful information needs to be extracted from such high-dimensional data. That is why dimension reduction ideas have been getting more important and many methodologies for dimension reduction has been developed so far. For example, methods such as principal component analysis, factor analysis, and canonical correlation analysis are in a long list of dimension reduction methodologies.

In the past decade *sufficient dimension reduction* (SDR) for regression has rapidly grown up. The basic idea of SDR is to replace the predictors $\mathbf{X} \in \mathbb{R}^p$ with a lower-dimensional linearly transformed predictor $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ without loss of information on selected aspects of the conditional distribution of $Y|\mathbf{X}$, where $\boldsymbol{\eta} \in \mathbb{R}^{p \times d}$ and $d \leq p$. SDR methods are nonparametric in the sense that they do not require a model for $Y|\mathbf{X}$, but unlike many local nonparametric approaches they can often avoid the curse of dimensionality, since estimates of dimension reduction subspaces are "global" and converge at the usual $\sqrt{n}$ rate.

When the conditional mean $E(Y|\mathbf{X})$ is of special interest (Cook and Li, 2002), the goal of an SDR inquiry might be usefully restricted to finding $\boldsymbol{\eta}$ such that

$$E(Y|\mathbf{X}) = E(Y|\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}). \tag{1.1}$$

Statement (1.1) implies that the d-dimensional linearly transformed predictor $\boldsymbol{\eta}^{\mathrm{T}}\mathbf{X}$ has the same amount of information about $E(Y|\mathbf{X})$ as the original p-dimensional predictor $\mathbf{X}$ has. If $d < p$ and such $\boldsymbol{\eta}$ can exist, a dimension reduction for $\mathbf{X}$ is attained without loss of information about $E(Y|\mathbf{X})$.

Our main interest is to recovering $\boldsymbol{\eta}$ satisfying (1.1) in multivariate regression of $\mathbf{Y} = (Y_1, \ldots, Y_r)^{\mathrm{T}} \in \mathbb{R}^r | \mathbf{X} \in \mathbb{R}^p$, where $r \geq 2$. Recently Cook and Setodji (2003) proposed a SDR method in this context. Yoo and Cook (2007) developed an optimal version of the former

The main goal of this article is to improve Yoo and Cook (2007) in both the estimation of $\boldsymbol{\eta}$ and predictor effect tests with no highly skewed predictors.

2. Dimension Reduction

2.1 Population structure of $\boldsymbol{\eta}$

Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_r)$, where $\beta_k = \boldsymbol{\Sigma}^{-1}\mathrm{cov}(\mathbf{X}, Y_k)$ is the population OLS coefficient vector of each coordinate regression $Y_k|\mathbf{X}$, and $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{X})$. We define the following relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$:

R1. $\mathrm{rank}(\boldsymbol{\beta}) = \mathrm{rank}(\boldsymbol{\eta})$ and $\boldsymbol{\beta} = \boldsymbol{\eta}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} \in \mathbb{R}^{d \times r}$.

The relation R1 between $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ indicates that $\boldsymbol{\beta}$ has the full information about $\boldsymbol{\eta}$ and holds under a multivariate linear regression of

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} + \boldsymbol{\delta}, \tag{2.1}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^r$ is an intercept vector, and $\boldsymbol{\delta} \in \mathbb{R}^r$ is an error vector that is independent of $\mathbf{X}$ and has mean $\mathbf{0} \in \mathbb{R}^r$.

According to Yoo and Cook (2007), the relation R1 also holds under the following two conditions:

Condition 1. Either (a) $E(\mathbf{X}|\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{X} = \nu_1)$ is is linear in $\nu_1 \in \mathbb{R}^1$ for all ks
or (b) $E(\mathbf{X}|\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} = \boldsymbol{\nu}_r)$ is is linear in $\boldsymbol{\nu}_r \in \mathbb{R}^r$.

Condition 2. $\mathrm{rank}(\boldsymbol{\eta}_k) \leq \mathrm{rank}(\boldsymbol{\beta})$ and $\boldsymbol{\eta}_k = \boldsymbol{\beta}\boldsymbol{\gamma}_k$, where each $\boldsymbol{\eta}_k \in \mathbb{R}^{p \times d_k}$
satisfies that $E(Y_k|\mathbf{X}) = E(Y_k|\boldsymbol{\eta}_k^{\mathrm{T}}\mathbf{X})$ and $\boldsymbol{\gamma}_k \in \mathbb{R}^{d_k \times r}$.

Condition 1, which is called *linearity condition*, is common in sufficient dimension reduction literature. If it does not hold, the predictors are often transformed to have normality in practice. The two statements of Condition 1 do not necessarily imply each other. Condition 1 implies that $\boldsymbol{\beta}$ can be expressed by a linear combination of $\boldsymbol{\eta}$ and Condition 2 indicates the reverse. For more about Conditions 1-2, readers can refer Yoo and Cook (2007).

Conditions 1-2 is to enable us to infer unknown parameter $\boldsymbol{\eta}$ without assuming any models via $\boldsymbol{\beta}$ whose estimator is known. Let $\hat{\boldsymbol{\beta}}$ be the usual moment estimator of $\boldsymbol{\beta}$. We will denote d as the true rank of $\boldsymbol{\eta}$. Emphasizing Conditions 1-2 for model-free approach, we explain Yoo and Cook (2007), which will be called *optimal approach*, with brief discussion about Cook and Setodji (2003) in the next section.

2.2 Optimal approach

The relation R1 enables us to estimate $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ with arguments $\hat{\mathbf{B}}$ and $\hat{\mathbf{C}}$ that minimize the following minimum discrepancy function over $\mathbf{B}$ and $\mathbf{C}$:

$$F_d(\mathbf{B}, \mathbf{C}) = \{\hat{\boldsymbol{\beta}}_{\mathrm{v}} - (\mathbf{BC})_{\mathrm{v}}\}^{\mathrm{T}} \mathbf{V}_n \{\hat{\boldsymbol{\beta}}_{\mathrm{v}} - (\mathbf{BC})_{\mathrm{v}}\}, \tag{2.2}$$

where $\mathbf{B} \in \mathbb{R}^{p \times d}$, $\mathbf{C} \in \mathbb{R}^{d \times r}$, $\mathbf{V}_n > 0$ is a $pr \times pr$ inner-product matrix and for a $p \times r$ matrix $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$, a notation of $\mathbf{A}_{\mathrm{v}}$ indicates a $pr \times 1$ vector constructed by stacking its columns: $\mathbf{A}_{\mathrm{v}} = (\mathbf{a}_1^{\mathrm{T}}, \ldots, \mathbf{a}_r^{\mathrm{T}})^{\mathrm{T}}$.

Any solution $\hat{\mathbf{B}}$ provides a consistent estimator of $\boldsymbol{\eta}$ for any choice of $\mathbf{V}_n > 0$, in the sense that $\hat{\mathbf{B}}$ converges to $\hat{\boldsymbol{\eta}}$. According to Yoo and Cook (2007), a best asymptotic choice for $\mathbf{V}_n$ is any consistent estimator of the inverse of the covariance matrix from the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathrm{v}} - \boldsymbol{\beta}_{\mathrm{v}})$, whose asymptotic distribution is a multivariate normal distribution with mean $\mathbf{0} \in \mathbb{R}^{pr}$ and covariance $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$. To explain a form of $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$ we define a $pr \times pr$ matrix of $\boldsymbol{\Gamma}$:

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Gamma}_{(1,1)} & \cdots & \boldsymbol{\Gamma}_{(1,r)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Gamma}_{(p,1)} & \cdots & \boldsymbol{\Gamma}_{(p,r)} \end{pmatrix} \tag{2.3}$$

where $\boldsymbol{\Gamma}_{(i,j)}$ is a $p \times p$ matrix.

Then $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$ has the same form as $\boldsymbol{\Gamma}$ in (2.3) with $\boldsymbol{\Gamma}_{(i,j)} = E[\varepsilon_i \varepsilon_j \boldsymbol{\Sigma}^{-1} \{\mathbf{X} - E(\mathbf{X})\}\{\mathbf{X} - E(\mathbf{X})\}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}]$, where $\varepsilon_i = Y_i - E(Y_i) - \boldsymbol{\beta}^{\mathrm{T}} \{\mathbf{X} - E(\mathbf{X})\}$.

We replace $\mathbf{V}_n$ in (2.2) with $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{-1}$, a consistent estimator of $\mathbf{V}_{\hat{\boldsymbol{\beta}}}^{-1}$, to obtain the optimal discrepancy function

$$F_d^{\mathrm{opt}}(\mathbf{B}, \mathbf{C}) = \{\hat{\boldsymbol{\beta}}_{\mathrm{v}} - (\mathbf{BC})_{\mathrm{v}}\}^{\mathrm{T}} \hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{-1} \{\hat{\boldsymbol{\beta}}_{\mathrm{v}} - (\mathbf{BC})_{\mathrm{v}}\}. \tag{2.4}$$

The estimation of $\boldsymbol{\eta}$ first requires the determination of its true rank d. It is done via a sequence of hypothesis tests (Li, 1991): Beginning with $m = 0$, test $H_0 : d = m$ vs. $H_1 : d > m$. If $H_0 : d = m$ is rejected, increment m by 1 and repeat the test, stopping the first time H_0 is not rejected and setting $\hat{d} = m$.

Define that $(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}}) = \arg_{\mathbf{B},\mathbf{C}} \min F_d^{\text{opt}}(\mathbf{B}, \mathbf{C})$ and that $\hat{F}_d^{\text{opt}} = F_d^{\text{opt}}(\hat{\boldsymbol{\eta}}, \hat{\boldsymbol{\gamma}})$. The optimal approach proposed $n\hat{F}_d^{\text{opt}}$ as a statistic for testing hypotheses of $H_0 : d = m$ versus $H_1 : d > m$. Then, under H_0, the statistic $n\hat{F}_d^{\text{opt}}$ is distributed asymptotically as $\chi^2_{\{(p-d)(r-d)\}}$, and the estimator $\hat{\boldsymbol{\eta}}$ is an efficient and consistent estimator of $\boldsymbol{\eta}$. The asymptotic efficiency of $\hat{\boldsymbol{\eta}}$ means that $\hat{\boldsymbol{\eta}}_{\text{v}}$ has minimum asymptotic variance within family (2.2).

The method by Cook and Setodji (2003) falls into family (2.2) with an inner-product matrix $\boldsymbol{\Gamma}$ in (2.3) with $\boldsymbol{\Gamma}_{(i,j)} = E(\varepsilon_i \varepsilon_j)\boldsymbol{\Sigma}$, and Cook and Setodji (2003) is optimal under model (2.1). Its asymptotic distribution under H_0 is either a weighted sum of independent χ_1^2 variables or $\chi^2_{\{(p-d)(r-d)\}}$.

2.3 Tests of predictor effects

Suppose that $\mathbf{X}$ is partitioned as $(\mathbf{X}_h, \mathbf{X}_{-h})$. Yoo and Cook (2007) tested the following three hypotheses related to $E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{Y}|\mathbf{X}_{-h})$.
Marginal predictor hypotheses:
$H_0 : E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{Y}|\mathbf{X}_{-h})$ vs. $H_1 : E(\mathbf{Y}|\mathbf{X}) \neq E(\mathbf{Y}|\mathbf{X}_{-h})$
Joint dimension-predictor hypotheses:
$H_0 : E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{Y}|\mathbf{X}_{-h})$ and $d = m$ vs. $H_1 : E(\mathbf{Y}|\mathbf{X}) \neq E(\mathbf{Y}|\mathbf{X}_{-h})$ or $d > m$
Conditional predictor hypotheses:
$H_0 : E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{Y}|\mathbf{X}_{-h})$ given $d = m$ vs. $H_1 : E(\mathbf{Y}|\mathbf{X}) \neq E(\mathbf{Y}|\mathbf{X}_{-h})$ given $d = m$.

If the statement of $E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{Y}|\mathbf{X}_{-h})$ holds, then a subset $\mathbf{X}_h$ does not contribute to $E(\mathbf{Y}|\mathbf{X})$. One of typical choices for $\mathbf{X}_h$ is an individual coordinate X_i of $\mathbf{X}$. Yoo and Cook (2007) shows that statistics for testing the three hypotheses are constructed by (2.4) and that all their asymptotic distributions are χ^2s.

3. Iterative Optimal Approach

We now consider choosing a consistent estimator of $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$. Any consistent estimator of $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$ would give the same asymptotic results, but non asymptotic behavior can be affected by the choice. A first consistent estimator of $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$ can be constructed by using its sample version $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$, substituting the usual estimates – $\hat{\boldsymbol{\Sigma}}, \bar{\mathbf{X}}, \bar{\mathbf{Y}}$, and $\hat{\boldsymbol{\beta}}$ – for the corresponding population quantities. Under the hypoth-

esis $d = m$, $\boldsymbol{\beta} = \boldsymbol{\eta}\boldsymbol{\gamma}$ with $\boldsymbol{\eta} \in \mathbb{R}^{p \times m}$ and thus $\varepsilon_i = Y_i - E(\mathbf{Y}_i) - \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\eta}^{\mathrm{T}}\{\mathbf{X} - E(\mathbf{X})\}$. Given an estimate $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ we can then estimate $\boldsymbol{\varepsilon}$ using the vector of residuals $\tilde{\varepsilon}_i$ from the OLS fit of Y_i on $\hat{\boldsymbol{\eta}}^{\mathrm{T}}\mathbf{X}$. Substituting $\tilde{\varepsilon}_i$ for ε_i, $\hat{\boldsymbol{\Sigma}}$ for $\boldsymbol{\Sigma}$, and $\bar{\mathbf{X}}$ for $E(\mathbf{X})$, leads to an alternative consistent estimate $\tilde{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ of $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$ and to a new estimate of $\boldsymbol{\eta}$. According to Carroll and Ruppert (1988), the number of such iterations should be two or more. Our simulation experience suggests that three iterations work well in the present context. We next present an algorithm for an iterative optimal method. During iteration, $\tilde{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ and the estimate of $\boldsymbol{\eta}$ are updated.

Algorithm for the iterative optimal approach

1. $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ is obtained as described above. Set $\tilde{\mathbf{V}}^0_{\hat{\boldsymbol{\beta}}} \leftarrow \hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$.

2. Estimate $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ by minimizing (2.4) with $\tilde{\mathbf{V}}^0_{\hat{\boldsymbol{\beta}}}$. Set $\hat{\boldsymbol{\eta}}_0 \leftarrow \hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\gamma}}_0 \leftarrow \hat{\boldsymbol{\gamma}}$.

3. Estimate $\tilde{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ using $\hat{\boldsymbol{\eta}}_0$ as an estimate of $\boldsymbol{\eta}$. Set $\tilde{\mathbf{V}}^1_{\hat{\boldsymbol{\beta}}} \leftarrow \tilde{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$.

4. Estimate $\boldsymbol{\eta}$ and $\boldsymbol{\gamma}$ by minimizing (2.4) with $\tilde{\mathbf{V}}^1_{\hat{\boldsymbol{\beta}}}$. Set $\hat{\boldsymbol{\eta}}_1 \leftarrow \hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\gamma}}_1 \leftarrow \hat{\boldsymbol{\gamma}}$.

5. Repeat steps 3 and 4 two more times.

Since this algorithm requires a specification of $d > 0$, it is available for the estimation of d and tests of the joint dimension-predictor and conditional predictor hypotheses.

4. Simulation and Case Studies

For illustration, we provide two examples; (1) a multivariate linear regression; (2) a regression with linear conditional means and heteroscedasticity. Example 2 mimicked what Cook and Setodji (2003, Table 4) used. In simulations, the total number of replication was 1000 and level $\alpha = 5\%$ was used for all tests. For notational conveniences, the proposed iterative optimal method and the Yoo-Cook method will be denoted as IOPT. and Y.C. respectively in tables and figures.

4.1 Example 1

We generate independently $\mathbf{X} = (X_1, \ldots, X_5)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_4)$ from standard normal distribution. And, $Y_k = \sum_{i=1}^{4} X_i + \varepsilon_k$ for $k = 1, \ldots, 4$. The hypothesis $H_0 : d = 0$ was always rejected by the two methods. In percentages of decision that $\hat{d} = 1$, with 100 sample sizes, the iterative optimal method (97.1%) outperformed the Yoo-Cook (67.3%), the former being slightly conservative in small

samples, while the latter is quite liberal. The two methods behaved similarly for the larger samples.

We next report simulation results from coordinate effect tests. We tested the statement of $E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{Y}|\mathbf{X}_{-i})$ jointly with $d = 1$, and conditionally given $d = 1$ for $i = 1, \ldots, 5$, where $\mathbf{X}_{-i}$ indicates a set of all predictors except the ith coordinate X_i. If the statement is not rejected, then it can be concluded that the ith coordinate X_i does not provide significant effect to $E(\mathbf{Y}|\mathbf{X})$. For $X_1, \ldots, X_4$, all hypotheses were rejected at least 99.7% of time with any sample sizes considered. For X_5, the results are summarized in Table 2. The results here reflect the characteristic behavior in our other simulations. The iterative joint and conditional predictor tests were better than the same tests by the Yoo-Cook method. The iterative conditional predictor test showed the best small sample behavior than the others. These differences in performance can be larger when there is curvature and heteroscedasticity in the simulation models. The iterative conditional predictor test was quite accurate even with relatively small samples.

Table 1: Estimated levels of 5% predictor effect tests for X_5 in Example 1; Joint dimension-predictor test with $d = 1$, Joint.; Conditional predictor test given $d = 1$, Cond.

	Y.C.		IOPT.	
n	Joint.	Cond.	Joint.	Cond.
100	34.8	14.7	2.7	5.5
200	18.2	8.4	4.8	4.8

4.2 Example 2

For Example 2, we generated V_1 and V_2 independently from t_6 and V_3, W_1, and W_2 independently from t_5. Define that $X_1 = W_1$, $X_2 = V_1 + W_2/2$, $X_3 = -V_1 + W_2/2$, $X_4 = V_2 + V_3$, and $X_5 = -V_2 + V_3$. Similar predictor configuration has been used in Cook and Setodji (2003) and Yoo and Cook (2007). This predictor configuration satisfies the linearity condition C1 for $\boldsymbol{\eta} = \{(1, 0, 0, 0, 0)^{\mathrm{T}}, (0, 1, 1, 0, 0)^{\mathrm{T}}\}$. Define that $Y_1 = 1 + X_1 + 0.5X_1\varepsilon_1$, $Y_2 = X_2 + X_3 + 0.5(X_2 + X_3)\varepsilon_2$, $Y_3 = 0.5X_3\varepsilon_3$, $Y_4 = 0.5X_4\varepsilon_4$ and $Y_5 = 0.5X_5\varepsilon_5$, with the ε_i's being independent standard normal random variables and independent of V_is and W_is.

Table 2: Estimated level of 5% tests of $d = 2$ for Example 2

n	Y.C.	IOPT.
100	17.5	3.3
400	13.3	3.9
800	9.9	4.5

Table 2 represents the qualitative behavior of the estimated size of the nominal 5% dimension test in our all of simulations with no highly skewed predictors. The Yoo-Cook method were generally liberal, and sometimes greatly so, while the iterative optimal method was observed to be slightly conservative with observed levels varying generally between 3 and 5%. A general message is that the Yoo-Cook method can be impacted easily by heteroscedasticity with smaller sample sizes. The small p-values for this method suggest that in such situations it will tend to overestimate d. Nevertheless, these results do not necessarily reflect the behavior of the three methods in estimating d, which is perhaps the primary use of tests on d.

Figure 1, which was constructed by computing the percentage of time that $\hat{d} = 2$, exhibits the characteristic performance of dimension estimation. A solid horizontal line in Figure 1 represent the true percentage of correct decisions, which converges to 95% as $n \to \infty$. With the smaller sample sizes the iterative optimal method underestimated the true dimension, but in brief, the iterative optimal method was always observed to yield the higher frequency of correct estimated $\hat{d} = d$.

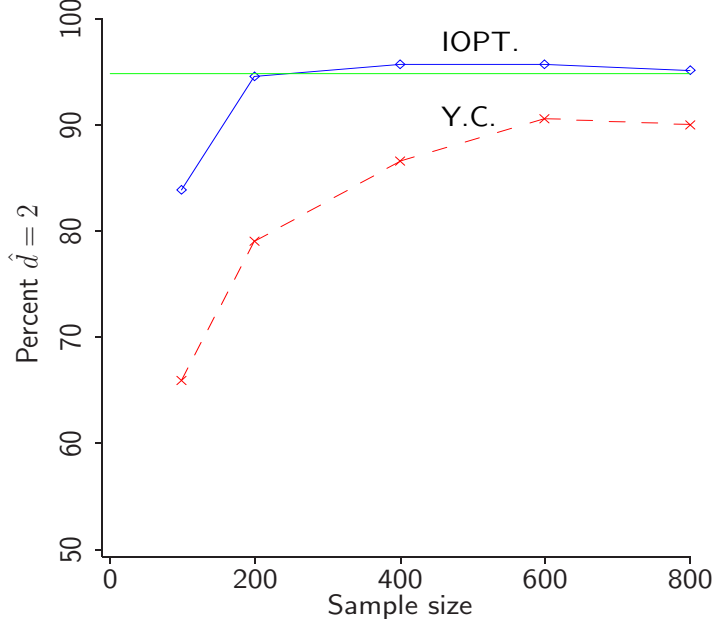Figure 1: Percentage of runs in which $\hat{d} = 2$ versus sample size for Example 2

4.3 Case study-Minneapolis school data

To illustrate a methodology introduced in the previous sections, we use data on the performance of students in $n = 63$ Minneapolis schools studied by Cook

(1998) and Cook and Setodji (2003). The $r = 4$ dimensional response vector $\mathbf{Y}$ consists of the percentages $P_{(\cdot)}$ of students in a school scoring above (A) and below (B) average on standardized fourth and sixth grade reading comprehension tests, $\mathbf{Y} = (P_{A4}, P_{B4}, P_{A6}, P_{B6})^{\mathrm{T}}$. Subtracting either pair of grade specific percentages from 100 gives the percentage of students scoring about average on the test. We used the same five predictors used by Cook and Setodji (2003): (1) the pupil teacher ratio (PT), and the square roots of (2) the percentage of children receiving Aid to Families with Dependent Children (AFDC), (3) the percentage of children not living with both biological parents (B), (4) the percentage of adults in the school area who completed high school (HS), (5) the percentage of persons in the area below the federal poverty level (PL).

With level 0.05 tests, the iterative optimal method gives $\hat{d} = 1$ (p-value 0.00 for $H_0 : d = 0$; p-value 0.15 for $H_0 : d = 1$), but the Yoo-Cook method decides $\hat{d} = 2$ (p-value 0.00 for $H_0 : d = 0$; p-value 0.02 for $H_0 : d = 1$; p-value 0.53 for $H_0 : d = 2$).

To gain useful information for deciding between $\hat{d} = 1$ and $\hat{d} = 2$, we construct a simulation as follows. Letting $\mathbf{X}_0 = \hat{\boldsymbol{\eta}}^{\mathrm{T}} \mathbf{X}$ be the estimated sufficient predictor from the iterative optimal method with $d = 1$, we generated new data sets from the model $Y_{k_i}^* = f_k(X_{0_i}) + \sigma_k \varepsilon_{k_i}$, $i = 1, 2, \ldots, 63$, $k = 1, 2, 3, 4$, where f_k is a LOWESS smooth of Y_k against X_0 using 0.7 as the tuning parameter, $\sigma_k^2 = 62^{-1} \sum_{i=1}^{63} \{Y_{k_i} - f_k(X_{0_i})\}^2$ and the ε_{k_i}'s are independent standard normal random variables. For each of 1000 data sets generated in this way, we tested the true null hypothesis $d = 1$ at nominal level 5% using each of the two methods under consideration. The rejection rates were 61.7% for the Yoo-Cook method and 5.6% for the iterative optimal method. Clearly, the Yoo-Cook method overestimates the true dimension, probably, because of the heteroscedasticity from a scatter plot matrix of the four responses and $\mathbf{X}_0$ (not shown), leading us to conclude that $d = 1$. Since the predictors in these data are not highly skewed and heteroscedasticity is present, the results agree with our previous observation that the iterative optimal method is best in such cases and generally performs well in smaller samples.

Next, we tested each coordinate effect conditionally given $d = 1$ using the iterative optimal method. The test results are given in Table 3. Row 2 in Table 3 indicates that $PL^{1/2}$ and PT contribute little to the mean function given the remaining predictors. However, as in linear regression, a sufficiently high correlation between $PL^{1/2}$ and PT might produce masking, and deleting either $PL^{1/2}$ or PT might cause the p-value for the remaining predictor to become significantly small. Row 3 in Table 3 gives the p-values for the same hypotheses after removing PT from the regression. Again we reach the same conclusion, and hence it can be determined that the mean function does depend on AFDC and

HS, and might also depend on B. Eliminating $PL^{1/2}$ and PT from the regression and marginally standardizing each of the remaining predictors to have a sample standard deviation of 1, the analysis might now be continued by plotting each of the four responses against the estimated sufficient predictor

$$\mathbf{X}_0 = 0.832AFDC^{1/2} - 0.382B^{1/2} - 0.403HS^{1/2}.$$

Table 3: p-values of the iterative conditional coordinate tests given $d = 1$.

	$AFDC^{1/2}$	$B^{1/2}$	$HS^{1/2}$	$PL^{1/2}$	PT
Conditional	0.00	0.06	0.01	0.64	0.49
Conditional (without PT)	0.00	0.11	0.00	0.48	–

5. Discussions

Throughout this article, we have considered no highly skewed predictors. With highly skewed predictors, simulation studies showed that the iterative optimal method often underestimated the true dimension of the multivariate conditional mean subspace with smaller samples, while the Yoo-Cook method was quite robust to skewness of the predictors, so it produced better results. With larger samples, however, the iterative optimal method often yielded the best performances. One remarkable result is that the iterative conditional predictor test showed the best performance among predictor effect tests regardless of predictor distributions.

In sufficient dimension reduction literature, predictors are often transformed for normality to satisfy linearity condition required. Therefore, the restriction of no highly skewed predictors in the proposed iterative optimal method does not seem heavy, and hence we expect that it has advantage over Yoo and Cook (2007) in many multivariate regression problems in practice. The code for the iterative optimal approach is available upon request from the author.

References

Carroll, R. J., and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, New York.

Cook, R. D. (1998). *Regression Graphics : Ideas for Studying Regressions Through Graphics.* Wiley, New York.

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

Cook, R. D. and Setodji, C. M. (2003), A model-free test for reduced rank in multivariate regression. *Journal of the American Statistical. Association* **98**, 340-351.

Li, K. C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association* **87**, 1025-1039.

Yoo, J.K. and Cook, R. D. (2007). Optimal sufficient dimension reduction for the conditional mean in multivariate regression. *Biometrika* **94**, 231-242.

Jae Keun Yoo
485 East Gray Street
Department of Bioinformatics and Biostatistics
School of Public Health and Information Sciences
University of Louisville
Louisville KY 40202 USA
peter.yoo@louisville.edu