Using Arc for Dimension Reduction and Graphical Exploration in Regression

R. Dennis Cook*

June 15, 2000

Abstract

Abstract here

Key Words: Central subspaces, Dimension reduction; Principal Hessian Directions; Regression; Regression graphics, Sliced inverse regression; Sliced average variance estimation.

1 Introduction

Graphical displays have always played a key role in regression studies, particularly when the number of predictors is small. For instance, in the case of a simple regession with a single predictor, a scatterplot of the response y versus the predictor x can provide useful information about curvature in the mean function, heterscedasticity and plausible models. Identifying the main pattern in the plot while simultaneously recognizing gross deviations from that pattern is often easy, and so visually finding outliers, influential cases and regression mixtures is possible without the need to pre-specify a parametric model. These ideas are easily generalized to problems with two predictors when viewed in a rotating three-dimensional (3D) plot (Cook and Weisberg, 1999, Ch. 18). Useful graphical displays of data from regessions with three predictors can be obtained by first relacing the response y with a discrete version $\tilde{y}$ constructed by partitioning its range, and then assigning

*School of Statistics, 1994 Buford Ave., University of Minnesota, St. Paul, MN 55108. This paper was supported in part by National Science Foundation grant DMS-9703777.

the predictors to the axes of a 3D and marking the points to correspond to the values of $\tilde{y}$ (see, for example, Cook 1996; 1998, Ch. 5).

It is generally not possible to construct a single comprehensive display of all the data for a regression with four or more predictors. Low dimensional views of convenience such as those available in a scatterplot matrix can provide useful information about the distribution of $y|\mathbf{x}$ but can also be quite misleading. A common procedure in regressions with many predictors is to first fit a plausible model and then view the data in low dimensional plots designed to provide information on specific aspects of the fit. Residual plots, added variable plots, CERES plots and many other types fall into this category (see, Cook and Weiserg 1999 for a review). However, even these well-established graphics can be misleading. Examples of this were presented by Cook (1994; 1998, Section 1.2).

Informative low-dimensional exploratory displays of the data are possible without the need to pre-specify a parametric model. The basic idea is to reduce the dimension of the predictor vector with no or little loss of information on the regression. Drawing on a body of recent literature for dimension reduction and visualization in regression, our focus in this article is on methodology and application. These recent advances in visualization methodology have the potential to take on a fundamental role in the exploratory and diagnostic stages of a regression analysis. Most of the discussion is at the level of Cook and Weisberg (1999).

All of the methodology described in this article is available in *Arc*, a regression computer program that integrates many standard regression method with recent developments in regression graphics (Cook and Weisberg 1999; hereafter CW). *Arc* may be obtained from the web address www.stat.umn.edu/arc. Many but not all of the visualization methods available in Arc are described in CW. One purpose of this article is to describe visualization methodology is available in *Arc* but is not described in CW.

OUTLINE HERE WITH FURTHER REFERENCES TO CW. mention that some extension of the language in CW is necessary.

2 Dimension Reduction

Consider a regression with univariate response variable y and p predictors $x_1, \ldots, x_p$ which we collect in the $p \times 1$ predictor vector $\mathbf{x} = (x_j)$. The overarching goal of a regression analysis is to understand how the conditional distribution of the response y given $\mathbf{x}$ depends on the value assumed by $\mathbf{x}$ (CW, Ch. 2). While attention is often restricted to the mean function $\mathrm{E}(y|\mathbf{x})$ and perhaps the variance function

Var$(y|\mathbf{x})$, in full generality the object of interest is the conditional distribution of $y|\mathbf{x}$.

The study of $y|\mathbf{x}$ in practice may be problematic without restrictive assumptions, except perhaps in situations where data are plentiful and the dimension of $\mathbf{x}$ is small. For example, the most common assumption is that y depends on $\mathbf{x}$ through a linear model (CW, Ch. 6–17), which severely restricts the way that y can depend on $\mathbf{x}$. In this article we take a different approach based on dimension reduction.

Dimension reduction without loss of information is a dominant theme of regression graphics: We try to reduce the dimension of $\mathbf{x}$ without losing information on $y|\mathbf{x}$, and without requiring a pre-specified parametric model for $y|\mathbf{x}$. Borrowing terminology from classical statistics, we call this *sufficient dimension reduction*. Sufficient dimension reduction leads to the pursuit of *sufficient summary plots* (CW, Ch. 18) which contain all of the information on the regression that is available from the sample.

There are a variety of approaches to the graphical exploration of regression data and the pursuit of "interesting" low-dimensional projections. The approach based on sufficient dimension reduction differs from others because it rests on views that contain all the regression information.

In regression graphics, we try to reduce the dimension of $\mathbf{x}$ by finding *smallest* number of linear combinations of $\mathbf{x}$, say $\boldsymbol{\beta}_1^T\mathbf{x}, \ldots, \boldsymbol{\beta}_d^T\mathbf{x}$, so that for all y and all values of $\mathbf{x}$

$$\begin{aligned}\Pr(y \leq y_0|\mathbf{x}) &= \Pr(y \leq y_0|\boldsymbol{\beta}_1^T\mathbf{x}, \ldots, \boldsymbol{\beta}_d^T\mathbf{x}) \\ &= \Pr(y \leq y_0|\mathbf{B}^T\mathbf{x}) \qquad (1)\end{aligned}$$

where for notational convenience $\mathbf{B}$ is the $p \times d$ matrix with columns $\boldsymbol{\beta}_j$, $j = 1, \ldots, d$. This equation places no restrictions on the regression since it is always true for some $d \leq p$. For example, taking $d = p$ and choosing $\boldsymbol{\beta}_j$ to be the $p \times 1$ vector with a 1 in the jth position and 0's elsewhere we obtain

$$\Pr(y \leq y_0|\mathbf{x}) = \Pr(y \leq y_0|x_1, \ldots, x_p)$$

which is always true. Thus (1) represents a very general way of summarizing the data.

If (1) is true then we can replace the $p \times 1$ predictor vector $\mathbf{x}$ with the $d \times 1$ vector of linearly combinations $\mathbf{B}^T\mathbf{x}$ without loss of information on the regression.

We call the individual linear combinations $\boldsymbol{\beta}_j^T\mathbf{x}$ *sufficient predictors*[1] because they contain all the regression information that $\mathbf{x}$ has about y. The minimal number d of sufficient predictors is called the *structural dimension* of the regression (CW, Section 18.3); we refer to regressions as having 1D,...,dD structure.

If d is at most 2 or 3, which often seems to be the case in practice, and $\mathbf{B}$ is known, then as described briefly in Section 1, we can construct a graphical display of y versus the sufficient predictors $\mathbf{B}^T\mathbf{x}$ that contains all of the regression information. Such *sufficient summary plots* can be very useful exploratory tools and used to formulate models, identify anomalies and generally guide the subsequent analysis (CW, Ch. 18). If an estimate $\hat{\mathbf{B}}$ of $\mathbf{B}$ is known then we can proceed similarly using the estimated linear combinations $\hat{\mathbf{B}}^T\mathbf{x}$. To help fix ideas, we next describe several models and how they relate to the dimension reduction representation (1). Methods for estimating $\mathbf{B}$ will be discussed later in this article.

2.1 Structural Dimension

If a regression has 0D structure ($d = 0$), then y is independent of $\mathbf{x}$, and the predictors contain no information about the response. In such cases, a simple histogram or other graphical representation of the marginal distribution of y is a sufficient summary plot. If $d = 1$ then all the information about y that is available from $\mathbf{x}$ is contained in a single linear combination $\boldsymbol{\beta}_1^T\mathbf{x}$, and a plot of y versus $\boldsymbol{\beta}_1^T\mathbf{x}$ is a sufficient summary plot for the regression. Under 2D structure all the information about y that is available from $\mathbf{x}$ is contained in two linear combinations $(\boldsymbol{\beta}_1^T\mathbf{x}, \boldsymbol{\beta}_2^T\mathbf{x})$, and a three-dimensional plot of y versus $(\boldsymbol{\beta}_1^T\mathbf{x}, \boldsymbol{\beta}_2^T\mathbf{x})$ is a sufficient summary plot for the regression.

Each of the following regression models is an example of a regression with

[1] The sufficient predictors are not unique since any linear combination $\sum_j a_j\boldsymbol{\beta}_j^T\mathbf{x}$ of the sufficient predictors is a sufficient predictor. More generally, if $\mathbf{B}^T\mathbf{x}$ is a vector of d sufficient predictors and $\mathbf{A}$ is any $d \times d$ full rank matrix then $\mathbf{A}\mathbf{B}^T\mathbf{x}$ is another set of d sufficient predictors. However, since the distribution of $y|\mathbf{A}\mathbf{B}^T\mathbf{x}$ is the same as the distribution of $y|\mathbf{B}^T\mathbf{x}$, a plot of y versus $\mathbf{A}\mathbf{B}^T\mathbf{x}$ contains the same statistical information as a plot of y versus $\mathbf{B}^T\mathbf{x}$ and consequently the nonuniqueness of sufficient predictors is not normally a worrisome issue in practice. Starting with any vector of sufficient predictors $\mathbf{B}^T\mathbf{x}$, it is often desirable to choose $\mathbf{A}$ to be a square root of $\mathbf{B}^T\text{Var}(\mathbf{x})\mathbf{B}$ since then the new vector $\mathbf{A}\mathbf{B}^T\mathbf{x}$ of sufficient predictors will be uncorrelated, often leading to improved resolution in plots.

The vector $\mathbf{B}^T\mathbf{x}$ contains a *minimal set of sufficient predictors*. If we add any linear combination $\boldsymbol{\alpha}^T\mathbf{x}$ to a set of sufficient predictors, we obtain another set of sufficient predictors, although it need not be minimal. While we always try to find a minimal set of sufficient predictors, substantial dimension reduction is often possible without minimality.

1D structure,

$$
\begin{aligned}
y &= \beta_0 + \boldsymbol{\beta}_1^T\mathbf{x} + \sigma\varepsilon & (2)\\
y^{(\lambda)} &= \beta_0 + \boldsymbol{\beta}_1^T\mathbf{x} + \sigma\varepsilon & (3)\\
y &= \mathsf{m}(\boldsymbol{\beta}_1^T\mathbf{x}) + \sigma\varepsilon & (4)\\
y &= \mathsf{m}(\boldsymbol{\beta}_1^T\mathbf{x}) + \mathsf{v}^{1/2}(\boldsymbol{\beta}_1^T\mathbf{x})\varepsilon & (5)\\
\log\left(\frac{\Pr(y=0)}{1-\Pr(y=0)}\right) &= \beta_0 + \boldsymbol{\beta}_1^T\mathbf{x} & (6)
\end{aligned}
$$

where ε is an error that is independent of $\mathbf{x}$. Model (2) is the standard linear regression model. In (3) we have a linear regression model after some response transformation represented by the unknown parameter λ. The mean function in (4) is $\mathrm{E}(y|\mathbf{x}) = \mathsf{m}(\boldsymbol{\beta}_1^T\mathbf{x})$, where the *kernel mean function* m may be unknown or known, but it depends on only one linear combination of the predictors and thus the regression still has 1D structure. Similarly, in (5) both the kernel mean function m and the kernel variance function $\mathsf{v}^{1/2}$ may depend on $\boldsymbol{\beta}_1^T\mathbf{x}$. Many of the usual generalized linear models represent regressions with 1D structure. This is illustrated in (6) which describes logistic regression ($y = 0$ or 1) with mean function that depends on a single linear combination of the predictors. The class of regression models with 1D structure is quite large and covers many models used in practice.

A model with 2D structure can be considerably more complex than a regression with 1D structure. The models

$$
\begin{aligned}
y &= \mathsf{m}(\boldsymbol{\beta}_1^T\mathbf{x}, \boldsymbol{\beta}_2^T\mathbf{x}) + \mathsf{v}^{1/2}(\boldsymbol{\beta}_1^T\mathbf{x}, \boldsymbol{\beta}_2^T\mathbf{x})\varepsilon\\
\log\left(\frac{\Pr(y=0)}{1-\Pr(y=0)}\right) &= \mathsf{m}(\boldsymbol{\beta}_1^T\mathbf{x}, \boldsymbol{\beta}_2^T\mathbf{x})
\end{aligned}
$$

are examples of regressions with 2D structure provided $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are not collinear. For further background on structural dimension and summary plots, see CW (Ch. 18) and Cook (1998, Ch. 4-6).

The structural dimension d of a regression is an index of its complexity. Regressions with $d = 0$ are trivial. Regressions with $d = 1$ can be much more complicated but are usually less complicated than regressions with 2D structure, and so on. The structural dimension and sufficient predictors $\mathbf{B}^T\mathbf{x}$ of a regression will generally be unknown. In the next section we discuss conditions on the marginal distribution of $\mathbf{x}$ that are necessary to estimate these quantities. Then in Section 4 we discuss methods for estimating them, leading eventually to *estimated sufficient summary plots*.

3 Constraints on $\mathbf{x}$ Necessary for Estimation

There are several methods for estimating sufficient summary plots or properties thereof. While there are no restrictive constraints placed on the conditional distribution of $y|\mathbf{x}$, some constraints on the marginal distribution of $\mathbf{x}$ are necessary for estimating sufficient predictors.

3.1 Linearly Related Predictors

The key condition that we will assume throughout the rest of this article is that

$$\mathrm{E}(X_j|\mathbf{B}^T\mathbf{x}) = a_j + \mathbf{b}_j^T(\mathbf{B}^T\mathbf{x}), \ \ j = 1, \ldots, p \tag{7}$$

The mean function for the regression of the jth original predictor on the sufficient predictors should be linear or approximately so (See CW, Ch. 19). This requirement constrains the marginal distribution of $\mathbf{x}$ and not the conditional distribution of $y|\mathbf{x}$ as is usual when modeling in practice. Li (1991) emphasized that this condition is not a severe restriction, since most low-dimensional projections of a high-dimensional data cloud are close to being normal (Diaconis and Freedman 1984; Hall and Li 1993). Condition (7) is required to hold only for the sufficient predictors $\mathbf{B}^T\mathbf{x}$. Since $\mathbf{B}$ is unknown, in practice we may require that it hold for all possible $\mathbf{B}$, which is equivalent to elliptical symmetry of the distribution of $\mathbf{x}$ (Eaton 1986) including the normal. We refer to predictors satisfying this expanded condition as *linearly related predictors* (CW, Ch. 19).

The requirement of linearly related predictors is not very crucial in practice, but substantial departures can cause problems. We have found that simultaneous coordinatewise power transformations of positive predictors can help induce linearly related predictors to a useful approximation (CW, Section 19.4). Letting $x^{(\lambda)} = (X^\lambda - 1)/\lambda$ for $\lambda \neq 0$ and $x^{(\lambda)} = \log x$ for $\lambda = 0$, the idea here is to find power transformations λ_j so that the vector

$$T(\mathbf{x}) = (X_j^{(\lambda_j)}), \ \ j = 1, \ldots, p$$

of transformed predictors is apporximately multivariate normal (Velilla 1993). A series of strictly monotonic coordinatewise transformations $T(\mathbf{x})$ of the predictor vector $\mathbf{x}$ does not cause any difficulty for the methods discussed in this article because the distribution of $y|\mathbf{x}$ is the same as that of $y|T(\mathbf{x})$. Since no model is being assumed for $y|\mathbf{x}$, the consequence of a predictor transformation is just to change the manner in which the conditional distributions are indexed.

3.2 Constant Covariances

In addition to linearly related predictors, some methods also require a *constant covariance condition* on the predictors:

$$\mathrm{Cov}(X_j X_k|\mathbf{B}^T\mathbf{x}) = \sigma_{jk}, \;\; j,k = 1,\ldots p \qquad (8)$$

This condition seems less important that the linearity condition, but again problems can arise if the departures are substantial. Using power transformations to multivariate normality helps insure the constant covariance condition in addition to linearity related predictors. Other methods for insuring both conditions will be disussed later in this article.

4 Estimation Methods

There are several methods for inferring about sufficient summary plots. All are based on estimating p linear combinations of the original predictor $\mathbf{x}$:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \rightarrow \mathbf{W} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1^T\mathbf{x} \\ \hat{\boldsymbol{\beta}}_2^T\mathbf{x} \\ \vdots \\ \hat{\boldsymbol{\beta}}_p^T\mathbf{x} \end{pmatrix}$$

The p estimated linear combinations $\hat{\boldsymbol{\beta}}_j^T\mathbf{x}$ are ordered on the “likelihood” that they will be needed in a sufficient summary plot. The first d linear combinations $\boldsymbol{\beta}_1^T\mathbf{x},\ldots,\boldsymbol{\beta}_d^T\mathbf{x}$ estimate the sufficient predictors. If $d = 0$ then none of the linear combinations is needed. If $d = 1$ then only the first is needed and a plot of y versus $\hat{\boldsymbol{\beta}}_1^T\mathbf{x}$ is the estimated summary plot. Similarly, if the structural dimension $d = 2$ then only the first two estimated linear combinations are needed and a three-dimensional plot of y versus $(\hat{\boldsymbol{\beta}}_1^T\mathbf{x}, \hat{\boldsymbol{\beta}}_2^T\mathbf{x})$ is the estimated summary plot.

There are both graphical and numerical methods for finding the vector $\mathbf{W}$ of linearly transformed predictors. With $p = 2$ predictors it is possible to use a single 3D plot to estimate d and $\mathbf{W}$ graphically. This methodology is discussed in CW (Ch. 18). When $p > 2$, a series of 3D plots can be used to perform the same tasks using a procedure called *graphical regression* (CW, Ch. 20). A scatterplot matrix in *Arc* can be used to check visually for linearly related predictors, find normalizing power transformations of the predictors if necessary, and decide if the data are consistent with the possibility that $d \leq 1$ (CW Ch. 19).

There are four primary numerical methods for producing the vector $\mathbf{W}$. While each is designed to estimate sufficient predictors, each has its own advantages and disadvantages depending on the complexity of the regression. The four methods are:

- OLS and other methods based on convex objective functions,
- Sliced Inverse Regression (SIR; Li 1991),
- Sliced Average Variance Estimation (SAVE, Cook and Weisberg 1991), and
- Principal Hessian Directions (pHd; Li 1992)

4.1 Ordinary least squares (OLS)

If the predictors are linearly related, $d = 1$ and the covariance between y and $\mathbf{x}$ is not zero then the coefficient $\boldsymbol{\beta}_1$ of the single sufficient predictor $\boldsymbol{\beta}_1^T\mathbf{x}$ can be estimated as the coefficient vector $\hat{\mathbf{b}}$ of $\mathbf{x}$ from the ordinary least squares fit of y on $\mathbf{x}$, including an intercept. The estimated sufficient summary plot is then just the plot of y versus the fitted values $\hat{y} = \hat{a} + \hat{\mathbf{b}}^T\mathbf{x}$ from this same fit. This conclusion, which is called the *1D estimation result* by CW (Section 19.3), seems quite important and may be one reason for the success of OLS estimation over the years. It is worth emphasizing that OLS is being used here only as a method for summarizing the data and that no model for $y|\mathbf{x}$ is assumed. If the structural dimension is greater than 1 then the OLS fitted values still estimate a sufficient predictor, although there is no way to tell which one.

Fitting methods other than least squares can be used as well, provided that they are based on objective functions that are convex in $\mathbf{b}^T\mathbf{x}$. For example, if the response is binary taking the values $y = 0$ and $y = 1$ then it may desirable to fit using logistic regression (6) as a method of summarizing the data. Additional discussion of this result, was given by Li and Duan (1989) and Cook (1998a, Proposition 8.1) .

4.2 Sliced Inverse Regression (SIR)

Sliced inverse regression as proposed by Li (1991) is a method for estimating sufficient predictors when $d > 1$. It requires linearly related predictors and begins by replacing the response y with a discrete version $\tilde{y}$ constructed by partitioning the range of y into H non-overlapping slices. SIR is then based on the regression of $\tilde{y}$ on $\mathbf{x}$. The number of slices H is a tuning parameter much like the tuning parameters encountered in the smoothing literature. If $\boldsymbol{\beta}^T\mathbf{x}$ is a sufficient predictor for the regression of $\tilde{y}$ on $\mathbf{x}$ then it is also a sufficient predictor for the regression of y on $\mathbf{x}$. However, the reverse need not be true. If $H < d$ then the set of sufficient predictors for $\tilde{y}$ on $\mathbf{x}$ will necessarily exclude some of the sufficient predictors for the regression of y on $\mathbf{x}$. Our experience indicates that good results are often obtained by choosing H to be some what larger than $d + 1$, trying a few different values of H as necessary. Choosing H very much larger than d should generally be avoided[2] Slicing may not be necessary when y is qualitative or takes on few values since then we can set $\tilde{y} = y$. For further discussion of foundations, the number of slices and implementation algorithms, see Cook (1998a, Ch. 11), Cook and Weisberg (1994) and Li (1991).

Having selected the number of slices, the next step is to form the $p \times p$ matrix

$$\hat{\mathbf{M}}_{SIR} = \sum_{s=1}^{H} f_s \bar{\mathbf{z}}_s \bar{\mathbf{z}}_s^T \tag{9}$$

where f_s is the fraction of observations falling in slice s, and $\bar{\mathbf{z}}_s$ is the average of the sample standardized predictor vector

$$\mathbf{z}_i = \widehat{\mathrm{Var}}(\mathbf{x})^{-\frac{1}{2}}(\mathbf{x}_i - \bar{\mathbf{x}}), \;\; i = 1, \ldots, n \tag{10}$$

for the y's in slice s, $s = 1, \ldots, H$. Here $\widehat{\mathrm{Var}}(\mathbf{x})$ is the usual estimate of the covariance matrix of $\mathbf{x}$, and $\bar{\mathbf{x}}$ is the sample mean of the predicor vector. Let $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_p$ be the eigenvectors of $\hat{\mathbf{M}}_{SIR}$ corresponding to its eigenvalues $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$. Then the estimated coefficient vectors $\hat{\boldsymbol{\beta}}_j$ are given by

$$\hat{\boldsymbol{\beta}}_j = \widehat{\mathrm{Var}}(\mathbf{x})^{-\frac{1}{2}}\hat{\mathbf{u}}_j, \;\; j = 1, \ldots, p \tag{11}$$

We call the corresponding estimated predictors $\hat{\boldsymbol{\beta}}_j^T\mathbf{x}$, $j = 1, \ldots, p$, the *SIR predictors* to distinguish them from other types discussed later in the article.

A plot of y versus the first two SIR predictors and a marked plot of $\tilde{y}$ versus the first three SIR predictors are usually informative in practice. However, to be most effective, we require a method for inferring about d since it will typically be unknown.

[2] In *Arc*, the default number of slices is 8.

4.2.1 Normal Theory Tests on d

Assuming that the SIR analysis does not miss any sufficient predictors in theory, we can estimate d by performing a series of hypothesis tests using the statistic

$$\hat{\Lambda}_m = n \sum_{j=m+1}^{p} \hat{\lambda}_j \tag{12}$$

Under the assumption of normally distributed predictors, Li (1991) proved $\hat{\Lambda}_d$ has an asymptotic chi-square distribution with $(p-d)(H-d-1)$ degrees of freedom. Consequently, we can estimate d as follows: Beginning with $m = 0$, compare $\hat{\Lambda}_m$ to the percentage points of its distribution under the hypothesis $d = m$. If it is smaller, there is no information to contradict the hypothesis. If it is larger, conclude that $d > m$, increment m by 1 and repeat the procedure. The estimate $\hat{d} = m$ follows when $\hat{\Lambda}_{m-1}$ is relatively large, implying that $d > m - 1$, while $\hat{\Lambda}_m$ is relatively small, so there is no information to contradict the hypothesis. For further discussion of this testing procedure and the distribution of $\hat{\Lambda}_m$ without the assumption of normal predictors, see Cook (1998a, Ch. 11). Bura and Cook (2000) studied extensions of Li's chi-square test that do not require normally distributed predictors.

4.2.2 Permutation Tests on d

In addition to the normal theory tests discussed in Section4.2.1, *Arc* also provides a nonparametric permutation test for d that does not require normally distributed predictors. To test the hypothesis $d = 0$ versus $d > 0$, the n values of the response are permuted randomly C times. For each permutation, the value $\hat{\Lambda}_0^{(k)}$ of test statistic (12) with $m = 0$ is computed, $k = 1, \ldots, C$. The p-value is then the fraction of the $\hat{\Lambda}_0^{(k)}$'s that exceed the value $\hat{\Lambda}_0$ of the test statistic for the actual data. The procedure for testing $d = 1$ versus $d > 1$ is the same except the indices i of the pairs $(y_i, \hat{\boldsymbol{\beta}}_1^T \mathbf{x}_i)$, $i = 1, \ldots, n$ are randomly permuted C times. Similarly, to test $d = d_0$ versus $d > d_0$, randomly permute the indices of the $(d_0 + 1) \times 1$ vectors $(y_i, \hat{\boldsymbol{\beta}}_1^T \mathbf{x}_i, \ldots, \hat{\boldsymbol{\beta}}_{d_0}^T \mathbf{x}_i)$. The procedure for estimating d is then the same as that described for the normal theory test.

In general, this permutation test procedure does not estimate d but instead estimates an upper bound on d. Consequently, the procedure may end with more predictors $\hat{\boldsymbol{\beta}}_j^T \mathbf{x}$ that are really necessary. However, if the predictors are normally distributed, then the permutation procedure does estimate d and the permutation estimate is very often the same as the normal theory estimate discussed in Section 4.2.1.

The permutation procedure for estimating d is quite general and can be used with the other numerical esitimation procedures discussed in this article, in addition to many others. Permutation tests were proposed by Cook and Weisberg (1991) and developed by Cook and Yin (2000).

4.2.3 Bivariate Responses

The theory behind SIR does not require that y be univariate, and it holds equally for multivariate responses. An implementation of SIR for bivariate responses is provided as a supplement to *Arc*. This may be useful in a number of contexts, including the study of net effects of single predictors x_j in linear models (Cook 1995).

4.2.4 Limitations

While SIR can be an effective procedure it can also miss important sufficient predictors. In particular, it is not effective at finding sufficient predictors that occur in symmetric relationships. For example, suppose $\mathbf{x}$ follows a standard normal distribution and that

$$y = x_1^2 + \sigma\varepsilon$$

Then $\bar{\mathbf{z}}_s$ estimates 0 in all slices and SIR will fail. The same type of situation happens in more complicated models like

$$y = \beta_0 + \boldsymbol{\beta}_1^T\mathbf{x} + (\boldsymbol{\beta}_2^T\mathbf{x})^2 + \sigma\varepsilon$$

Now SIR should detect the first sufficient predictor $\boldsymbol{\beta}_1^T\mathbf{x}$ but miss the second sufficient predictor $\boldsymbol{\beta}_2^T\mathbf{x}$. The SAVE method discussed in the next section has the ability to find more of the sufficient predictors and is more is comprehensive than SIR.

4.3 Sliced average variance estimation (SAVE)

SAVE was originally proposed by Cook and Weisberg (1991) and developed further by Cook and Lee (1998) and Cook and Critchley (2000). A discussion of basic methodology was given by Cook (2000). It requires both linearly related predictors and the constant covariance condition.

Like SIR, SAVE begins by constructing a sliced version $\tilde{y}$ of y. Next, construct the $p \times p$ matrix

$$\hat{\mathbf{M}}_{SAVE} = \sum_{s=1}^{H} f_s (I - \hat{\mathbf{\Sigma}}_s)^2, \tag{13}$$

where $\hat{\mathbf{\Sigma}}_s$ the estimated covariance matrix for the vector of standardized predicors (10) within slice s. The remaining calculations under SAVE now parallel those for SIR. Let $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_p$ be the eigenvectors of $\hat{\mathbf{M}}_{SAVE}$ corresponding to its eigenvalues $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$. Then the estimated coefficient vectors $\hat{\boldsymbol{\beta}}_j$ are again given by

$$\hat{\boldsymbol{\beta}}_j = \widehat{\text{Var}}(\mathbf{x})^{-\frac{1}{2}} \hat{\mathbf{u}}_j, \;\; j = 1, \ldots, p \tag{14}$$

The corresponding estimated predictors $\hat{\boldsymbol{\beta}}_j^T \mathbf{x}$, $j = 1, \ldots, p$, are called the *SAVE predictors*.

Like SIR, a plot of y versus the first two SAVE predictors and a marked plot of $\tilde{y}$ versus the first three SAVE predictors are usually informative in practice. Asymptotic normal theory test procedures for estimating d under SAVE are available only in special cases (Cook and Lee 1999). Generally, we recommend estimating d by using permutation tests. These estimates are constructed as outlined in Section 4.2.2 in the context of SIR. In particular, the same test statistic (12) is used except the eigenvalues are computed from $\hat{\mathbf{M}}_{SAVE}$.

4.4 Principal Hessian directions (pHd)

pHd was proposed by Li (1992), and extended by Cook (1998b). Like SAVE it requires both the linearity condition and the constant covariance condition.

To find the *pHd predictors*, let $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_p$ be the eigenvectors corresponding to the squared eigenvalues $\hat{\lambda}_1^2 \geq \ldots \geq \hat{\lambda}_p^2$ of the $p \times p$ matrix

$$\hat{\mathbf{M}}_{pHd} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}) \mathbf{z}_i \mathbf{z}_i^T \tag{15}$$

where $\mathbf{z}$ is the standardized predictor (10). Then the estimated coefficient vectors $\hat{\boldsymbol{\beta}}_j$ are again given by

$$\hat{\boldsymbol{\beta}}_j = \widehat{\text{Var}}(\mathbf{x})^{-\frac{1}{2}} \hat{\mathbf{u}}_j, \;\; j = 1, \ldots, p \tag{16}$$

The corresponding pHd predictors $\hat{\boldsymbol{\beta}}_j^T \mathbf{x}$ are used in the general same way as the SIR or SAVE predictors.

pHd is better at estimating sufficient predictors that correspond to nonlinear trends in the data than it is at estimating sufficient predictors that correspond to linear trends (Cook 1998b). For this reason pHd seems to be most effective when used as a model diagnostic, replacing the response y with the residuals r from a fitted linear model. When used in this way d is the minimum number of sufficient predictors for the regression of the residuals r on $\mathbf{x}$.

The following test statistic is used to estimate d under the pHd method in the same way as the SIR and SAVE test statistics are used:

$$\hat{\Lambda}_m(y) = \frac{n}{2\widehat{\mathrm{Var}}(y)} \sum_{j=m+1}^{p} \hat{\lambda}_j^2 \tag{17}$$

where $\widehat{\mathrm{Var}}(y)$ is the usual estimate of the marginal variance of y. This test statistic is written as a function of y so that we can write $\hat{\Lambda}_m(r)$ when using residuals. Assuming normally distributed predictors and that pHd does not miss any sufficient predictors in theory, $\hat{\Lambda}_m(r)$ is distributed asymptotically as a chi-squared random variable with $(p-d)(p-d+1)/2$ degrees of freedom. This then is the reference distribution for estimating the structural dimension of the regession of r on $\mathbf{x}$ using the general test procedure described in Section 4.2. If the model producing r is true then the structural dimension of the regression of r on $\mathbf{x}$ is zero.

The asymptotic distribution of $\hat{\Lambda}_m(y)$ is a linear combination of chi-squared random variables with unknown coefficients that must be estimated from the data for use in practice. Since this distribution is relatively complicated and there may be doubt about the accuracy of the approximation it provides, we recommend the use of the permutation test procedure for estimating the structural dimension of the regression of y on $\mathbf{x}$ under pHd.

4.5 Discussion

4.6 Checking the Conditions

It is prudent at the outset of an analysis to check for linearly related predictors and constant covariances. This can be done by viewing a scatterplot matrix of the predictors (CW, Section 19.2). If there are no strong nonlinear relationships or clearly nonconstant variances in the individual plots of the matrix, the conditions are probably satisfied to a reasonable approximation. A scatterplot matrix provides a necessary but not sufficient check. Nevertheless, it seems to work well in practice.

Power transformations of positive predictors to approximate joint normality is often an effective remedy when there is clear curvature or heteroscedasticity in the scatterplot matrix (CW, Section 19.4). In addition, it may be possible to move closer to multivariate normality by introducing predictor weights. The idea here, as developed by Cook and Nachtsheim (1994), is to modify the jump heights of the empirical distribution of the standardized predictor $\mathbf{z}$ so that the weighted empirical distirbution matches a target multivariate normal with covariance matrix $\sigma^2 I$, where σ is a use-selected standard deviation. The mean of this target distribution should be at the center of the data, typically the sample mean or L_1 median. Values of σ between 0.5 and 1 seem to work the best in practice. The weights can be determined using a Monte Carlo routine that is available as a supplement to *Arc*. The weighed empirical predictor distribution can then used in OLS, SAVE, SIR and pHd.

4.7 General Operating Characteristics

While it has been demonstrated that SAVE, SIR and pHd can yield very useful results in practice, they are all fallible in the right situations. SAVE is most comprehensive. In theory, it will find all the sufficient predictors found by SIR and pHd plus any others present in the regression. This increased flexibility comes with a price, however. In effect, SIR and pHd looks through certain classes of linear combinations in their search for sufficient predictors. SAVE looks through a larger class of linear combinations that includes those considered by SIR and pHd. Relatively straightforward structure that can be detected with SIR or pHd may be harder to detect with SAVE, particularly when p is large. In such situations the summary plots produced by SAVE may not be as crisp or informative as those based on SIR or pHd. We nearly always use both SIR and SAVE in practice and usually find that comparing their solutions is informative.

SIR does well at finding directions $\boldsymbol{\beta}_j$ in which there are linear trends in the mean function $\mathrm{E}(y|\mathbf{x})$. The sample correlation between the OLS fitted values and the first SIR predictor is typically quite large. SIR can also find directions in which the variance function $\mathrm{Var}(y|\mathbf{x})$ is not constant. This property may be useful when using SIR as a model diagnostic, replacing the response y with residuals r from a linear model. However, SIR is generally ineffective at finding curvature in the residual mean function $\mathrm{E}(r|\mathbf{x})$.

On the other hand, pHd is a relatively specialized method that seems to work particularly well for finding curvature in the residual mean function $\mathrm{E}(r|\mathbf{x})$. It does not seem to work as well as SIR or SAVE for finding directions in which the residual variance function Var$(r|\mathbf{x})$ is not constant. It can be applied with y as well but its operating characteristics are not a well understood.

4.8 Nature of the Response

There are no restrictions placed on the response y by any of the methods. It can be continuous, discrete or qualitative. Indeed, SIR and SAVE seem to work particularly well in discriminant analysis where the response is often a classification like sex, location or species (Cook and Yin 2000). Also, the SIR and SAVE predictors will stay the same when the response is replaced with any strictly monotonic transformation. In some regressions the visual resolution of summary plots might be improved by transforming the response at the outset.

4.9 Estimating the Structural Dimension

The permutation test procedure for estimating the structural dimension of a regression is quite general and can be applied with any of the methods discussed in this article. The large sample distribution of $\hat{\Lambda}_d$ is know for SIR applied to either y or r and for pHd applied to r. For other situations we recommend the permutation test.

It is also possible to estimate the structural dimension of a regression visually. This involves using the SAVE, SIR or pHd predictors in graphical regession (CW, Ch. 20). In effect, these methods are use as pre-processors to construct linear combinations of $\mathbf{x}$ that are ordered based the likelihood that they are estimating sufficient predictors. The method-specific predictors are then studied graphically using a series of 3D plots, ending with a summary plot and an estimate of the structural dimension.

4.10 How the Methods Work

The information that SIR and SAVE use to form their predictors comes from the column of p *inverse response plots*, x_j versus y, in a scatterplot matrix of the predictors and the response. CW (Section 19.5) describe how to use the inverse response plots to assess visually if the data are consistent with the possibility that $d \leq 1$ or if $d > 1$ is indicated. SAVE and SIR use the same data to extract additional information on the structural dimension and to estimate sufficient predictors. They measure the shape of the inverse response plots by first replacing the response y with a discrete version $\tilde{y}$ constructed by partitioning the range of y into

H slices. SIR then gauges the shapes of the inverse mean functions $\mathrm{E}(\mathbf{x}|\tilde{y})$, while SAVE gauges the shapes of the inverse functions $\mathrm{E}(\mathbf{x}|\tilde{y})$ and the inverse variance functions $\mathrm{Var}(\mathbf{x}|\tilde{y})$ simultaneously.

pHd gains its information from the covariances between the response (or residuals) and the products of the standardized predictors $z_j z_k$, $j, k = 1, \ldots, p$. Thus it is a marginal method that does not require slicing. pHd does not have a tuning parameter, a property that might be seen as an advantage. Assuming that the population standardized predictor $\mathbf{z} = \mathrm{Var}(\mathbf{x})^{1/2}(\mathbf{x} - \mathrm{E}(\mathbf{x}))$ is normally distributed, the pHd matrix $\hat{\mathbf{M}}_{pHd}$ is an estimate of the average Hessian matrix

$$E\left(\frac{\partial^2 \mathrm{E}(y|\mathbf{z})}{\partial \mathbf{z} \partial \mathbf{z}^T}\right)$$

for the mean function (Li 1992). Thus the eigenvectors of $\hat{\mathbf{M}}_{pHd}$ estimate the principal directions of the Hessian matrix. The name "principal Hessian directions" may not be applicable when $\mathbf{z}$ is not normal because the connection with the Hessian matrix may be lost.

5 Illustration

In this section we illustrate how to use *Arc* for the methodology discussed in this article. Familiarity with *Arc* at the level of CW is assumed.

5.1 Supplemental *Arc* Files

Three supplemental files, all available at the *Arc* Internet site, are needed to implement some of the methods discussed in this article. They should be loaded before loading the data set to be analyzed.

PermTest.lsp This file contains the code to perform the permuations tests on the structural dimesion of the regression. It causes a new item – *Permutation tests* – to appear in inverse regression model menus. Selecting this item will produce a dialog box for the number of permutations. After completing this dialog, the output, which is illustrated later in this section, will appear in the text window. The computation of the permutation tests may take several minutes if the data set or number of permutations is large.

BivSir.lsp This file contains the code for bivariate SIR. It causes the item *Fit Bivariate SIR* to appear in the graph&fit menu. Use of this menu item will be dicussed later a bit later.

Reweight.lsp The code for determining predictor weights for multivariate normality, as discussed in Section 4.6, is contained in this file. It produces the menu item *Reweight for Normality* in the data set menu. The dialog produced by this menu item allows the user to select the predictors for reweighting, specify the value of σ and the number of Monte Carlo trials, and choose either the sample mean or L_1 median as the center of the target normal distribution. After the calculations are finished, the weights are added to the data set and are then available for use in any inverse regression method. The weights added to the data set have names of the form "pwt-sigma" where "pwt" stands for predictor weight and "sigma" is the value of σ supplied in the dialog.

We use the Australian athletes data (CW, Section 19.6) to illustrate how *Arc* can be used to produce summary plots. The response variable is an athletes lean body mass, *LBM.* There are three predictors, red cell count *RCC*, height *Ht* and *Wt.*

5.2 Checking the Conditions

We first inspected a scatterplot matrix of the predictors. We concluded that case 165 is outlying and that it may distort the results. We deleted that from the data for all calculations with the intention of restoring it in the final summary plot. There didn't seem to be any strong nonlinearities among the predictors, but some nonconstant variance seemed visually evident. Consequently, we used *Arc* to find normalizing power transformations of the predictors (CW, Section 19.4). The results indicated that *Wt* should be transformed to logs, while no transformation of the other two predictors was indicated. Consequently, we used the three predictors $\mathbf{x} = (RCC, Ht, \log(Wt))^T$ for the analysis.

5.3 Dimension Reduction

The setup and output for SAVE, SIR and pHd are quite similar. Consequently we use SIR for the primary illustration and indicate how the output for SAVE and pHd differs.

The item Inverse regression in the Graph&Fit menu produces a dialog box for selecting the predictors and the response, and for choosing the method by using the radio buttons on the right of the plot. Three methods are available by default: SIR, SAVE and pHd applied to the residuals from the OLS regression of y on $\mathbf{x}$.

Default Inverse Regression Output from *Arc*.

```
Inverse Regression SIR
Name of Dataset = AIS
Name of Fit = I18.SIR
Response = LBM
Predictors = (RCC Ht log[Wt])
Deleted cases are:
(165)

Number of slices = 6
Slices sizes are:  (35 34 34 35 34 29)
Std. coef. use predictors scaled to have SD equal to one.
Coefficients    Lin Comb 1      Lin Comb 2      Lin Comb 3
Predictors      Raw     Std.    Raw     Std.    Raw     Std.
RCC           -0.143  -0.301   0.441   0.729  -0.116  -0.160
Ht            -0.006  -0.272   0.006   0.217   0.025   0.783
log[Wt]       -0.990  -0.914  -0.897  -0.649  -0.993  -0.601

Eigenvalues             0.832          0.145           0.016
Rˆ2(OLS| SIR lin comb) 0.999          0.999           1.000

Approximate Chi-squared test statistics based on partial
sums of eigenvalues times 201

Number of    Test
Components   Statistic     df p-value
   1         199.58        15   0.000
   2         32.295         8   0.000
   3         3.1889         3   0.363
```

Permutation Tests.

```
Permutation pvalues for
sums of eigenvalues times 201
Number of permutations: 500

Number of    Test               Permutation
Components   Statistic            p-value
```

```
  1         199.58              0.000
  2         32.295              0.000
  3         3.1889              0.142
```

Availability. SIR and SAVE are available through a point-and-click interface in the regression program *Arc* (Cook and Weisberg 1999) which is available at the internet site http://www.stat.umn.edu/arc. The program includes much related methodology as well.

References

Chen, C-H, and Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, 8, 289–316.

Cheng, C-S. and Li, K. C. (1995). A study of the method of principal Hessian directions for analysis of data from designed experiments. *Statistica Sinica*, 5, 617–637.

Cook, R. D. (1994a). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89, 177–190.

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

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

Cook, R. D. (1998b). Principal Hessian directions revisited (with discussion). *Journal of the American Statistical Association*, 93, 84–100.

Cook, R. D. and Critchley, F. (2000). Identifying outliers and regression mixtures graphically. *Journal of the American Statistical Association*, to appear.

Cook, R. D. and Lee, H. (1999). Dimension reduction in regressions with a binary response. *Journal of the American Statistical Association*, in press.

Cook, R. D. and Nachtsheim, C. J. (1994). Re-weighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, 89, 592–599.

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

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

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

taxonomic

Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approach.* London: Chapman and Hall.

FerrT̂e, L. (1998). Determining the dimension of sliced inverse regression and related methods. *Journal of the American Statistical Association*, 93, 132–140.

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

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single index models. *Annals of Statistics*, 21, 157–178.

Hsing, T. and Carroll (1992). An asymptotic theory of sliced inverse regression. *Annals of Statistics*, 20, 1040–1061.

Kent, J. T. (1991). Discussion of Li (1991). *Journal of the American Statistical Association*, 86, 336–337.

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

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *Journal of the American Statistical Association*, 87, 1025–1039.

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

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley.

Seber, G.A.F. (1984). *Multivariate Observations*. New York: Wiley.