DEPARTMENT OF STATISTICS
North Carolina State University
2501 Founders Drive, Campus Box 8203
Raleigh, NC 27695-8203

Institute of Statistics Mimeo Series No. 2607

SIR^3: Dimension Reduction in the Presence of Linearly or Nonlinearly Related Predictors

Lexin Li
Department of Statistics, North Carolina State University

R. Dennis Cook
School of Statistics, University of Minnesota

Christopher J. Nachtsheim
Carlson School of Management, University of Minnesota

First draft: May 14, 2004
Second draft: December 13, 2007

Abstract

Sufficient dimension reduction (SDR) is an effective tool for reducing high-dimensional predictor spaces in regression problems. SDR achieves dimension reduction without loss of any regression information and without the need to assume any particular parametric form of a model. This is particularly useful for high-dimensional applications such as data mining, marketing, and bioinformatics. However, most SDR methods require a linearity condition on the predictor distribution, and that restricts the applications of SDR. In this article, we propose a new SDR method, SIR^3, which does not require the linearity condition, and which we show to be effective when nonlinearly-related predictors are present. SIR^3 is an extension of a representative SDR method sliced inverse regression (SIR), and it is shown that SIR^3 reduces to SIR when the linearity condition holds. A simulation study and a real data application are presented to demonstrate the effectiveness of the proposed method.

KEY WORDS: Sufficient dimension reduction; Sliced inverse regression; The linearity condition.

1 Introduction

The goal of a regression of a univariate response y, which may be continuous or categorical, on a $p \times 1$ vector of predictor variables X is, in full generality, to develop inferences about the conditional distribution of $y \,|\, X$. Sufficient dimension reduction (SDR, Cook, 1998, Li, 2000) seeks to replace the p-dimensional vector of predictors X with q linear combinations $\eta^{\mathsf{T}} X$ (η is a $p \times q$ matrix, $q \leq p$), while preserving all information in $y \,|\, X$. The relevant information is preserved when $y \perp\!\!\!\perp X \,|\, \eta^{\mathsf{T}} X$, where $\perp\!\!\!\perp$ stands for independence. In practice, it is often true that q is far less than p, in which case a substantial reduction in dimension is achieved.

We first define a unique inferential object for SDR. A subspace $\mathcal{S}$ that satisfies $y \perp\!\!\!\perp X \,|\, P_{\mathcal{S}} X$ is called a dimension reduction subspace, where $P_{(\cdot)}$ denotes a projection operator with respect to the standard inner product. The intersection of all dimension reduction subspaces, under mild conditions (Cook, 1996), is a dimension reduction subspace itself. The resulting subspace is named the central subspace (CS), and is denoted by $\mathcal{S}_{y|X}$. The CS, by definition, is a unique and parsimonious subspace that preserves all information of $y \,|\, X$, and thus is the main object of interest in our dimension reduction inquiry. We will assume the existence of $\mathcal{S}_{y|X}$ throughout this article, and call its dimension, $d = \dim(\mathcal{S}_{y|X})$, the structural dimension of the regression.

There are a number of model-free methods for estimating $\mathcal{S}_{y|X}$, for instance, sliced inverse regression (SIR, Li, 1991), sliced average variance estimation (SAVE, Cook and Weisberg, 1991), and parametric inverse regression (Bura and Cook, 2002). Among them, SIR is perhaps the most popular method, and having proven successful in many real data applications. For instance, Naik, Hagerty, and Tsai (2000) introduced SIR to data-rich direct marketing environments, Chiaromonte and Martinelli (2002) applied SIR to tumor classification in microarray gene data, and Li and Li (2004) used SIR to predict patients' survival time given gene expressions.

SIR, as well as most other SDR methods, imposes no traditional assumption on $y \,|\, X$ such as $\mathrm{E}(y \,|\, X)$ being linear in X, or $y \,|\, X$ following a specified distribution. However, all of these methods require a *linearity condition* on the marginal distribution of X, i.e.,

$$\mathrm{E}(X \,|\, \eta^{\mathsf{T}} X = u) = A_0 + A_1 u, \tag{1}$$

where A_0 is a $p \times 1$ vector and A_1 is some $p \times q$ matrix. Simply stated, the linearity condition requires that the conditional mean $\mathrm{E}(X \,|\, \eta^{\mathsf{T}} X)$ be linear in $\eta^{\mathsf{T}} X$. As we will demonstrate below, strong nonlinearity among predictors can significantly degrade estimation accuracy of SIR and may produce completely

misleading results. Thus, in this article, we propose an extension of SIR which does not require the linearity condition, and we show that SIR is a special case of the proposed method when the linearity condition holds. This extension is one of the first such attempts to relax the linearity condition, and it also lays ground work for a number of future developments.

The rest of the article is organized as follows. Section 2 briefly reviews SIR, and provides an example to show how SIR fails when the linearity condition is not met. SIR^3, our extension of SIR, is proposed in Section 3, and the connection with SIR is established. Inference on the structural dimension associated with SIR^3 is addressed in Section 4. The effectiveness of the proposed method is demonstrated via a simulation study in Section 5, and the method is applied to the Los Alamos environmental contamination data in Section 6. Section 7 summarizes the paper and discusses the future research.

2 Sliced Inverse Regression

2.1 Review of SIR

To facilitate presentation, we work mostly in terms of the standardized predictors $Z = \Sigma_x^{-1/2}(X - \mathrm{E}(X))$, where $\Sigma_x = \mathrm{Cov}(X)$ is assumed to be positive definite. There is no loss of generality by working on the Z-scale, because $\mathcal{S}_{y|X} = \Sigma_x^{-1/2}\mathcal{S}_{y|Z}$ (Cook, 1998, Proposition 6.1). So $\mathcal{S}_{y|Z}$ can always be back-transformed to $\mathcal{S}_{y|X}$.

Sliced inverse regression was proposed by Li (1991) who approached sufficient dimension reduction through the inverse regression of Z on y. Define the inverse mean subspace

$$\mathcal{S}_{E(Z|y)} = \mathrm{Span}\{\mathrm{E}(Z \,|\, y) : y \in \Omega_y\},$$

where Ω_y denotes the sample space for response y.

Li (1991) and Cook (1998) showed, assuming the linearity condition (1),

$$\mathcal{S}_{E(Z|y)} \subseteq \mathcal{S}_{y|Z}.$$

It is often further assumed that $\mathcal{S}_{E(Z|y)}$ coincides with $\mathcal{S}_{y|Z}$, i.e., $\mathcal{S}_{E(Z|y)} = \mathcal{S}_{y|Z}$. This is called the coverage condition, which is often satisfied in practice. So we do not generally consider this condition to be worrisome. Since $\mathcal{S}_{E(Z|y)} = \text{Span}\{\text{Var}(\text{E}(Z \,|\, y))\}$ (Cook, 1998, Proposition 11.1), the eigenvectors of $M = \text{Var}(\text{E}(Z \,|\, y))$ corresponding to its non-zero eigenvalues form a basis for the central subspace accordingly. The matrix M is called the SIR kernel matrix.

To estimate M, we follow the SIR protocol by assuming that the response has been discretized by constructing h slices so that y takes values in $\{1, \ldots, h\}$. The j-th value of y is called the j-th slice. The number of slices h can be viewed as a tuning parameter in SIR. The choice of h usually does not affect the SIR estimate, as long as $h > d$ and n is large enough for the asymptotics to provide useful approximations. When the response is discrete, this slicing step may be unnecessary and h may simply take the number of classes in the response. The sample version of the SIR kernel matrix is then

$$\hat{M} = \sum_{j=1}^{h} \frac{n_j}{n} \, \hat{\text{E}}(Z \,|\, y = j) \, \hat{\text{E}}(Z \,|\, y = j)^{\intercal},$$

where n_j is the number of observations in the j-th slice, n is the total number of observations, and $\hat{\text{E}}(Z \,|\, y = j)$ is the average of Z in the j-th slice,

Letting $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ denote the eigenvalues of $\hat{M}$, inference on the structural dimension of the regression $d = \dim(\mathcal{S}_{y|Z})$ is based on the statistic

$$\hat{\Lambda}_d = n \sum_{j=d+1}^{p} \hat{\lambda}_j. \tag{2}$$

Estimation of d is obtained by performing a series of tests of hypotheses $d = m$ versus $d > m$ for $m = 0, \ldots, p-1$ (Li, 1991). The estimate of d is taken as the minimum m such that the null hypothesis $d = m$ is not rejected. Asymptotically

$\hat{\Lambda}_d$ is distributed as a linear combination of χ^2 random variables (Bura and Cook, 2002). Cook and Yin (2001) also developed a permutation test based on (2), which is particularly useful when the sample size is relatively small.

2.2 Linearity Condition

We next examine the linearity condition (1) more closely. First, it is important to note that, unlike parametric regression analysis, SIR does not impose any traditional assumption on the conditional distribution of $y \mid Z$. Instead, SIR places a constraint on the marginal distribution of predictors Z to assure the conditional mean $\mathrm{E}(Z \mid y)$ lies in the central subspace $\mathcal{S}_{y|Z}$.

Let columns of the $p \times d$ matrix γ be a basis for $\mathcal{S}_{y|Z}$, the linearity condition is equivalent to requiring that (Cook, 1998, Proposition 4.2),

$$\mathrm{E}(Z \mid P_\gamma Z) = P_\gamma Z, \tag{3}$$

where P_γ stands for a projection matrix onto the space spanned by columns of γ with respect to the standard inner product. Since γ is unknown in practice, we may require the linearity condition to hold for all possible subspaces of $\mathbb{R}^p$, which in turn is equivalent to requiring that the predictor distribution is elliptically symmetric (Eaton, 1986). The multivariate normal is a special case of an elliptically symmetric distribution.

Li (1991) argued that the linearity condition is not a severe restriction, because most low-dimensional projections of a high-dimensional data are close to normal (Diaconis and Freedman, 1984, Hall and Li, 1993). However, estimation via SIR may be impaired when there is a strong nonlinear relation among predictors. We demonstrate this in a simulation example below.

2.3 Illustration

Consider $p = 10$ independent standard normal predictors $X = (x_1, \ldots, x_{10})^\intercal$, with sample size $n = 100$. The response model is

$$y = \exp(0.5 \times \beta^\intercal X + 1) + \varepsilon,$$

where $\beta = (1, -1, -1, 0, \ldots, 0)^\intercal$, and ε is a standard normal error independent of X. In this example the central subspace is $\mathcal{S}_{y|X} = \text{Span}(\beta)$ with the structural dimension $d = 1$. Applying SIR to these data, we found that the p-values for testing the null hypotheses $d = 0$, $d = 1$, and $d = 2$ versus the alternatives that $d > 0$, $d > 1$, and $d > 2$, respectively, are $0.00, 0.40$ and 0.76. This strongly suggests that the estimated structural dimension $\hat{d} = 1$. The SIR estimate of β is $\hat{\beta} = (0.57, -0.59, -0.56, 0.05, 0.04, -0.02, 0.01, 0.08, 0.01, -0.01)^\intercal$, and the angle between β and $\hat{\beta}$ is $6.21°$. For reference, if we use a randomly generated vector as the estimate of β, the angle between this random estimate and the true direction is expected to be $74.64°$. SIR successfully uncovers both the structural dimension and the vector spanning the central subspace in this example where the linearity condition holds.

Next we introduce a nonlinear relation between the first two predictors by letting x_1 follow a uniform distribution between 0 and 1, and defining

$$x_2 = \log(x_1) + e, \text{ where } e \sim \text{Uniform}(-0.3, 0.3).$$

The remaining predictors $x_3, \ldots, x_{10}$ and the error ε follow the standard normal distribution as before, and the model and sample size are unchanged. In this instance, the linearity condition no longer holds.

Applying SIR to these data, the p-values for testing the null hypotheses $d = 0$, $d = 1$, and $d = 2$ are $0.00, 0.06$ and 0.42 respectively. It is not clear from these results whether the estimated structural dimension should be $\hat{d} = 1$ or $\hat{d} = 2$ because the p-value for hypotheses $d = 1$ versus $d > 1$ is marginal

at 0.06. Suppose we conclude correctly that $\hat{d} = 1$, the SIR estimate now becomes $(0.02, -0.53, -0.84, 0.00, -0.01, 0.02, -0.03, 0.04, -0.01, -0.03)^\intercal$. The angle between this estimate and β is 36.30°. The example demonstrates that a nonlinear relation among the predictors degrade both inference on d and estimation accuracy of vectors in $\mathcal{S}_{y|X}$. In the next section, we will develop an estimation approach that is unrestricted by the presence of nonlinearly related predictors.

3 SIR^3: Extension of SIR

3.1 Development

Given the structural dimension d, we propose to minimize the following objective function

$$L_d(B) = \mathrm{E}\left(||\mathrm{E}(Z \,|\, y) - \mathrm{E}(\mathrm{E}(Z \,|\, B^\intercal Z) \,|\, y)||^2\right) \tag{4}$$

over all $p \times d$ orthonormal matrices B in the Stiefel manifold (Murihead, 1982, p.67). Letting γ denote a basis for $\mathcal{S}_{y|Z}$, we then note that γ is a population minimizer of (4). This is true because, by definition, $y \perp\!\!\!\perp Z \,|\, \gamma^\intercal Z$, which in turn implies that $Z \,|\, (y, \gamma^\intercal Z) \stackrel{d}{=} Z \,|\, \gamma^\intercal Z$, where $\stackrel{d}{=}$ denotes the two distributions are identical. Consequently,

$$\mathrm{E}(Z \,|\, y) = \mathrm{E}(\mathrm{E}(Z \,|\, y, \gamma^\intercal Z) \,|\, y) = \mathrm{E}(\mathrm{E}(Z \,|\, \gamma^\intercal Z) \,|\, y). \tag{5}$$

Moreover, our experiences from simulations suggest that the minimizer of (4) generally corresponds to a basis of the central subspace. Therefore we propose to estimate $\mathcal{S}_{y|Z}$ by solving the optimization problem (4).

Given n observations $\{(X_1, y_1), \ldots, (X_n, y_n)\}$, let $\bar{X}$ and $\hat{\Sigma}_x$ denote the sample versions of $\mathrm{E}(X)$ and $\mathrm{Cov}(X)$ respectively. The standardized predictor is $Z_i = \hat{\Sigma}_x^{-1/2}(X_i - \bar{X})$, $i = 1, \ldots, n$. The sample version of objective $L_d(B)$ is

then

$$\hat{L}_d(B) = \sum_{j=1}^{h} n_j \, ||\frac{1}{n_j} \sum_{i=1}^{n_j} Z_i - \frac{1}{n_j} \sum_{i=1}^{n_j} \hat{\mathrm{E}}(Z \,|\, B^{\mathsf{T}} Z_i)||^2 \tag{6}$$

where $\hat{\mathrm{E}}(Z \,|\, B^{\mathsf{T}} Z_i)$ denotes the estimate of $\mathrm{E}(Z \,|\, B^{\mathsf{T}} Z)$ that are evaluated at the observations $Z_1, \ldots, Z_{n_j}$ within the j-th slice $y = j$.

An estimate of $\mathrm{E}(Z \,|\, B^{\mathsf{T}} Z)$ can be obtained by smoothing the individual components of Z over the d-dimensional covariates $B^{\mathsf{T}} Z$. We first note that the choice of a specific smoother is flexible. In our studies, we have tried both a local loess smoother (Cleveland and Devlin, 1988) and global polynomial smoothers of varying orders. For the latter, given the order q of the global polynomial smoother, for a univariate variable v and d covariates $u_1, \ldots, u_d$, we estimate $\mathrm{E}(v \,|\, u_1, \ldots, u_d)$ by

$$\beta_0 + \sum_{i_1} \beta_{i_1} \, u_{i_1} + \sum_{i_1, i_2} \beta_{i_1 i_2} \, u_{i_1} u_{i_2} + \ldots + \sum_{i_1, \ldots, i_q} \beta_{i_1 \ldots i_q} \, u_{i_1} \ldots u_{i_q}$$

where $i_1, \ldots, i_q$ are indices taking values in $\{1, \ldots, d\}$, and the β's are obtained by the least squares. Our extensive simulations have suggested that a third order global polynomial smoother, i.e., $q = 3$, often yields satisfactory empirical results. For this reason, we recommend to choose $q = 3$ in practice, and we call the resulting estimation method as SIR^3. For a discussion on the order of polynomial smoother in a related context, see also Yin and Cook (2002).

Given the nonlinear nature of both objective function and the constraint that B being orthonormal, a sequential quadratic programming (SQP) algorithm (Gill, Murray, and Wright, 1981) is appropriate for the minimization of $\hat{L}_d(B)$. The SQP algorithm *NPSOL*, released by the System Optimization Laboratory of Stanford University, is employed. Details of usage of *NPSOL* can be found in Li (2003) and Gill, Murray, Saunders, and Wright (1986). The matrix that minimizes $\hat{L}_d(B)$ is taken as the estimate $\hat{\gamma}$ of γ, the basis for the central subspace $\mathcal{S}_{y|Z}$. To obtain an estimated basis for $\mathcal{S}_{y|X}$ in the original X-scale, we transform $\hat{\gamma}$ to $\hat{\Sigma}_x^{-1/2} \hat{\gamma}$.

In the above procedure, the structural dimension d is assumed known. Inference on d associated with this estimation method via the permutation test will be discussed in detail in Section 4.

We applied SIR^3 to the example in Section 2.3 where the linearity condition does not hold. The p-values of the permutation test of SIR^3 for the null hypotheses $d = 0$, $d = 1$, and $d = 2$ are $0.00, 0.41$ and 0.78 respectively, which suggests that the structural dimension is 1. The SIR^3 estimate is $(0.75, -0.53, -0.39, -0.01, -0.02, 0.02, 0.05, -0.03, -0.01, 0.00)^\intercal$, with angle to the true direction equal to $15.44°$. It is clear that substantial improvement has been achieved compared with SIR in this example.

3.2 Relation to SIR

We next examine the connections between SIR^3 and SIR. For the basis γ of $\mathcal{S}_{y|Z}$, from (5) we know γ is the population minimizer of $L_d(B)$ in (4). If the linearity condition that $\mathrm{E}(Z \,|\, P_\gamma Z) = P_\gamma Z$ holds, we have

$$\mathrm{E}(Z \,|\, y) = \mathrm{E}(\mathrm{E}(Z \,|\, P_\gamma Z) \,|\, y) = \mathrm{E}(P_\gamma Z \,|\, y) = P_\gamma \,\mathrm{E}(Z \,|\, y).$$

That is, the projection of $\mathrm{E}(Z \,|\, y)$ to the space spanned by columns of γ is $\mathrm{E}(Z \,|\, y)$ itself, therefore, $\mathrm{E}(Z \,|\, y) \in \mathcal{S}_{y|Z}$. So SIR^3 produces exactly the same population solution as SIR when the linearity condition holds.

Furthermore, the linearity condition (1) implies that $\mathrm{E}(Z \,|\, \gamma^\intercal Z)$ can be well approximated by a global polynomial smoother with order $q = 1$. If the marginal distribution of X satisfies the linearity condition, we expect the coefficients associated with the polynomial smoother with order greater than one will be close to 0, thus SIR^3 should give solutions similar to SIR. This will be shown using simulation. On the other hand, if there is a nonlinear relation among $\mathrm{E}(Z \,|\, \gamma^\intercal Z)$, the polynomial smoother with order greater than one will then capture the nonlinearity and yield better estimates than SIR. In this sense, the global third order polynomial smoother provides a natural relaxation of the linearity condition.

3.3 Handling of Local Optima

Numerical optimization of objective function $\hat{L}_d(B)$ may be complicated by the presence of multiple local optima when the predictors are nonlinearly related. We consider the following response model as an illustration. The sample size is $n = 200$. The structural dimension is $d = 2$, and two basis vectors of the central subspace are $(1, 0, 0, 0, 0)^\tau$ and $(0, 0, 1, 0, 0)^\tau$.

Response model 1:

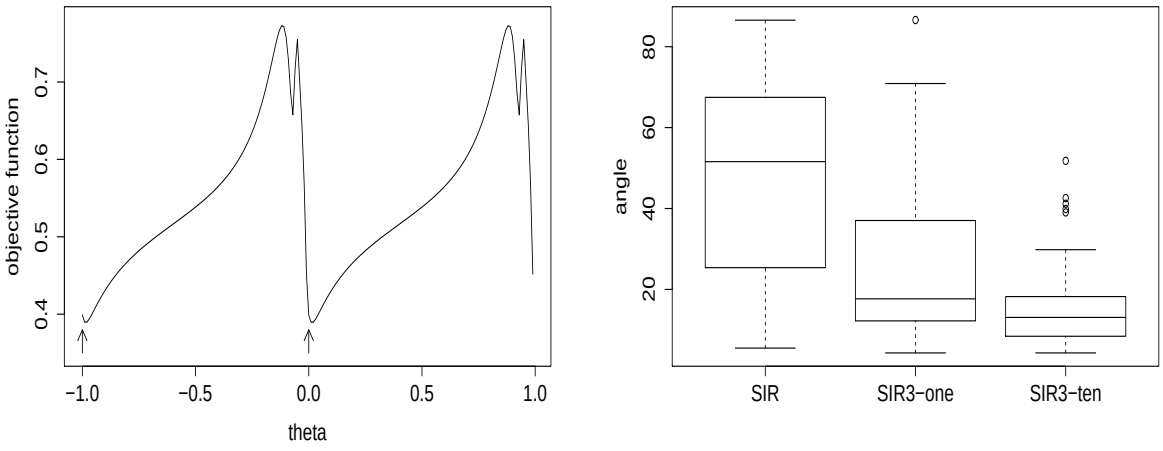
$x_1 \sim \text{Uniform}(0, 1)$, $x_2 = \log(x_1) + e$, where $e \sim \text{Uniform}(-0.3, 0.3)$,

$x_3, x_4, x_5 \sim \text{Normal}(0, 1)$,

$y = \log(x_1)^2 + x_3^3 + \varepsilon$, where $\varepsilon \sim \text{Normal}(0, 0.2^2)$.

To examine the nature of objective function, we created a two-dimensional view of the objective function along a fixed direction. Figure 1(a) shows the evaluated objective function $\hat{L}_d(B)$ as B moves along $(\cos(\theta\pi), \sin(\theta\pi), 0, 0, 0)^\tau$ with $-1.0 \leq \theta < 1.0$. One basis direction of the central subspace $(1, 0, 0, 0, 0)^\tau$ is marked out by arrows in the plot. It shows that the global minimum of objective function agrees with the basis vector of the central subspace; meanwhile, there exist local non-global minima. In our extensive simulation study of examples where there is nonlinear relation among predictors, we have observed that, as the number of predictors increases, the objective tends to have more local minima. On the other hand, as the sample size increases, the number of local optima appears to decrease (Li, 2003).

Nevertheless, the existence of multiple local optima suggests that multiple searches, each from a different randomly selected starting point, or the use of a global minimizer such as simulated annealing, should be considered in an effort to identify the global minimum. We have examined many examples, and found that the multiple searches with ten starting values often yield satisfactory results. Consider response model 1, Figure 1(b) shows the box plot of angles between the true basis and the estimate produced from SIR, SIR^3 with a single



(a) Objective function plot (b) Box plot

Figure 1: Response model 1: (a) Objective function $\hat{L}_d(B)$ versus $(\cos(\theta\pi), \sin(\theta\pi), 0, 0, 0)^\mathsf{T}$ with $-1.0 \leq \theta < 1.0$; (b) Box plot of angles out of 50 data replications. Three boxes corresponds to estimates of SIR, SIR^3 with a single starting value, and SIR^3 with ten starting values.

starting value, and SIR^3 with ten starting values respectively. Improvement in estimation accuracy from multiple starting values is clearly seen in the plot. For instance, the median of angles for SIR is 51.58°. It improves to 17.67° for SIR^3 with a single starting value, and it further drops to 13.08° for SIR^3 with ten starting values. In what follows, we will always employ ten starting values in our simulations.

4 Test for Structural Dimension

We next address the issue of inference on the structural dimension d of $\mathcal{S}_{y|Z}$. When the linearity condition does not hold, neither an asymptotic test nor a permutation test based on SIR works properly. One such example was shown in Section 2.3. We next propose a permutation test based on SIR^3, which can

estimate the structural dimension d without requiring the linearity condition.

Given the hypothesized dimension m, B is a $p \times m$ orthonormal matrix. Denote the sample objective function as shown in (6) by $\hat{L}_m(B)$. Let $\hat{B}_m$ be the matrix which minimizes $\hat{L}_m(B)$, and $\hat{L}_m$ be the minimum objective function value obtained at $B = \hat{B}_m$. For a test of the hypotheses $d = m$ versus $d > m$, we consider the following test statistic

$$\hat{\Lambda}_m^E = \hat{L}_m - \hat{L}_p = \hat{L}_m$$

where the last equality comes from the fact that $\hat{B}_p = I_p$, the $p \times p$ identity matrix, and $\hat{L}_p = 0$. We then propose the permutation test procedure based on $\hat{\Lambda}_m^E$ as shown in Table 1. To estimate the structural dimension d, we repeat the above procedure for $m = 0, \ldots, p-1$. The estimated dimension $\hat{d}$ is the smallest integer that the null hypothesis $d = \hat{d}$ can not be rejected.

It is informative to investigate the link between the proposed test statistic $\hat{\Lambda}_m^E$ and the SIR-based statistic $\hat{\Lambda}_m$ as shown in (2). Assuming the linearity condition is satisfied for any orthonormal matrix B, i.e., $\mathrm{E}(Z \mid B^\intercal Z) = P_B Z = BB^\intercal Z$, one can show that

$$\begin{aligned}
\hat{L}_m(B) &= \sum_{j=1}^{h} n_j \, || \, \hat{\mathrm{E}}(Z \mid y) - \hat{\mathrm{E}}(\, \hat{\mathrm{E}}(Z \mid B^\intercal Z) \mid y \,) \, ||^2 \\
&= \sum_{j=1}^{h} n_j \, || \, \hat{\mathrm{E}}(Z \mid y) - P_B \, \hat{\mathrm{E}}(Z \mid y) \, ||^2 \\
&= n \sum_{j=1}^{h} \frac{n_j}{n} \, \hat{\mathrm{E}}(Z \mid y)^\intercal \, \hat{\mathrm{E}}(Z \mid y) - n \sum_{j=1}^{h} \frac{n_j}{n} \, \hat{\mathrm{E}}(Z \mid y)^\intercal \, BB^\intercal \, \hat{\mathrm{E}}(Z \mid y) \\
&= n \, \mathrm{trace}(\hat{M}) - n \, \mathrm{trace}(B^\intercal \hat{M} B),
\end{aligned}$$

where $\hat{M}$ is the sample SIR kernel matrix. Thus, to minimize $\hat{L}_m(B)$, it is sufficient to maximize $\mathrm{trace}(B^\intercal \hat{M} B)$ subject to the constraint $B^\intercal B = I_m$. This maximum value is equal to $\sum_{j=1}^{m} \hat{\lambda}_j$, where $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_m$ are the largest m

1. Compute $\hat{B}_m = \arg\min_B \hat{L}_m(B) = (\hat{b}_1, \ldots, \hat{b}_m)$ based on original data y_i and Z_i, $i = 1, \ldots, n$. Obtain the test statistic $\hat{\Lambda}_m^E$.

2. Construct the complementary basis $(\hat{b}_{m+1}, \ldots, \hat{b}_p)$ of $\hat{B}_m$ in $\mathbb{R}^p$, and construct vectors $\hat{V}_i = (\hat{b}_1^\intercal Z_i, \ldots, \hat{b}_m^\intercal Z_i)$, and $\hat{V}_i^c = (\hat{b}_{m+1}^\intercal Z_i, \ldots, \hat{b}_p^\intercal Z_i)$, for $i = 1, \ldots, n$.

3. Randomly permute the indices i of the $\hat{V}_i^c$ to obtain the permuted set $\hat{V}_i^{c*}$. Form the permuted data $Z_i^* = (\hat{V}_i, \hat{V}_i^{c*})^\intercal$.

4. Compute $\hat{B}_m^* = \arg\min_B \hat{L}_m(B)$ based on the original data y_i along with the permuted data Z_i^*. Obtain the test statistic $\hat{\Lambda}_m^{E*}$.

5. Repeat steps 3 - 4 N times, where N is the number of permutations. The p-value of the hypothesis testing is the fraction of $\hat{\Lambda}_m^{E*}$ that exceed $\hat{\Lambda}_m^E$.

Table 1: Permutation test for SIR^3.

eigenvalues of $\hat{M}$. Therefore,

$$\hat{\Lambda}_m^E = n \sum_{j=1}^{p} \hat{\lambda}_j - n \sum_{j=1}^{m} \hat{\lambda}_j = n \sum_{j=m+1}^{p} \hat{\lambda}_j = \hat{\Lambda}_m$$

That is, the proposed test statistic equals the SIR statistic when the linearity condition holds.

A power study of the permutation test is carried out in Section 5.3.

5 Performance of SIR^3

5.1 Nonlinear Predictors

We first compare SIR and SIR^3 in a number of examples where the linearity condition does not hold. The box plot of angles between the true and the

estimated basis based on 50 data replications will be used as a measure for estimation accuracy. In addition, we compute the *percentage reduction* resulting from the use of SIR^3 as $\Delta_{pr} = 100\,(a_{\text{SIR}} - a_{\text{SIR}^3})/a_{\text{SIR}}$, where a_{SIR} and a_{SIR^3} are the median angle of SIR and SIR^3 respectively.

In addition to response model 1 in Section 3.3, we consider the following models.

Response model 2:

$x_1 \sim \text{Uniform}(0, 1)$, $x_2 = \log(x_1) + e$, where $e \sim \text{Uniform}(-0.3, 0.3)$,

$x_3, x_4, x_5 \sim \text{Normal}(0, 1)$,

$y = \exp(0.5\,(x_1 - x_2 - x_3) + 1) + \varepsilon$, where $\varepsilon \sim \text{Normal}(0, 1)$.

Response model 3:

$x_1, \ldots, x_8 \sim \text{Exponential}(1)$,

$y = \exp(x_1 - x_2) + \varepsilon$, where $\varepsilon \sim \text{Normal}(0, 0.5^2)$.

Response model 4:

$x_1 \sim \text{Normal}(0, 1)$, $x_2 = x_1^2$, $x_3, \ldots, x_8 \sim \text{Normal}(0, 1)$,

$\varepsilon_1, \varepsilon_2 \sim \text{Normal}(0, 1)$,

$y = \text{sign}(x_1 + x_2 + x_3 + x_4 + 0.1\,\varepsilon_1) \times \log(|x_1 - x_2 + x_7 - x_8 + 5 + 0.1\,\varepsilon_2|)$.

For models 1 and 4, the structural dimension $d = 2$, and the sample size is set as $n = 200$. For models 2 and 3, $d = 1$, and $n = 100$. The linearity condition does not hold in models 1, 2, and 4 because of the nonlinear relation between the first two predictors x_1 and x_2. The linearity condition is not met in model 3 because $\text{E}(X \mid \eta^\intercal X = u)$ is nonlinear in u, where $\eta = (1, -1, 0, \ldots, 0)^\intercal$ denotes the basis of corresponding central subspace. Figure 2 shows the box plots of angles for the four response models, and Table 2 summarizes the median angles as well as the percentage reduction Δ_{pr} resulting from the use of SIR^3. It is clear that SIR^3 brings considerable amount of improvement in estimation accuracy over SIR.

Method	Model 1	Model 2	Model 3	Model 4
SIR	51.58°	27.60°	18.14°	30.09°
SIR^3	13.08°	12.76°	11.77°	17.11°
Δ_{pr}	74.64%	53.77%	35.12%	43.14%

Table 2: Comparison of SIR^3 and SIR. Listed are the median of angles of SIR and SIR^3 estimates based on 50 data replications, and the percentage reduction Δ_{pr} resulting from the use of SIR^3.

For instance, the median angle produced by SIR is 27.60° in response model 2, whereas the median angle drops to 12.76° for SIR^3, resulting in the percentage reduction 53.77%.

5.2 Linear Predictors

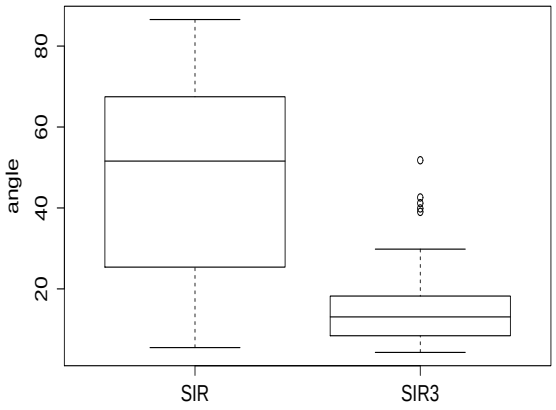
We next examine the performance of SIR^3 when the linearity condition holds. The following response model is examined, with the sample size $n = 100$.

Response model 5:

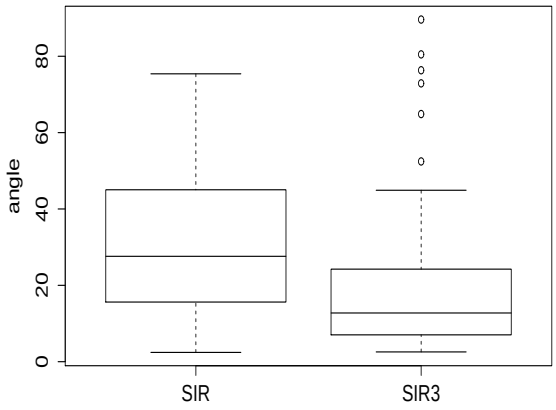
$x_1, x_2 \sim \text{Normal}(0, 1)$,

$y = \exp(0.5x_1) + \varepsilon$, where $\varepsilon \sim \text{Normal}(0, 1)$.

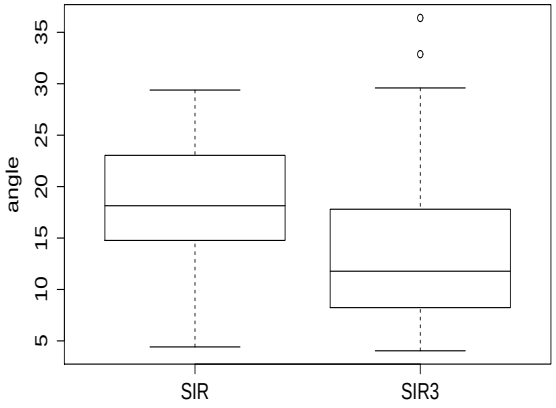
Since the predictors follow a normal distribution, the linearity condition holds. Figure 3 shows the box plot of angles based on 50 replications. In addition to SIR^3, where the order of the polynomial smoother is three, we also compared the method of minimizing $\hat{L}_d(B)$ with a linear polynomial smoother (a method which may be called SIR^1). We first observed that our extension of SIR with a polynomial smoother of order $q = 1$ essentially produces the same results as SIR. This agrees with our expectation. In addition, SIR^3 yields estimates close to SIR when the linearity condition is met. The computation involved in SIR^3, however, is inevitably more expensive than SIR.



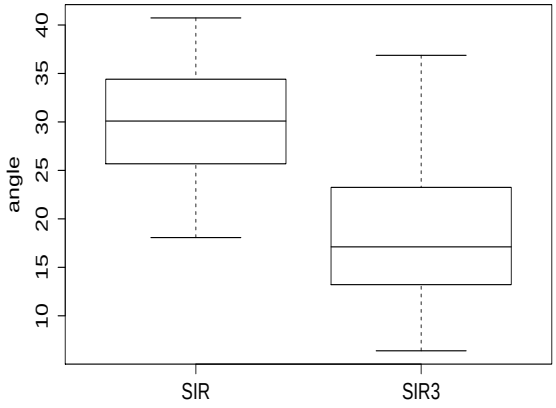
(a) Model 1



(b) Model 2



(c) Model 3



(d) Model 4

Figure 2: Comparison of SIR^3 and SIR. Box plots of angles for response models 1 - 4, based on 50 data replications, where the linearity condition is not satisfied.

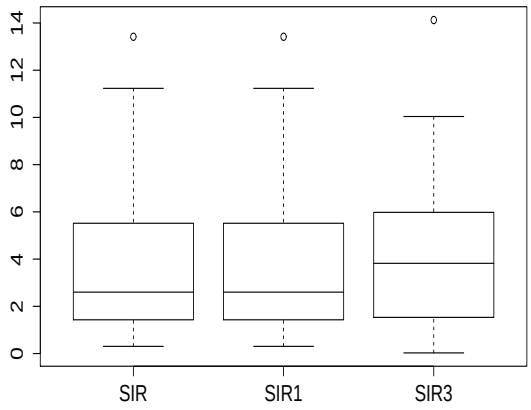


Figure 3: Comparison of SIR^3 and SIR. Box plot of angles for response model 5, based on 50 data replications, where the linearity condition is satisfied.

5.3 Power Study of Dimension Test

In this section we summarize the results of a power study of the proposed test for inference on d. We examined three response models that we have considered before: response model 3 with $d = 1$ and model 4 with $d = 2$, both of which do not meet the linearity condition; and response model 5 where $d = 1$, and the linearity condition holds. For each example, 1000 replications are produced. The p-values of both SIR-based permutation test and our proposed SIR^3-based test are computed. Given the type I error rate α, we report the observed rejection rate, i.e., the ratio of number of times that the p-value of the test is less than α out of 1000 replications. For a problem with a structural dimension d and the hypotheses $d = m$ versus $d > m$, this rejection rate is the estimated power of the test for $m < d$, and the estimated type I error rate for $m = d$. Table 3 summarizes the results.

From the table, we observe that, when the linearity condition holds, our proposed SIR^3-based permutation test has comparable power and type I error

		$d = 0$ vs $d > 0$		$d = 1$ vs $d > 1$		$d = 2$ vs $d > 2$	
		$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$
Model 3	SIR	1.000	1.000	0.494	0.634	-	-
$(d = 1)$	SIR^3	1.000	1.000	0.058	0.108	-	-
Model 4	SIR	1.000	1.000	1.000	1.000	0.261	0.393
$(d = 2)$	SIR^3	1.000	1.000	1.000	1.000	0.047	0.085
Model 5	SIR	1.000	1.000	0.057	0.115	-	-
$(d = 1)$	SIR^3	1.000	1.000	0.059	0.119	-	-

Table 3: Rejection rates per 1000 Monte Carlo replications of the SIR-based permutation test and the SIR^3-based permutation test.

rate as the SIR-based test. For instance, for hypotheses $d = 1$ versus $d > 1$ in model 5, SIR^3-based test produced Type I error rate of 0.059, compared with 0.057 by SIR, when $\alpha = 0.05$. On the other hand, when the linearity condition is not met, SIR-based permutation test yields type I error rate far from the nominal level. For instance in model 3, with $\alpha = 0.05$, the error rate for SIR is 0.494. In contrast, the error rate for SIR^3 is 0.058, which is consistent with the nominal level.

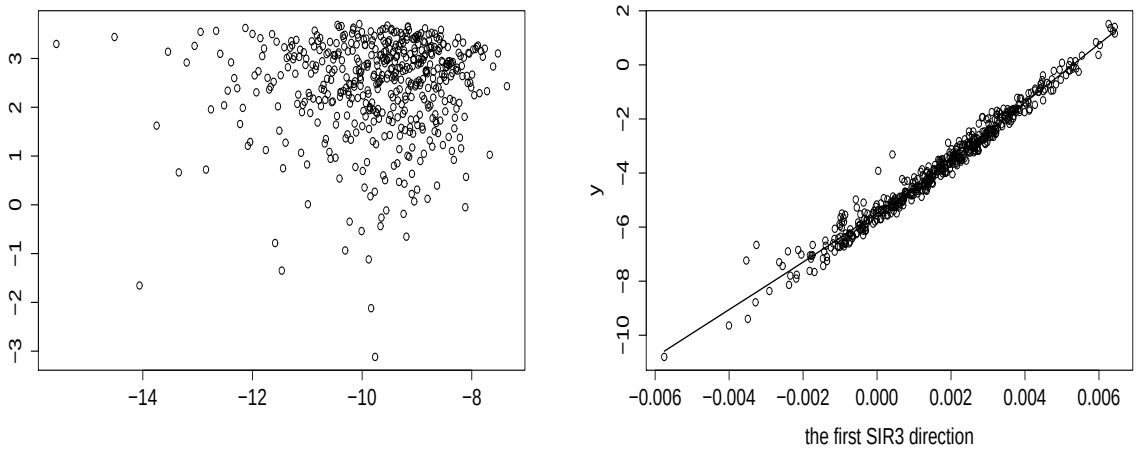
6 Application: Los Alamos Environmental Contamination Data

We consider data from Los Alamos National Laboratory on the fate of an environmental contaminant introduced into an ecosystem. The response y is the logarithm of the amount of contamination in the terrestrial invertebrates at day 5000, and we use the 16 active predictors identified by Cook (1994). We took the

logarithm of the positive predictors with ratio between the largest and smallest values larger than 10, leaving 7 transformed predictors. The data contain $n = 501$ samples.

Verification of the linearity condition for the 16 variables is not a simple task. Partial information can be obtained by inspecting scatter plots for predictor pairs, although there are 120 such plots to examine here. For example, Figure 4(a) provides a plot of two log-transformed predictors, which seems to exhibit some nonlinearity. In any event, it is not clear if the linearity condition holds for the data, and this adds uncertainty to inference from SIR. The proposed SIR^3 method, however, frees us from the restriction of the linearity condition. As a comparison, we applied both SIR and SIR^3 to these data. The p-values of the permutation tests of SIR^3 are 0.00, 0.26, and 0.33 for $d = 0, 1,$ and 2 respectively, while the p-values of SIR tests are 0.00, 0.05, and 0.52. So SIR^3 suggests that the structural dimension of regression is $d = 1$, but SIR concludes that $d = 2$. Furthermore, the first SIR^3 direction agrees exactly with the first direction obtained by SIR, although such a similarity between the estimated directions is irrelevant unless the estimated structural dimensions are the same. These results imply that, for the environmental contamination data, SIR yields a 2D subspace that contains the 1D SIR^3 subspace. It is then essential to check the validity of tests of SIR^3 and SIR in order to infer the true structural dimension d.

For this purpose, we simulated a one-dimensional model based on the environmental contamination data. Specifically, let $\hat{\beta}$ denote the extracted first direction by both SIR^3 and SIR, and let X denote the vector of 16 predictors. Figure 4(b) shows a plot of the response y versus the extracted component $\hat{\beta}^\intercal X$. We first fitted a local linear smoother to the data in Figure 4(b), resulting in a smooth function $s(\hat{\beta}^\intercal X)$ of $\hat{\beta}^\intercal X$. We next generated new response from the model $y_{new} = s(\hat{\beta}^\intercal X) + 0.2\varepsilon$, where the same predictor values of X are used, and



(a) Two log-transformed predictors (b) y versus the first SIR^3 direction

Figure 4: Analysis of environmental contamination data. (a) Marginal plot of two log-transformed predictors; (b) Response versus the first SIR^3 direction; the solid line is the local linear smoother with bandwidth 0.75.

ε is a standard normal error independent of X. The structural dimension of the simulated data is $d = 1$. SIR and SIR^3 were then applied to the new data, with the sample size $n = 501$. The p-values for the SIR tests are 0.00, 0.06, and 0.51, while the p-values for SIR^3 are 0.00, 0.35, and 0.63. Therefore we have strong evidence showing that SIR concludes the structural dimension larger than the true value, while SIR^3 provides a more reliable estimate of d. This result echoes the general qualitative patterns observed in simulations, indicating that SIR is responding to the nonlinearity in the predictors.

In summary, SIR^3 can serve as a diagnostic when there is uncertainty about inference from SIR. It provides an effective replacement for SIR without the burden of checking the linearity condition. This is particularly useful for high-dimensional data, where checking the linearity can be difficult.

7 Conclusion

Sliced inverse regression has proven to be an effective dimension reduction tool for high-dimensional data analysis as well as data visualization. However, the intrinsic linearity condition imposed by SIR, and in fact by most existing sufficient dimension reduction methods, has largely impeded its applications to many regression problems where the predictors are nonlinearly related. In this article, we have proposed SIR^3, an extension of SIR, that permits both linearly and nonlinear related predictors. Our simulation results demonstrated that the performance of SIR^3 is comparable to SIR when the linearity condition holds, and is superior to SIR when the linearity condition is not satisfied. Additionally, SIR^3 was shown to reduce to SIR in the population level given the linearity condition.

The idea to extend SIR in this article is useful and may be applied to other SDR estimation approaches, e.g. SAVE, to reduce the dependence on the linearity condition. Work alone this line is currently under investigation.

References

Bura, E., and Cook, R.D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of Royal Statistical Society*, Series B, **63**, 393-410.

Chiaromonte, F., and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, **176**, 123-144.

Cleveland, W.S., and Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**, 596-610.

Cook, R.D, (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. *Proceedings of the Section on Physical Engineering Sciences*, pp.18-25. Alexandria, Virginia, The American Statistical Association.

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

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

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

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

Cook, R.D., and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley, New York.

Cook, R.D., and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis. *Australian and New Zealand Journal of Statistics*, **43**, 147-177.

Diaconis, P., and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics*, **12**, 793-815.

Eaton, M. (1983). *Multivariate Statistics: A Vector Space Approach*. Wiley, New York.

Eaton, M. (1986). A characterization of spherical distributions. *Journal of Multivariate Analysis*. **20**, 272-276.

Gill, P.E., Murray, W., and Wright, M.H. (1981). *Practical optimization*. Academic Press, New York.

Gill, P.E., Murray, W., Saunders, M.A., and Wright, M.H. (1998). User's guide for NPSOL 5.0: a Fortran package for nonlinear programming. Stanford Systems Optimization Laboratory Software, Inc.

Hall, P., and Li, K.C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Annals of Statistics*, **21**, 867-889.

Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316-327.

Li, L. (2003). Sufficient dimension reduction in high-dimensional data. Ph.D. Dissertation, School of Statistics, University of Minnesota.

Li, L., and Li, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20**, 3406-3412.

Murihead, R. (1982). *Aspects of multivariate statistical theory.* Wiley, New York.

Naik, P., Hagerty, M.R., and Tsai, C-L. (2000). A new dimension reduction approach for data-rich marketing environments: sliced inverse regression. *Journal of Marketing Research*, **37**, 88-101.

Yin, X., and Cook, R.D. (2002). Dimension reduction for the conditional kth moment in regression. *Journal of Royal Statistical Society, Series B.* **64**, 159-175.