Resistant Dimension Reduction

Jing Chang and David J. Olive*

Culver Stockton College and Southern Illinois University

October 21, 2007

Abstract

Existing dimension reduction (DR) methods such as ordinary least squares (OLS) and sliced inverse regression (SIR) often perform poorly in the presence of outliers. Ellipsoidal trimming can be used to create outlier resistant DR methods that can also give useful results when the assumption of linearly related predictors is violated. Theory for SIR and OLS is reviewed, and it is shown that several important hypothesis tests for an important class of regression models can be done using OLS output originally meant for multiple linear regression.

KEY WORDS: sliced inverse regression; outliers; robust regression.

*Jing Chang is Assistant Professor of Mathematics, Culver Stockton College, One College Hill, Canton, MO 63435, USA, E-mail address: jchang@culver.edu, and David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Mailcode 4408, Carbondale, IL 62901-4408, USA, E-mail address: dolive@math.siu.edu. This work was supported by the National Science Foundation grant DMS-0600933. The authors are grateful to the associate editor and referees for a number of helpful suggestions for improvement in the article.

1 INTRODUCTION

Regression is the study of the conditional distribution $Y|\boldsymbol{x}$ of the response Y given the $(p-1)\times 1$ vector of nontrivial predictors $\boldsymbol{x}$. *Dimension reduction* (DR) replaces $\boldsymbol{x}$ with a lower dimensional $d \times 1$ vector $\boldsymbol{w}$ without loss of information on the regression $Y|\boldsymbol{x}$. Following Cook and Li (2004), if there is a $(p-1)\times k$ matrix $\boldsymbol{B} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_k]$ such that the k linear combinations $\boldsymbol{B}^T\boldsymbol{x}$ fully describe the conditional distribution $Y|\boldsymbol{x}$, then the subspace spanned by the columns of $\boldsymbol{B}$ is a dimension reduction subspace and Y is independent of $\boldsymbol{x}$ given $\boldsymbol{B}^T\boldsymbol{x}$, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{B}^T\boldsymbol{x} \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{x}, ..., \boldsymbol{\beta}_{\mathrm{k}}^{\mathrm{T}}\boldsymbol{x}. \tag{1.1}$$

If model (1.1) is valid, then $\boldsymbol{\beta}_1^T\boldsymbol{x}, ..., \boldsymbol{\beta}_k^T\boldsymbol{x}$ is called a set of *sufficient predictors.*

The *structural dimension* d is the smallest value of k such that model (1.1) holds and $\boldsymbol{w} = \boldsymbol{B}^T\boldsymbol{x}$. If $d = 0$ then $Y \perp\!\!\!\perp \boldsymbol{x}$, and $0 \le d \le p-1$ since $Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{I}_{p-1}\boldsymbol{x}$ where $\boldsymbol{I}_{p-1}$ is the $(p-1)\times(p-1)$ identity matrix. If $k = d$, let the minimum dimension reduction subspace $S(\boldsymbol{B})$ be the span of the columns of $\boldsymbol{B}$.

In a *1D regression model*, $d = 1$ and Y is conditionally independent of $\boldsymbol{x}$ given a single linear combination $\boldsymbol{\beta}^T\boldsymbol{x}$ of the predictors, written

$$Y \perp\!\!\!\perp \boldsymbol{x}|\boldsymbol{\beta}^T\boldsymbol{x} \quad \text{or} \quad \mathrm{Y} \perp\!\!\!\perp \boldsymbol{x}|\alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}. \tag{1.2}$$

Many of the most commonly used regression models are 1D regression models, and the additive error *single index model* has the form

$$Y = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e, \tag{1.3}$$

where e is a zero mean error that is independent of $\boldsymbol{x}$. Important theoretical results for the single index model were given by Brillinger (1977, 1983), and Li and Duan (1989) extended these results to models of the form $Y = g(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}, e)$ where g is a bivariate inverse link function.

A key condition for many theoretical results in dimension reduction is the condition of *linearly related predictors* which holds if $E(\boldsymbol{x}|\boldsymbol{B}^T\boldsymbol{x})$ is a linear function of $\boldsymbol{B}^T\boldsymbol{x}$. This condition holds if $\boldsymbol{x}$ is nondegenerate and elliptically contoured (EC) with second moments. Hall and Li (1993) show that the linearity condition often approximately holds in large dimensions even if $\boldsymbol{x}$ is not EC.

In the following sections ordinary least squares (OLS), sliced inverse regression (SIR), (residual based) principal Hessian directions (PHD), and sliced average variance estimation (SAVE) are used. These well known DR methods were chosen since the Weisberg (2002) `dr` library allows computation in `R`. Following Cook and Li (2002), OLS and SIR can be shown to be useful when $\boldsymbol{x}$ is EC. SAVE and PHD theory has an additional constant covariance condition which is satisfied when $\boldsymbol{x}$ is multivariate normal (MVN).

Further information about dimension reduction methods can be found, for example, in Chen and Li (1998), Cook (1998ab, 2004), Cook and Critchley (2000), Cook and Ni (2005), Cook and Weisberg (1991), Li (1991, 1992, 2000), Li and Duan (1989), Li and Zhu (2007) and Xia, Tong, Li and Zhu (2002). Outlier resistance is studied by Prendergast (2005), Heng-Hui (2001) and Čížek and Härdle (2006) who replace local least squares by local one step M or L smoothers. Gather, Hilker and Becker (2001, 2002) robustify SIR by replacing the sample covariance estimator by an S estimator that is impractical to compute.

Section 2 reviews DR theory for OLS and SIR. Section 3 presents a general method for obtaining outlier resistant DR methods, while Section 4 gives examples and simulations.

2 Some DR Theory

The following results relating to OLS and SIR dimension reduction will be useful. Let $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. The population coefficients from an OLS regression of Y on $\boldsymbol{x}$ are $\alpha_{OLS} = E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x})$ and $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$.

Let the data be $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. Let the $p \times 1$ vector $\boldsymbol{\eta} = (\alpha, \boldsymbol{\beta}^T)^T$, let $\boldsymbol{X}$ be the $n \times p$ OLS design matrix with ith row $(1, \boldsymbol{x}_i^T)$, and let $\boldsymbol{Y} = (Y_1, ..., Y_n)^T$. Then the OLS estimator $\hat{\boldsymbol{\eta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$. The sample covariance of $\boldsymbol{x}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \quad \text{where the sample mean} \quad \overline{\boldsymbol{x}} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \boldsymbol{x}_{\mathrm{i}}.$$

Similarly, define the sample covariance of $\boldsymbol{x}$ and Y to be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i Y_i - \overline{\boldsymbol{x}}\,\overline{Y}.$$

The first result shows that $\hat{\boldsymbol{\eta}}$ is a consistent estimator of $\boldsymbol{\eta}$.

i) Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid random vectors such that $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exist. Then

$$\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T \overline{\boldsymbol{x}} \xrightarrow{D} \alpha_{OLS}$$

and

$$\hat{\boldsymbol{\beta}}_{OLS} = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \xrightarrow{D} \boldsymbol{\beta}_{OLS} \quad \text{as} \quad \mathrm{n} \to \infty.$$

The next result shows that the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ can be useful for dD regression.

ii) Cook (1994, p. 184): If $\boldsymbol{x}$ follows a nondegenerate elliptically contoured distribution with second moments, then $\boldsymbol{\beta}_{OLS} \in S(\boldsymbol{B})$.

The following results will be for 1D regression and some notation is needed. Many 1D regression models have an error e with $\sigma^2 = \text{Var}(e) = E(e^2)$. Let the population OLS residual $v = Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T \boldsymbol{x}$ with

$$\tau^2 = E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T \boldsymbol{x})^2] = E(v^2), \tag{2.1}$$

and let the OLS residual be $r = Y - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T \boldsymbol{x}$. Typically the OLS residual r is not estimating the error e and $\tau^2 \neq \sigma^2$, but the following results show that the OLS residual is of great interest for 1D regression models.

Assume that a 1D model holds, $Y \perp\!\!\!\perp \boldsymbol{x} | \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, which is equivalent to $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{\beta}^T \boldsymbol{x}$. Then under regularity conditions, results iii) – v) below hold.

iii) Li and Duan (1989): $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$ for some constant c.

iv) Li and Duan (1989) and Chen and Li (1998):

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \boldsymbol{C}_{OLS}) \tag{2.2}$$

where

$$\boldsymbol{C}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} E[(Y - \alpha_{OLS} - \boldsymbol{\beta}_{OLS}^T \boldsymbol{x})^2 (\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}. \tag{2.3}$$

v) Chen and Li (1998): Let $\boldsymbol{A}$ be a known full rank constant $k \times (p-1)$ matrix. If the null hypothesis H_0: $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ is true, then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} - c\boldsymbol{A}\boldsymbol{\beta}) = \sqrt{n}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T)$$

and

$$\boldsymbol{A}\boldsymbol{C}_{OLS}\boldsymbol{A}^T = \tau^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T. \tag{2.4}$$

If the multiple linear regression (MLR) model holds or if $E[v^2(\boldsymbol{x}-E(\boldsymbol{x}))(\boldsymbol{x}-E(\boldsymbol{x}))^T] = E(v^2)\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then $\boldsymbol{C}_{OLS} = \tau^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$. If the MLR model holds, $\tau^2 = \sigma^2$. To create test statistics, the estimator

$$\hat{\tau}^2 = \mathrm{MSE} = \frac{1}{\mathrm{n-p}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{r}_{\mathrm{i}}^2 = \frac{1}{\mathrm{n-p}}\sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{Y_i} - \hat{\alpha}_{\mathrm{OLS}} - \hat{\boldsymbol{\beta}}_{\mathrm{OLS}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{i}})^2$$

will be useful. The estimator

$$\hat{\boldsymbol{C}}_{OLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}[(Y_i - \hat{\alpha}_{OLS} - \hat{\boldsymbol{\beta}}_{OLS}^T\boldsymbol{x}_i)^2(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T]\right]\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tag{2.5}$$

can also be useful. Notice that for general 1D regression models, the OLS MSE estimates τ^2 rather than the error variance σ^2.

vi) Chen and Li (1998): A test statistic for $H_0 : \boldsymbol{A\beta} = \boldsymbol{0}$ is

$$W_{OLS} = n\hat{\boldsymbol{\beta}}_{OLS}^T\boldsymbol{A}^T[\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS}/\hat{\tau}^2 \xrightarrow{D} \chi_k^2, \tag{2.6}$$

the chi–square distribution with k degrees of freedom.

Before presenting the next result, some results from OLS MLR theory are needed. Let the known $k \times p$ constant matrix $\tilde{\boldsymbol{A}} = [\boldsymbol{a}\ \boldsymbol{A}]$ where $\boldsymbol{a}$ is a $k \times 1$ vector, and let $\boldsymbol{c}$ be a known $k \times 1$ constant vector. Following Seber and Lee (2003, pp. 99–106), the usual F statistic for testing $H_0 : \tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{c}$ is

$$F_0 = \frac{(SSE(H_0) - SSE)/k}{SSE/(n-p)} = (\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})^T[\tilde{\boldsymbol{A}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\tilde{\boldsymbol{A}}^T]^{-1}(\tilde{\boldsymbol{A}}\hat{\boldsymbol{\eta}} - \boldsymbol{c})/(k\hat{\tau}^2) \tag{2.7}$$

where $MSE = \hat{\tau}^2 = SSE/(n-p)$, $SSE = \sum_{i=1}^n r_i^2$ and $SSE(H_0) = \sum_{i=1}^n r_i^2(H_0)$ is the minimum sum of squared residuals subject to the constraint $H_0 : \tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{c}$. Recall that if H_0 is true, the MLR model holds and the errors e_i are iid $N(0, \sigma^2)$, then $F_0 \sim F_{k,n-p}$, the F distribution with k and $n-p$ degrees of freedom. Also recall that if a random variable $Z_n \sim F_{k,n-p}$, then as $n \to \infty$

$$Z_n \xrightarrow{D} \chi^2_k/k. \tag{2.8}$$

Theorem 2.1 below and (2.8) suggest that OLS output, originally meant for testing with the MLR model, can also be used for testing with many 1D regression data sets. Li and Duan (1989) suggest that OLS F tests are asymptotically valid if $\boldsymbol{x}$ is multivariate normal and if $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \boldsymbol{0}$. Freedman (1981), Brillinger (1983) and Chen and Li (1998) also discuss $\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS})$. Let the 1D model $Y \perp\!\!\!\perp \boldsymbol{x}|\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ be written as $Y \perp\!\!\!\perp \boldsymbol{x}|\alpha_R + \boldsymbol{\beta}_R^T\boldsymbol{x}_R + \boldsymbol{\beta}_O^T\boldsymbol{x}_O$ where the reduced model is $Y \perp\!\!\!\perp \boldsymbol{x}|\alpha_R + \boldsymbol{\beta}_R^T\boldsymbol{x}_R$ and $\boldsymbol{x}_O$ denotes the terms outside of the reduced model. Notice that OLS ANOVA F test corresponds to H_0: $\boldsymbol{\beta} = \boldsymbol{0}$ and uses $\boldsymbol{A} = \boldsymbol{I}_{p-1}$. The tests for H_0: $\beta_i = 0$ use $\boldsymbol{A} = (0, ..., 0, 1, 0, ..., 0)$ where the 1 is in the ith position and are equivalent to the OLS t tests. The test H_0: $\boldsymbol{\beta}_O = \boldsymbol{0}$ uses $\boldsymbol{A} = [\boldsymbol{0}\ \boldsymbol{I}_j]$ if $\boldsymbol{\beta}_O$ is a $j \times 1$ vector, and the test statistic (2.7) can be computed by running OLS on the full model to obtain SSE and on the reduced model to obtain $SSE(R) \equiv SSE(H_0)$.

In the theorem below, it is crucial that H_0: $\boldsymbol{A\beta} = \boldsymbol{0}$. Tests for H_0: $\boldsymbol{A\beta} = \boldsymbol{1}$, say, may not be valid.

Theorem 2.1. Assume that a 1D regression model (1.2) holds and that Equation (2.6) holds when $H_0 : \boldsymbol{A\beta} = \boldsymbol{0}$ is true. Then as $n \to \infty$, the test statistic (2.7) satisfies

$$F_0 = \frac{n-1}{kn} W_{OLS} \xrightarrow{D} \chi^2_k/k.$$

To see this, notice that by (2.6), the result follows if $F_0 = (n-1)W_{OLS}/(kn)$. Let $\tilde{\boldsymbol{A}} = [\boldsymbol{0}\ \boldsymbol{A}]$ so that H_0:$\tilde{\boldsymbol{A}}\boldsymbol{\eta} = \boldsymbol{0}$ is equivalent to H_0:$\boldsymbol{A\beta} = \boldsymbol{0}$. Following Seber and Lee

(2003, p. 150),

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \tag{2.9}$$

where the $(p-1)\times(p-1)$ matrix

$$\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1). \tag{2.10}$$

Using $\tilde{\boldsymbol{A}}$ and (2.9) in (2.7) shows that

$$F_0 = (\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS})^T \left[[\boldsymbol{0}\ \boldsymbol{A}] \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{0}^T \\ \boldsymbol{A}^T \end{pmatrix} \right]^{-1} \boldsymbol{A}\hat{\boldsymbol{\beta}}_{OLS}/(k\hat{\tau}^2),$$

and the result follows from (2.10) after algebra.

Following Chen and Li (1998), SIR produces eigenvalues $\hat{\lambda}_i$ and associated SIR directions $\hat{\boldsymbol{\beta}}_{i,SIR}$ for $i = 1, ..., p-1$. The SIR directions $\hat{\boldsymbol{\beta}}_{i,SIR}$ for $i = 1, ..., d$ are used for dD regression. The following theory for a SIR t type test holds under regularity conditions.

vii) Chen and Li (1998): For a 1D regression and vector $\boldsymbol{A}$, a test statistic for H_0 : $\boldsymbol{A}\boldsymbol{\beta}_1 = \boldsymbol{0}$ is

$$W_S = n\hat{\boldsymbol{\beta}}_{1,SIR}^T\boldsymbol{A}^T[\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{1,SIR}/[(1-\hat{\lambda}_1)/\hat{\lambda}_1] \xrightarrow{D} \chi_1^2. \tag{2.11}$$

3 Resistant DR Methods

Ellipsoidal trimming can be used to create outlier resistant DR methods that can give useful results when the assumption of linearly related predictors is violated. To perform ellipsoidal trimming, a robust estimator of multivariate location and dispersion $(T, \boldsymbol{C})$ is computed and used to create the squared Mahalanobis distances $D_i^2 \equiv D_i^2(T, \boldsymbol{C}) =$

$(\boldsymbol{x}_i - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x}_i - T)$ for each vector of observed predictors $\boldsymbol{x}_i$. If the ordered distance $D_{(j)}$ is unique, then j of the $\boldsymbol{x}_i$'s are in the ellipsoid $\{\boldsymbol{x} : (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) \leq D^2_{(j)}\}$. The ith case $(Y_i, \boldsymbol{x}_i)$ is trimmed if $D_i > D_{(j)}$. For example, if $j \approx 0.9n$, then about M% = 10% of the cases are trimmed, and a DR method can be computed from the cases $(Y_M, \boldsymbol{x}_M)$ that remain.

Several authors have noted that applying DR methods to a subset $(Y_M, \boldsymbol{x}_M)$ of the data with the $\boldsymbol{x}_M$ distribution closer to being elliptically contoured is an effective method for making DR methods such as OLS and SIR resistant to the presence of strong nonlinearities. See Li and Duan (1989, p. 1011), Brillinger (1991), Cook (1994, p. 188; 1998a, p. 152) and Li, Cook and Nachtsheim (2004).

The Olive (2004a) MBA estimator $(T_{MBA}, \boldsymbol{C}_{MBA})$ will be used for $(T, \boldsymbol{C})$. To compute MBA, first compute the classical estimator $(T_1, \boldsymbol{C}_1)$ and the classical estimator $(T_2, \boldsymbol{C}_2)$ computed from the half set of cases closest to the coordinatewise median in Euclidean distance. The distances based on $(T_i, \boldsymbol{C}_i)$ are computed and the classical estimator is computed from the half set with the smallest distances. This step is iterated 5 times resulting in two estimators. Then the estimator with the smallest determinant is scaled to be consistent at multivariate normal data. The scaled estimator is the MBA estimator which is not affine equivariant.

Cook and Nachtsheim (1994) and Olive (2002, 2004b) used ellipsoidal trimming with alternative estimators, so it is important to explain why MBA should be used. First, the MBA estimator is far faster than the Rousseeuw and Van Driessen (1999) FAST-MCD estimator used by Olive (2002). Second, Olive (2007, ∮ 10.7) shows that the MBA estimator is a $\sqrt{n}$ consistent high breakdown estimator of the same quantity estimated by the minimum covariance determinant (MCD) estimator if $\boldsymbol{x}$ has a nondegenerate elliptically contoured distribution with second moments. Third, MBA can be regarded as a high breakdown 0-1 weighting method for transforming data towards an EC distribution since if the data distribution is not EC, the MBA estimator finds a compact 50% covering ellipsoid such that the distribution of the covered cases is closer to being EC. No other published estimator of multivariate location and dispersion has these three properties.

The choice of M is important, and the Rousseeuw and Van Driessen (1999) DD plot of classical Mahalanobis distances MD_i vs MBA distances RD_i can be used to choose M. The MD_i use $(T, \boldsymbol{C}) = (\overline{\boldsymbol{x}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})$. Olive (2002) shows that the plotted points in the DD plot will follow the identity line with zero intercept and unit slope if the predictor distribution is MVN, and will follow a line with zero intercept but non–unit slope if the distribution is EC with second moments but not MVN. Delete M% of the cases with the largest MBA distances so that the remaining cases follow the identity line (or some line through the origin) closely. Let $(Y_{Mi}, \boldsymbol{x}_{Mi})$ denote the data remaining after trimming where $i = 1, ..., n_M$. Then apply the DR method on these n_M cases.

As long as M is chosen only using the predictors, DR theory will apply if the data $(Y_M, \boldsymbol{x}_M)$ satisfies the regularity conditions. Let $\phi_M = \lim_{n\to\infty} n/n_M$, let c_M be a constant and let $\hat{\boldsymbol{\beta}}_{DM}$ denote a DR estimator applied to $(Y_{Mi}, \boldsymbol{x}_{Mi})$ with

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{DM} - c_M\boldsymbol{\beta}) = \frac{\sqrt{n}}{\sqrt{n_M}}\sqrt{n_M}(\hat{\boldsymbol{\beta}}_{DM} - c_M\boldsymbol{\beta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \phi_M \boldsymbol{C}_{DM}). \tag{3.1}$$

If $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{0}$ is true and $\hat{\boldsymbol{C}}_{DM}$ is a consistent estimator of $\boldsymbol{C}_{DM}$, then

$$W_{DM} = n_M \hat{\boldsymbol{\beta}}_{DM}^T \boldsymbol{A}^T [\boldsymbol{A}\hat{\boldsymbol{C}}_{DM}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}_{DM} \xrightarrow{D} \chi^2_k.$$

For example, if the MLR model is valid and the errors are iid $N(0, \sigma^2)$, then the OLS estimator

$$\hat{\boldsymbol{\eta}}_M = (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T \boldsymbol{Y}_M \sim N_p(\boldsymbol{\eta}, \sigma^2 (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1}).$$

A tradeoff is that low amounts of trimming may not work while large amounts of trimming may be inefficient (see (3.1)). For 1D models, Olive (2002, 2004b, 2005) suggested plotting $\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}$ versus Y for $M = 0, 10, ..., 90$ and choosing M_{TV} such that the plot (called a trimmed view or estimated sufficient summary plot) has a smooth mean function and the smallest variance function. Suppose $\sqrt{n}(\hat{\boldsymbol{\beta}}_M - c_M \boldsymbol{\beta}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{C}_M)$ for $M = 0, 10, ..., 90$. Then $\hat{\boldsymbol{\beta}}_{M,TV}$ is $\sqrt{n}$ consistent if $c_M \equiv c_0$, e.g., for MLR $c_M \equiv 1$. But if $\hat{\boldsymbol{\beta}}_{M,TV}$ oscillates between $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_{10}$, then $\hat{\boldsymbol{\beta}}_{M,TV}$ need not be asymptotically normal. If there is oscillation and the c_M are not equal, then $\hat{\boldsymbol{\beta}}_{M,TV}$ is inconsistent.

Adaptive trimming can be used to obtain an asymptotically normal estimator that may avoid large efficiency losses. First, choose an initial amount of trimming M_I by using, e.g., the DD plot or trimmed views. Let $\hat{\boldsymbol{\beta}}$ denote the first direction of the DR method. Next compute $|corr(\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}, \hat{\boldsymbol{\beta}}_{M_I}^T \boldsymbol{x})|$ for $M = 0, 10, ..., 90$ and find the smallest value $M_A \leq M_I$ such that the absolute correlation is greater than 0.95. If no such value exists, then use $M_A = M_I$. The resulting adaptive trimming estimator is asymptotically equivalent to the estimator that uses 0% trimming if $\hat{\boldsymbol{\beta}}_0$ is a consistent estimator of $c_0 \boldsymbol{\beta}$ and if $\hat{\boldsymbol{\beta}}_{M_I}$ is a consistent estimator of $c_{M_I} \boldsymbol{\beta}$ for $c_0 \neq 0$ and $c_{M_I} \neq 0$.

4 Examples and Simulations

Example 4.1. The Buxton (1920) data consists of measurements taken on 87 men. Let *height* be the response. Figure 1a shows the DD plot made from the four predictors *head length, nasal height, bigonal breadth,* and *cephalic index.* The five massive outliers correspond to head lengths that were recorded to be around 5 feet. Figure 1b shows that after deleting these points, the predictor distribution is much closer to a multivariate normal distribution. Now DR methods can be used to investigate the regression.

In a small simulation, the clean data $Y = (\alpha + \boldsymbol{\beta}^T \boldsymbol{x})^3 + e$ where $\alpha = 1, \boldsymbol{\beta} = (1,0,0,0)^T, e \sim N(0,1)$ and $\boldsymbol{x} \sim N_4(\boldsymbol{0}, \boldsymbol{I}_4)$. The outlier percentage γ was either 0% or 49%. The 2 clusters of outliers were about the same size and had $Y \sim N(0,1)$, $\boldsymbol{x} \sim N_4(\pm 10(1,1,1,1)^T, \boldsymbol{I}_4)$. Table 1 records the averages of $\hat{\beta}_i$ over 100 runs where the DR method used $M = 0$ or $M = 50\%$ trimming. SIR, SAVE and PHD were very similar except when $\gamma = 49$ and $M = 0$. When outliers were present, the average of $\hat{\boldsymbol{\beta}}_{F,50} \approx c_F(1,0,0,0)^T$ where c_F depended on the DR method and F was OLS, SIR, SAVE or PHD. The sample size $n = 1000$ was used although OLS gave reasonable estimates for much smaller sample sizes. The collection of functions `rpack`, available from (www.math.siu.edu/olive/rpack.txt) contains a function *drsim7* that can be used to duplicate the simulation in `R`. Olive (2007, ∮ 14.2) explains how to use `rpack` and how to download the Weisberg (2002) `dr` library from (www.r-project.org/#doc).

The following example shows that ellipsoidal trimming can be useful for DR when $\boldsymbol{x}$ is not EC. There is a myth that transforming predictors is free, but using a log transformation for the example below will destroy the 1D structure.

Example 4.2. The artificial data set `sinc.lsp` is available from (www.math.siu.edu/olive/ol-bookp.htm). It contains 200 trivariate vectors $\boldsymbol{x}_i$ such that the marginal distributions are iid lognormal. The response $Y_i = \sin(\boldsymbol{\beta}^T \boldsymbol{x}_i)/\boldsymbol{\beta}^T \boldsymbol{x}_i$ where $\boldsymbol{\beta} = (1,2,3)^T$. Trimming with $M = 0, 10, ..., 90$ was examined. Weisberg (2002) was used (with 4 slices) to produce the SIR, PHD and SAVE estimators. Table 2 shows the estimated coefficients $\hat{\boldsymbol{\beta}}$ when the true coefficients are $c(1,2,3)^T$. Trimming greatly improved the SIR, SAVE and PHD estimators of $c\boldsymbol{\beta}$. Figure 2 shows the trimmed views for 0% trimming and Figure 3 shows the trimmed views estimators where ESP $= \hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}$ or $\hat{\alpha}_M + \hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}$.

The following simulation study is extracted from Chang (2006) who used eight types of predictor distributions: d1) $\boldsymbol{x} \sim N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1})$, d2) $\boldsymbol{x} \sim 0.6N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1}) + 0.4N_{p-1}(\mathbf{0}, 25\boldsymbol{I}_{p-1})$, d3) $\boldsymbol{x} \sim 0.4N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1}) + 0.6N_{p-1}(\mathbf{0}, 25\boldsymbol{I}_{p-1})$, d4) $\boldsymbol{x} \sim 0.9N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1}) + 0.1N_{p-1}(\mathbf{0}, 25\boldsymbol{I}_{p-1})$, d5) $\boldsymbol{x} \sim LN(\mathbf{0}, \boldsymbol{I})$ where the marginals are iid lognormal(0,1), d6) $\boldsymbol{x} \sim MVT_{p-1}(3)$, d7) $\boldsymbol{x} \sim MVT_{p-1}(5)$ and d8) $\boldsymbol{x} \sim MVT_{p-1}(19)$. Here $\boldsymbol{x}$ has a multivariate t distribution $\boldsymbol{x}_i \sim MVT_{p-1}(\nu)$ if $\boldsymbol{x}_i = \boldsymbol{z}_i/\sqrt{W_i/\nu}$ where $\boldsymbol{z}_i \sim N_{p-1}(\mathbf{0}, \boldsymbol{I}_{p-1})$ is independent of the chi–square random variable $W_i \sim \chi^2_\nu$. Of the eight distributions, only d5) is not elliptically contoured. The MVT distribution gets closer to the MVN distribution d1) as $\nu \to \infty$. The MVT distribution has first moments for $\nu \geq 3$ and second moments for $\nu \geq 5$. See Johnson and Kotz (1972, pp. 134-135). All simulations used 1000 runs.

The simulations for single index models used $\alpha = 1$. Let the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. Then the seven models considered were m1) $Y = SP + e$, m2) $Y = (SP)^2 + e$, m3) $Y = \exp(SP) + e$, m4) $Y = (SP)^3 + e$, m5) $Y = \sin(SP)/SP + 0.01e$, m6) $Y = SP + \sin(SP) + 0.1e$ and m7) $Y = \sqrt{|SP|} + 0.1e$ where $e \sim N(0,1)$. Models m2), m3) and m4) can result in large $|Y|$ values which can cause numerical difficulties for OLS if $\boldsymbol{x}$ is heavy tailed.

First, coefficient estimation was examined with $\boldsymbol{\beta} = (1,1,1,1)^T$, and for OLS the sample standard deviation (SD) of each entry $\hat{\beta}_{Mi,j}$ of $\hat{\boldsymbol{\beta}}_{M,j}$ was computed for $i = 1,2,3,4$ with

$j = 1, ..., 1000$. For each of the 1000 runs, the Chen and Li (1998) formula $SE_{cl}(\hat{\beta}_{Mi}) = \sqrt{n_M^{-1}(\hat{\boldsymbol{C}}_M)_{ii}}$ was computed where $\hat{\boldsymbol{C}}_M$ is the estimate (2.5) applied to $(Y_M, \boldsymbol{x}_M)$. The average of $\hat{\boldsymbol{\beta}}_M$ and of $\sqrt{n}SE_{cl}$ were recorded as well as $\sqrt{n}SD$ of $\hat{\beta}_{Mi,j}$ under the labels $\overline{\boldsymbol{\beta}}_M$, $\sqrt{n}\ \ \overline{SE}_{cl}$ and $\sqrt{n}SD$. Under regularity,

$$\sqrt{n}\ \ \overline{SE}_{cl} \approx \sqrt{n}SD \approx \sqrt{\frac{1}{1 - \frac{M}{100}} \operatorname{diag}(\boldsymbol{C}_{\mathrm{M}})}$$

where $\boldsymbol{C}_M$ is (2.3) applied to $(Y_M, \boldsymbol{x}_M)$.

For MVN $\boldsymbol{x}$, MLR and 0% trimming, all three recorded quantities were near (1,1,1,1) for $n = 60, 500$, and 1000. For 90% trimming and $n = 1000$, the results were $\overline{\boldsymbol{\beta}}_{90} = (1.00, 1.00, 1.01, 0.99)$, $\sqrt{n}\ \ \overline{SE}_{cl} = (7.56, 7.61, 7.60, 7.54)$ and $\sqrt{n}SD = (7.81, 8.02, 7.76, 7.59)$, suggesting that $\hat{\boldsymbol{\beta}}_{90}$ is asymptotically normal but inefficient.

For other distributions, Chang (2006) recorded results for 0 and 10% trimming as well as a "good" trimming value M_B. Results are "good" if all of the entries of both $\overline{\boldsymbol{\beta}}_{M_B}$ and $\sqrt{n}\ \overline{SE}_{cl}$ were approximately equal and if the theoretical $\sqrt{n}\ \overline{SE}_{cl}$ was close to the simulated $\sqrt{n}SD$. The results were good for MVN $\boldsymbol{x}$ and all seven models, and the results were similar for $n = 500$ and $n = 1000$. The results were good for models m1 and m5 for all eight distributions. Model m6 was good for 0% trimming except for distribution d5 and model m7 was good for 0% trimming except for distributions d5, d6 and d7. Trimming usually helped for models m2, m3 and m4 for distributions d5 - d8. For $n = 500$ and OLS, Table 3 shows that $\hat{\boldsymbol{\beta}}_M$ estimates $c_M\boldsymbol{\beta}$ and the average of the Chen and Li (1998) SE is often close to the simulated SD.

For SIR with $h = 4$ slices $\overline{\boldsymbol{\beta}}_M$ was recorded. Chang (2006) shows that the SIR results were similar to those for OLS, but often more trimming and larger sample sizes were needed than those for OLS. Much of the literature suggests that SIR is insensitive to h, but our simulations suggest that the results depend on h in that the largest sample sizes were needed for $h = 2$ slices and then for 3 slices.

Next testing was considered. Let F_M denote the OLS statistic (2.7) applied to the n_M cases $(Y_M, \boldsymbol{x}_M)$ that remained after trimming. H_0 was rejected if $F_M > F_{k,n_M-p}(0.95)$. Let W_M denote the SIR statistic (2.11) except that $\boldsymbol{A}$ is the same matrix used for OLS. Then H_0 was rejected if $W_M > \chi^2_k(0.95)$ although theory is only given for $k = 1$. As h increased from 2 to 3 to 4, $\hat{\lambda}_1$ and the SIR chi–square test statistic W_0 rapidly increased. For $h > 4$ the increase was much slower. For 1D models, 2 slices were used since otherwise H_0 was rejected too often.

For testing the nominal level was 0.05, and we recorded the proportion $\hat{p}$ of runs where H_0 was rejected. Since 1000 runs were used, the count $1000\hat{p} \sim$ binomial(1000, $1 - \delta_n$) where $1 - \delta_n$ converges to the true large sample level $1 - \delta$. The standard error for the proportion is $\sqrt{\hat{p}(1-\hat{p})/1000} \approx 0.0069$ for $p = 0.05$. An observed coverage $\hat{p} \in (0.03, 0.07)$ suggests that there is no reason to doubt that the true level is 0.05.

Let $Y = m(\alpha, \boldsymbol{\beta}_1^T\boldsymbol{x}, ..., \boldsymbol{\beta}_{p-1}^T\boldsymbol{x}) + e$. For a 0D regression, this reduces to $Y = m(\alpha, 0, ..., 0)$ $+ e = c_\alpha + e$. The 0D assumption can be tested with $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{\beta} \neq \boldsymbol{0}$ (use $\boldsymbol{\beta}_1$ for SIR), and the OLS F statistic (2.7) and SIR W statistic (2.11) are invariant with respect to a constant. Hence this test is interesting because the results do not depend on the model, but only on the distribution of $\boldsymbol{x}$ and the distribution of e. Since $\boldsymbol{\beta}_{OLS} \in S(\boldsymbol{B})$, power can be good if $\boldsymbol{\beta}_{OLS} \neq \boldsymbol{0}$. The OLS test is equivalent to the ANOVA F test from MLR of Y on $\boldsymbol{x}$. Under H_0, the test should perform well provided that the design matrix is nonsingular and the error distribution and sample size are such that the central limit theorem holds. Table 4 shows the results for OLS and SIR for $n = 100, 500$ and for the eight different distributions. Since the true model was linear and normal, the exact OLS level is 0.05 even for $n = 10$. Table 4 shows that OLS performed as expected while SIR only gave good results for MVN $\boldsymbol{x}$. Notice that Table 4 shows that OLS is useful for testing 0D structure even if a dD additive error model is assumed.

Next the test $H_0 : \beta_2 = 0$ was considered. The OLS test is equivalent to the t test from MLR of Y on $\boldsymbol{x}$. The true model used $\alpha = 1$ and $\boldsymbol{\beta} = (1, 0, 1, 1)^T$. To simulate adaptive trimming, $|corr(\hat{\boldsymbol{\beta}}_M^T \boldsymbol{x}, \boldsymbol{\beta}^T \boldsymbol{x})|$ was computed for $M = 0, 10, ..., 90$ and the initial trimming proportion M_I maximized this correlation. This process should be similar to choosing the best trimmed view by examining 10 plots. The rejection proportions were recorded for $M = 0, ..., 90$ and for adaptive trimming. Chang (2006) used the seven models, eight distributions and sample sizes $n = 60, 150$, and 500. Table 5 shows some results for $n = 150$.

For OLS, the test that used adaptive trimming had proportions ≤ 0.072 except for model m4 with distributions d2, d3, d4, d6, d7 and d8; m2 with d4, d6 and d7 for n = 500 and d6 with n = 150; m6 with d4 and n = 60, 150; m5 with d7 and n = 500 and m7 with d7 and n = 500. With the exception of m4, when the adaptive $\hat{p} > 0.072$, then 0% trimming had a rejection proportion near 0.1. Occasionally adaptive trimming was conservative with $\hat{p} < 0.03$. The 0% trimming worked well for m1 and m6 for all eight distributions and for d1 and d5 for all seven models. Models m2 and m3 usually benefited from adaptive trimming. For distribution d1, the adaptive and 0% trimming methods had identical $\hat{p}$ for $n = 500$ except for m3 where the values were 0.038 and 0.042. Table 5 supports the claim that the adaptive trimming estimator can be asymptotically equivalent to the non-resistant DR method (0% trimming) and that trimming can greatly improve the type I error.

For SIR results were not as good. Adaptive trimming failed more often than it worked, and failed for model m1. 0% trimming performed well for all seven models for the MVN distribution d1, and there was always an M such the W_M did not reject H_0 too often.

The `rpack` function *drsim5* can be used to simulate the OLS and SIR tests while *drsim6* can be used to simulate tests based on adaptive and 0% trimming for OLS.

5 Conclusions

For the first time, there are robust multivariate location and dispersion estimators (e.g., MBA), which have been shown to be both fast and to have good theoretical properties. For example MCD has computational complexity greater than $O(n^p)$ while no theory for FAST-MCD has been given. By using MBA, outliers in the predictors can be found, the assumption that $\boldsymbol{x}$ is EC can be checked, and resistant DR methods that have theory similar to the non–resistant method can be created.

The DD plot should be used to detect outliers and influential cases for regressions with continuous predictors as well as for multivariate analysis. The DD plot is also a diagnostic for the linearity condition since for EC data the plotted points will follow a line through the origin, and for MVN data the plotted points will follow the identity line. In the case of no outliers, power transformations may be used to remove nonlinearities from the predictors and to transform the predictor distribution towards a MVN distribution. A DD plot and scatterplot matrix (when p is not too large) may then be useful tools for determining the success of such transformations.

Ellipsoidal trimming can be used to make many DR methods resistant to $\boldsymbol{x}$ outliers. For 1D regression, the plot of $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$ versus Y is crucial for visualizing the regression $Y|\boldsymbol{x}$ (where $\hat{\alpha} \equiv 0$ may be used). For 2D regression, plot Y versus the two estimated sufficient predictors $\hat{\boldsymbol{\beta}}_1^T \boldsymbol{x}$ and $\hat{\boldsymbol{\beta}}_2^T \boldsymbol{x}$ and spin the plot. Trimming combined with these plots makes the DR methods resistant to Y and $\boldsymbol{x}$ outliers.

For 1D regression models, Theorem 2.1 and the simulations suggest that OLS software is useful for fast exploratory data analysis. To use the OLS output, the assumption that OLS is a useful estimator for the 1D model needs to be checked. In addition to plotting the ESP versus Y, additional methods for checking OLS are suggested by Olive and Hawkins (2005) who showed that variable selection methods, originally meant for MLR and based on OLS and the Mallows' C_p criterion, can also be used for 1D models. Since the C_p statistic is a one to one function of the F statistic for testing the submodel, Theorem 2.1 provides additional support for using OLS for variable selection for 1D models. After using OLS for exploratory analysis, alternative 1D methods can be tried.

For single index models with EC $\boldsymbol{x}$, OLS can fail if m is symmetric about the median θ of the distribution of $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$. If m is symmetric about a, then OLS may become effective as $|\theta - a|$ gets large. This fact is often overlooked in the literature and is demonstrated by models m7), m5) and m2) where $Y = (SP)^2 + e$ with $\theta = \alpha = 1$. OLS has trouble with $Y = (SP - a)^2 + e$ as a gets close to $\theta = 1$. The type of symmetry where OLS fails is easily simulated, but may not occur often in practice.

Tests developed for parametric models such as the deviance tests for GLMs will often have more power than the "model free" OLS tests. Simonoff and Tsai (2002) suggest tests for single index models while Cook (2004) develops "model free" tests for model (1.1).

There are many DR methods and methods tailored for 1D regression. As software becomes available, the three *drsim* functions from `rpack` can be modified to examine the effect of ellipsoidal trimming on these methods. Power of the OLS tests can also be examined by modifying the data so that the null hypothesis does not hold. Chang (2006) has much more extensive simulation results, including a simulation for a 2D model.

The simulations suggest that for 1D models, much larger sample sizes are needed to estimate the sufficient predictor for SIR, SAVE and PHD than for OLS. The Chen and Li (1998) t type tests for SIR were ineffective unless $\boldsymbol{x}$ was MVN and the number of slices h was small. Some earlier studies suggest that SIR is insensitive to the value of h. Wang, Ni and Tsai (2007) showed that p-values for SIR t type tests were accurate for MVN $\boldsymbol{x}$, but not for other EC distributions. Their technique of contour projection followed by SIR greatly improved inference.

6 References

Brillinger, D.R., 1977. The identification of a particular nonlinear time series. Biometrika 64, 509-515.

Brillinger, D.R., 1983. A generalized linear model with "Gaussian" regressor variables. In: Bickel, P.J., Doksum, K.A., Hodges, J.L. (Eds.), A Festschrift for Erich L. Lehmann. Wadsworth, Pacific Grove, CA, 97-114.

Brillinger, D.R., 1991. Comment on 'Sliced inverse regression for dimension reduction,' by K.C. Li. J. Amer. Statist. Assoc. 86, 333.

Buxton, L.H.D., 1920. The anthropology of Cyprus. J. R. Anthropological Institute of Great Britain and Ireland 50, 183-235.

Chang, J., 2006. Resistant Dimension Reduction. Ph.D. Thesis, Southern Illinois University, online at (www.math.siu.edu/olive/sjingth.pdf).

Chen, C.H., Li, K.C., 1998. Can SIR be as popular as multiple linear regression? Statist. Sinica 8, 289-316.

Čížek, P., Härdle, W., 2006. Robust estimation of dimension reduction space. Computat. Statist. Data Anal. 51, 545-555.

Cook, R.D., 1994. On the interpretation of regression plots. J. Amer. Statist. Assoc. 89, 177-189.

Cook, R.D., 1998a. Regression Graphics: Ideas for Studying Regression Through Graphics. Wiley, New York.

Cook, R.D., 1998b. Principal hessian directions revisited. J. Amer. Statist. Assoc. 93, 84-100.

Cook, R.D., 2004. Testing predictor contributions in sufficient dimension reduction. Ann. Statist. 32, 1062-1092.

Cook, R.D., Critchley, F., 2000. Identifying outliers and regression mixtures graphically. J. Amer. Statist. Assoc. 95, 781-794.

Cook, R.D., Li, B., 2002. Dimension reduction for conditional mean in regression. Ann. Statist. 30, 455-474.

Cook, R.D., Li, B., 2004. Determining the dimension of iterative hessian transformation. Ann. Statist. 32, 2501-2531.

Cook, R.D., Nachtsheim, C.J., 1994. Reweighting to achieve elliptically contoured covariates in regression. J. Amer. Statist. Assoc. 89, 592-599.

Cook, R.D., Ni, L., 2005. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. J. Amer. Statist. Assoc. 100, 410-428.

Cook, R.D., Weisberg, S., 1991. Comment on 'Sliced inverse regression for dimension reduction' by K.C. Li. J. Amer. Statist. Assoc. 86, 328-332.

Freedman, D.A., 1981. Bootstrapping regression models. Ann. Statist. 9, 1218-1228.

Gather, U., Hilker, T., Becker, C., 2001. A robustified version of sliced inverse regression. In: Fernholtz, T.L., Morgenthaler, S. Stahel, W. (Eds.) Statistics in Genetics and in the Environmental Sciences. Birkhäuser, Basel, Switzerland, pp. 145-157.

Gather, U., Hilker, T., Becker, C., 2002. A note on outlier sensitivity of sliced inverse regression. Statistics 36, 271-281.

Hall, P., Li, K.C., 1993. On almost linearity of low dimensional projections from high dimensional data. Ann. Statist. 21, 867-889.

Heng-Hui, L., 2001. A study of sensitivity analysis on the method of principal hessian directions. Computat. Statist. 16, 109-130.

Johnson, N.L., Kotz, S., 1972. Distributions in Statistics: Continuous Multivariate Distributions. 1st ed. Wiley, New York.

Li, K.C., 1991. Sliced inverse regression for dimension reduction. J. Amer. Statist. Assoc. 86, 316-342.

Li, K.C., 1992. On principal hessian directions for data visualization and dimension reduction: another application of Stein's lemma. J. Amer. Statist. Assoc. 87, 1025-1040.

Li, K.C., 2000. High Dimensional Data Analysis via the SIR/PHD Approach. Unpublished manuscript, online at (www.stat.ucla.edu/ kcli/).

Li, K.C., Duan, N., 1989. Regression analysis under link violation. Ann. Statist. 17, 1009-1052.

Li, L., Cook, R.D., Nachtsheim, C.J., 2004. Cluster-based estimation for sufficient dimension reduction. Computat. Statist. Data Anal. 47, 175-193.

Li, Y., Zhu, L.-X., 2007. Asymptotics for sliced average variance estimation. Ann. Statist. 35, 41-69.

Olive, D.J., 2002. Applications of robust distances for regression. Technometrics 44, 64-71.

Olive, D.J., 2004a. A resistant estimator of multivariate location and dispersion. Computat. Statist. Data Anal. 46, 99-102.

Olive, D.J., 2004b. Visualizing 1D regression. In: Hubert, M., Pison, G., Struyf, A., Van Aelst, S. (Eds.), Theory and Applications of Recent Robust Methods. Birkhäuser, Basel, Switzerland, 221-233.

Olive, D.J., 2005. Two simple resistant regression estimators. Computat. Statist. Data Anal. 49, 809-819.

Olive, D.J., 2007. Applied Robust Statistics. Unpublished text, online at (www.math.siu.edu/olive/).

Olive, D.J., Hawkins, D.M., 2005. Variable selection for 1D regression models. Technometrics 47, 43-50.

Prendergast, L.A., 2005. Influence functions for sliced inverse regression. Scandinavian J. Statist. 32, 385-404.

Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212-223.

Seber, G.A.F., Lee, A.J. 2003. Linear Regression Analysis. second ed. Wiley, New York.

Simonoff, J.S., Tsai, C., 2002. Score tests for the single index model. Technometrics 44, 142-151.

Wang, H., Ni, L., Tsai, C.-L., 2007. Improving dimension reduction via contour–projection. Statist. Sinica, to appear.

Weisberg, S., 2002. Dimension reduction regression in R.
J. Statist. Software 7, online at (www.jstatsoft.org/v07/i01/drdoc.pdf).

Xia, Y., Tong, H., Li, W.K., Zhu, L.-X., 2002. An adaptive estimation of dimension reduction space. J. Royal Statist. Soc. B 64, 363-410.

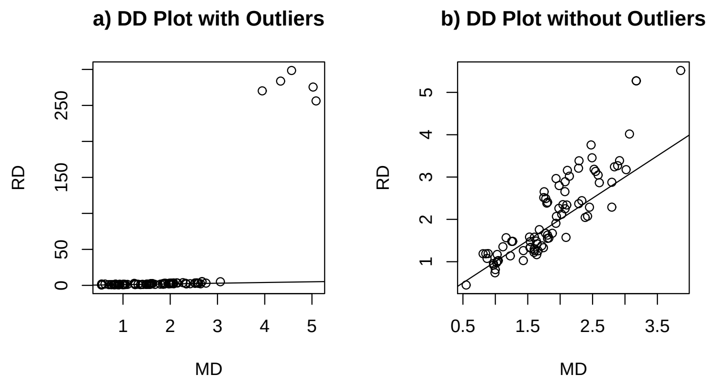

Figure 1: DD Plots for Buxton Data

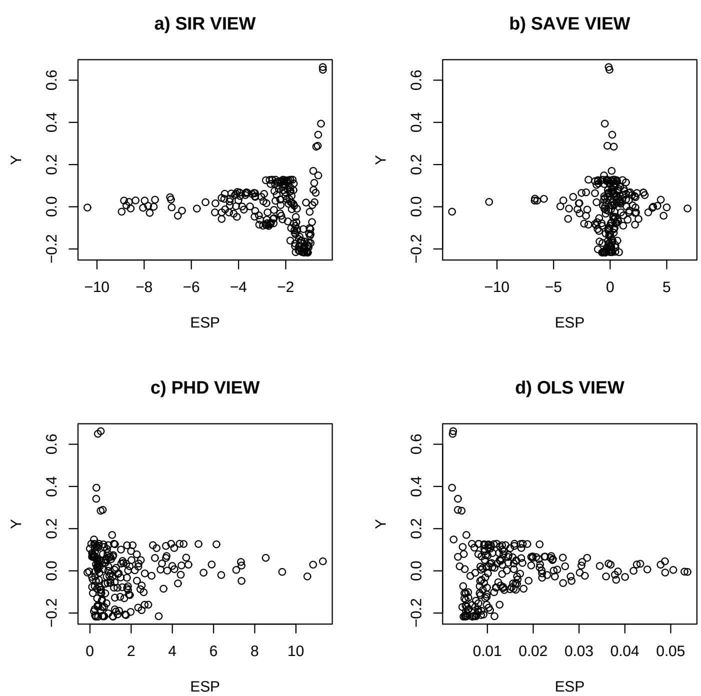

Figure 2: Estimated Sufficient Summary Plots Without Trimming

Table 1: DR Coefficient Estimation with Trimming

type	γ	M	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
SIR	0	0	.0400	.0021	−.0006	.0012
SIR	0	50	−.0201	−.0015	.0014	.0027
SIR	49	0	.0004	−.0029	−.0013	.0039
SIR	49	50	−.0798	−.0014	.0004	−.0015
SAVE	0	0	.0400	.0012	.0010	.0018
SAVE	0	50	−.0201	−.0018	.0024	.0030
SAVE	49	0	−.4292	−.2861	−.3264	−.3442
SAVE	49	50	−.0797	−.0016	−.0006	−.0024
PHD	0	0	.0396	−.0009	−.0071	−.0063
PHD	0	50	−.0200	−.0013	.0024	.0025
PHD	49	0	−.1068	−.1733	−.1856	−.1403
PHD	49	50	−.0795	.0023	.0000	−.0037
OLS	0	0	5.974	.0083	−.0221	.0008
OLS	0	50	4.098	.0166	.0017	−.0016
OLS	49	0	2.269	−.7509	−.7390	−.7625
OLS	49	50	5.647	.0305	.0011	.0053

Table 2: DR Coefficient Estimation of $c(1,2,3)^T$

method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
OLS	0.0032	0.0011	0.0047
M = 90% OLS	0.0321	0.0366	0.2329
SIR	-0.4066	-0.3916	-0.8254
M=10% SIR	0.3032	0.5003	0.8110
SAVE	0.0845	-0.7280	0.6804
M = 60% SAVE	-0.2116	-0.5657	-0.7970
PHD	0.9995	0.0097	-0.0316
M = 60% PHD	-0.2928	-0.6154	-0.7318

Table 3: OLS Coefficient Estimation with Trimming

m	$\boldsymbol{x}$	M	$\overline{\boldsymbol{\beta}}_M$	$\sqrt{n}\,\overline{SE}_{cl}$	$\sqrt{n}\,SD$
m2	d1	0	2.00,2.01,2.00,2.00	7.81,7.79,7.76,7.80	7.87,8.00,8.02,7.88
m3	d2	50	9.06,9.05,9.04,9.08	37.56,37.00,37.31,37.41	55.35,54.02,53.35,55.03
m4	d3	0	291.9,294.0,293.7,292.1	859.7,866.6,877.9,850.8	933.0,957.9,964.9,957.2
m5	d4	0	$-.03, -.03, -.03, -.03$	.30,.30,.30,.30	.31,.32,.33,.31
m6	d5	0	1.04,1.04,1.04,1.04	.36,.36,.37,.37	.41,.42,.42,.40
m7	d6	10	.11,.11,.11,.11	.58,.57,.57,.57	.60,.58,.62,.61

Table 4: Rejection Proportions for H_0: $\boldsymbol{\beta} = \mathbf{0}$

$\boldsymbol{x}$	n	F	SIR	n	F	SIR
d1	100	0.041	0.057	500	0.050	0.048
d2	100	0.050	0.908	500	0.045	0.930
d3	100	0.047	0.955	500	0.050	0.930
d4	100	0.045	0.526	500	0.048	0.599
d5	100	0.055	0.621	500	0.061	0.709
d6	100	0.042	0.439	500	0.036	0.472
d7	100	0.054	0.214	500	0.047	0.197
d8	100	0.044	0.074	500	0.060	0.077

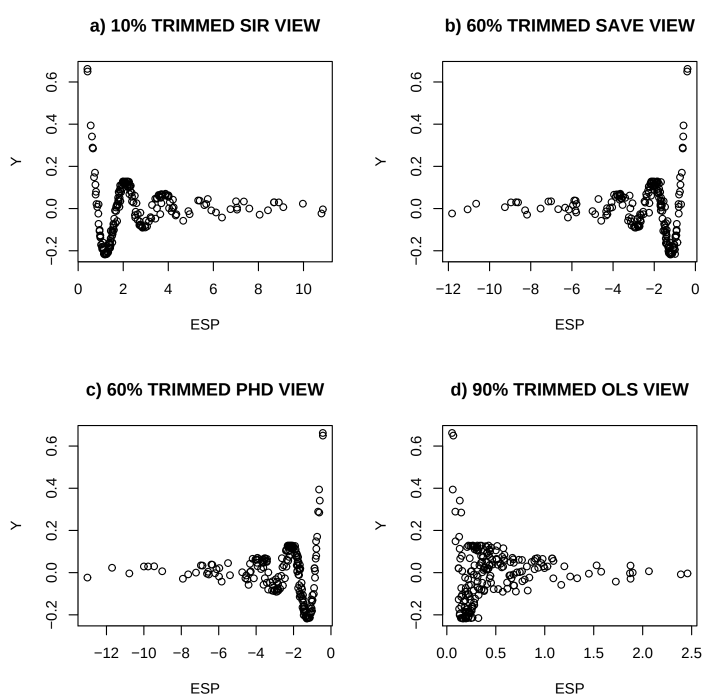

Figure 3: 1D Regression with Trimming

Table 5: Rejection Proportions for H_0: $\beta_2 = 0$

m	$\boldsymbol{x}$	Test	90	80	70	60	50	40	30	20	10	0	ADAP
1	1	F	.065	.073	.061	.056	.062	.051	.046	.050	.044	.043	.043
1	1	W	.004	.001	.007	.013	.015	.020	.027	.032	.045	.056	.056
5	1	F	.025	.017	.019	.023	.019	.019	.020	.022	.027	.037	.029
5	1	W	.006	.001	.002	.003	.006	.005	.010	.014	.025	.055	.026
2	2	F	.045	.033	.023	.024	.026	.070	.183	.182	.142	.166	.040
2	2	W	.010	.012	.007	.010	.021	.067	.177	.328	.452	.576	.050
4	3	F	.044	.032	.027	.058	.096	.081	.071	.057	.062	.123	.120
4	3	W	.024	.028	.028	.069	.152	.263	.337	.378	.465	.541	.539
6	4	F	.040	.023	.026	.024	.030	.032	.028	.044	.051	.088	.088
6	4	W	.009	.006	.012	.009	.013	.016	.030	.040	.076	.386	.319
7	5	F	.056	.053	.058	.058	.053	.054	.046	.044	.051	.037	.037
7	5	W	.003	.002	.001	.000	.005	.005	.034	.080	.118	.319	.250
3	6	F	.041	.030	.021	.024	.019	.025	.025	.034	.080	.374	.036
3	6	W	.004	.005	.003	.008	.007	.021	.019	.041	.084	.329	.264
6	7	F	.041	.032	.027	.032	.023	.041	.047	.053	.052	.055	.055
6	7	W	.009	.003	.007	.006	.013	.022	.019	.025	.054	.176	.169