Applications of Robust Distances for Regression

David J. Olive

Department of Mathematics

Southern Illinois University

Carbondale, IL 62901-4408, USA

(dolive@math.siu.edu)

July 21, 2003

Abstract

The DD plot, introduced by Rousseeuw and Van Driessen (1999), is a plot of classical vs robust Mahalanobis distances: MD_i vs RD_i. The DD plot can be used as a diagnostic for multivariate normality and elliptical symmetry, and to assess the success of numerical transformations towards elliptical symmetry. In the regression context, many procedures can be adversely affected if strong nonlinearities are present in the predictors. Even if strong nonlinearities are present, the robust distances can be used to help visualize important regression models such as generalized linear models.

KEY WORDS: Elliptically Contoured Distributions; GLM; Regression Graphics.

1 INTRODUCTION

Consider the multivariate model where the n iid observations $\boldsymbol{x}_i$ are $p \times 1$ vectors from a distribution with location/dispersion parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ is a $p \times 1$ vector and $\boldsymbol{\Sigma}$ is a $p \times p$ symmetric positive definite matrix. Let $\boldsymbol{X}$ be the $n \times p$ matrix with ith row $\boldsymbol{x}_i^T$, let $T(\boldsymbol{X})$ be a $p \times 1$ multivariate location estimator, and let the $p \times p$ symmetric positive definite matrix $C(\boldsymbol{X})$ be a covariance estimator. Then the ith squared Mahalanobis distance is the scalar

$$D_i^2 = D_i^2(T(\boldsymbol{X}), C(\boldsymbol{X})) = (\boldsymbol{x}_i - T(\boldsymbol{X}))^T C^{-1}(\boldsymbol{X})(\boldsymbol{x}_i - T(\boldsymbol{X})) \tag{1.1}$$

for each observation $\boldsymbol{x}_i$. The classical Mahalanobis distance uses the sample mean $\bar{\boldsymbol{x}}$ and sample covariance matrix S for (T, C) and will be denoted by MD_i. When $T(\boldsymbol{X})$ and $C(\boldsymbol{X})$ are alternative estimators, D_i will sometimes be denoted by RD_i (Rousseeuw and van Zomeren 1990).

The DD plot, introduced by Rousseeuw and Van Driessen (1999), is a plot of the MD_i vs the RD_i. Assume that $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and that the covariance matrix of $\boldsymbol{x}$ is $c_{\boldsymbol{x}}\boldsymbol{\Sigma}$ for some constant $c_{\boldsymbol{x}} > 0$. Then the classical estimator $(\bar{\boldsymbol{x}}, S)$ is a consistent estimator for $(\boldsymbol{\mu}, c_{\boldsymbol{x}}\boldsymbol{\Sigma})$. Section 2 shows that if the alternative estimator (T, C) is a consistent estimator for $(\boldsymbol{\mu}, a\boldsymbol{\Sigma})$ where $a > 0$ is some constant, then the plotted points will cluster tightly about the line through the origin with unit slope (identity line).

Regression is the study of the conditional distribution of y given predictors $\boldsymbol{x}$. An important class of regression models has the form

$$y = g(\boldsymbol{\beta}^T\boldsymbol{x}, e) \tag{1.2}$$

where g is the link function, $\boldsymbol{\beta}$ is a $p \times 1$ vector, and e is an error. Li and Duan (1989, p. 1014) note that this class of models includes generalized linear models (GLM), transformation models, dichotomous regression models, censored regression models, and projection pursuit models. Multiple linear regression and many nonlinear regression models are also included.

Dimension reduction attempts to reduce the dimension of the vector of predictors $\boldsymbol{x}$ without losing information about the conditional distribution of $y|\boldsymbol{x}$. The central subspace $S_{y|\boldsymbol{x}}(\eta)$ is the subspace spanned by the columns of η, where y is independent of $\boldsymbol{x}$ given $\eta^T\boldsymbol{x}$ and η is a $p \times d$ matrix with the smallest possible value of d. The central subspace is a super parameter that is used to characterize $y|\boldsymbol{x}$, and greater information reductions are attained with smaller values of d (Cook 1996, 1998b).

The assumption that the predictor distribution is elliptically contoured (symmetric) is often used in regression theory. Following Johnson (1987, p. 107-108), if $\boldsymbol{x}$ has density

$$f(\boldsymbol{x}) = k_p|\boldsymbol{\Sigma}|^{-1/2}g[(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})] \tag{1.3}$$

for some constant k_p and for some function g, then $\boldsymbol{x}$ has an elliptically contoured $EC_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ distribution. The characteristic function of $\boldsymbol{x} - \boldsymbol{\mu}$ is

$$\phi_{\boldsymbol{x}-\boldsymbol{\mu}}(\boldsymbol{t}) = \exp(i\boldsymbol{t}^T\boldsymbol{\mu})\psi(\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}) \tag{1.4}$$

for some function ψ. If the second moments exist, then

$$E(\boldsymbol{x}) = \boldsymbol{\mu} \text{ and } \text{Cov}(\boldsymbol{x}) = c_{\boldsymbol{x}}\boldsymbol{\Sigma} \tag{1.5}$$

where

$$c_{\boldsymbol{x}} = -2\psi'(0). \tag{1.6}$$

The population squared Mahalanobis distance

$$W \equiv D^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{1.7}$$

has the univariate density

$$h(w) = \frac{\pi^{p/2}}{\Gamma(p/2)} k_p w^{p/2-1} g(w). \tag{1.8}$$

A spherically symmetric distribution is an $EC_p(\mathbf{0}, \boldsymbol{I}, g)$ distribution, and for the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, $h(w)$ has the Chi-square χ^2_p density, $k_p = (2\pi)^{-p/2}$, and $g(a) = \exp(-a/2)$.

Under the assumption that the predictors $\boldsymbol{x}$ follow an EC distribution, inverse regression can be used to suggest response transformations (Cook 1998b, p. 21) and to identify semiparametric regression functions (Cook 1998b, pp. 56-57), as well as to determine the central subspace dimension d (Cook 1998b, pp. 144, 188, 191, and 197). The assumption is also used to show that sliced inverse regression (SIR), principal Hessian directions (pHd), and sliced average variance estimation (SAVE) provide information about the central subspace (Cook 1998b, pp. 204, 225, and 250 respectively) and to derive the asymptotic theory of associated statistics (Cook 1998b, pp. 211, 228, 230). See also Li (1991), Cook (1998a), Cook and Critchley (2000), and Cook and Lee (1999).

Cook (1993) and Cook and Croos-Dabrera (1998) show that partial residual plots perform best when the predictor distribution is EC. "Backfitting" uses partial residual plots for fitting models, with applications including projection pursuit regression, generalized additive models, additive spline models, and smoothing spline ANOVA. See Buja, Hastie, and Tibshirani (1989), Ansley and Kohn (1994), Luo (1998), and Wand (1999).

Many of these complex regression procedures seem to work well as long as there are no strong nonlinearities in the predictors. If the distribution of $\boldsymbol{x}$ is EC, then strong nonlinearities are not present since the conditional expectation $E(\boldsymbol{x}|\phi^T\boldsymbol{x})$ is linear for *all* conforming matrices ϕ (Eaton 1986, Cook 1998b, p. 130). Li and Duan (1989) and Cook (1998b) show that these procedures can provide useful results if a subset of the data can be selected such that the distribution of the predictors in the subset is closer to being EC.

Section 2 justifies the assertion that the DD plot will look like the identity line if the data follow a target EC distribution with finite second moments, and Section 3 examines possible estimators for computing the RD_i. Section 4 suggests how to make certain regression procedures resistant to nonlinearities in the predictors.

2 CONSTRUCTING THE DD PLOT

The following proposition shows that if consistent estimators are used to construct the distances, then the DD plot will tend to cluster tightly about the line segment through $(0, 0)$ and $(\mathrm{MD}_{n,\alpha}, \mathrm{RD}_{n,\alpha})$ where $0 < \alpha < 1$ and $\mathrm{MD}_{n,\alpha}$ is the α sample percentile of the MD_i. Let $K > 0$ be a large constant, e.g. the 99th percentile of the χ^2_p distribution.

Proposition 1. Assume that $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid observations from a distribution with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix. Let $a_j > 0$ and assume that $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$ are consistent estimators of $(\boldsymbol{\mu}, a_j\boldsymbol{\Sigma})$ for $j = 1, 2$. Let $D_{i,j}$ be the ith Mahalanobis distance computed from $(\hat{\boldsymbol{\mu}}_{j,n}, \hat{\boldsymbol{\Sigma}}_{j,n})$. Consider the cases in the region $R = \{i|0 \leq D_{i,j} \leq K,\ j = 1, 2\}$. Let r_n denote the correlation between $D_{i,1}$ and $D_{i,2}$ for

the cases in R (thus r_n is the correlation of the distances in lower left corner of the DD plot). Then $r_n \to 1$ as $n \to \infty$.

Proof. Let B_n denote the subset of the sample space on which both $\hat{\boldsymbol{\Sigma}}_{1,n}$ and $\hat{\boldsymbol{\Sigma}}_{2,n}$ have inverses. Then $P(B_n) \to 1$ as $n \to \infty$. The result follows since for fixed $\boldsymbol{x}$

$$D_j^2 \equiv (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) = \frac{1}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$$

$$+ \frac{2}{a_j} (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j) + \frac{1}{a_j} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_j) + \frac{1}{a_j} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T [a_j \hat{\boldsymbol{\Sigma}}_j^{-1} - \boldsymbol{\Sigma}^{-1}] (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \tag{2.1}$$

on B_n, and the last three terms converge to zero in probability. QED

To prove that the correlation tends to one for all of the distances requires more restrictions. Hardin and Rocke (1999) show that there exist distances computed from robust estimators that have an asymptotic χ_p^2 distribution if the underlying distribution of $\boldsymbol{x}$ is multivariate normal. Nevertheless, the variability in the DD plot may increase with the distances.

An algorithm estimator (T_A, C_A) (where the subscript "A" stands for "algorithm") of $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ can be constructed so that the DD plot follows the identity line. Let $\mathrm{RD}_i(A)$ denote the distances constructed using (T_A, C_A). By proposition 1, the plot of MD_i vs $\mathrm{RD}_i(A)$ will follow the line segment defined by the origin $(0, 0)$ and the point of medians, $(\mathrm{med}(\mathrm{MD}_i), \mathrm{med}(\mathrm{RD}_i(A)))$. This line segment has slope $\mathrm{med}(\mathrm{RD}_i(A))/\mathrm{med}(\mathrm{MD}_i)$ which is generally not one. Let $\mathrm{RD}_i = \tau \mathrm{RD}_i(A)$ denote the distances actually used in the DD plot where $\tau > 0$ is some constant; i.e., the estimator $(T_A, C_A/\tau^2)$ is used to construct the RD_i. Using the notation from Proposition 1, let $(a_1, a_2) = (a_M, a_A)$ (where "M" stands for "Mahalanobis"). The classical estimator is a consistent estimator of $(\boldsymbol{\mu}, a_M \boldsymbol{\Sigma})$

where $\boldsymbol{\mu} = E(\boldsymbol{x})$ and $\text{Cov}(\boldsymbol{x}) = a_M\boldsymbol{\Sigma}$, and the algorithm estimator (T_A, C_A) tends to be consistent for $(\boldsymbol{\mu}, a_A\boldsymbol{\Sigma})$ on the class of EC distributions and biased otherwise. The constant τ can be chosen so that the DD plot is *simultaneously* a diagnostic for elliptical symmetry and a diagnostic for the target distribution. That is, the plotted points follow the identity line if the data arise from a target EC distribution such as the multivariate normal distribution, but the points follow a line with non-unit slope if the data arise from an alternative EC distribution. In addition the DD plot can often detect departures from elliptical symmetry such as outliers, the presence of two groups, or the presence of a mixture distribution. These facts make the DD plot a useful alternative to other graphical diagnostics for target distributions. See Easton and McCulloch (1990), Li, Fang, and Zhu (1997), and Liu, Parelius, and Singh (1999) for references.

As an example, first assume that the target is the multivariate normal $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Then the $(\text{MD}_i)^2$ are asymptotically χ^2_p random variables. If $\boldsymbol{x} = \boldsymbol{\mu}$, then both the classical and the algorithm distances should be close to zero. Since the target distribution is Gaussian, let

$$\text{RD}_i = \frac{\sqrt{\chi^2_{p,0.5}}}{\text{med}(\text{RD}_i(A))}\text{RD}_i(A) \tag{2.2}$$

where $\chi^2_{p,0.5}$ is the median of the χ^2_p distribution.

Note that the DD plot can be tailored to any target elliptically contoured distribution that has 2nd moments. If it is known that $\text{med}(\text{MD}_i) \approx \text{MED}$ where MED is the target population analog (obtained, for example, via simulation, or from the actual target distribution as in equations (1.6) and (1.8)), then we use

$$\text{RD}_i = \frac{\text{MED}}{\text{med}(RD_i(A))}\text{RD}_i(A). \tag{2.3}$$

3 CHOICE OF THE ROBUST DISTANCES

The choice of the algorithm used to produce the estimator is important. Ideally, the (hyper) ellipsoids determined by the classical and alternative estimators should be approximately concentric if the data distribution is EC, but otherwise far from concentric for as large a class of non-EC-distributions as possible. Moreover, if the underlying distribution of the data is not EC, then the algorithm should try to select a subset of the data that is much more elliptically contoured than the data set as a whole. Let $(T_R, C_R) = (T_A, C_A/\tau^2)$ denote the scaled estimators used to construct the DD plot. In this plot, the points below the hth ordered distance $RD_{(h)}$ correspond to cases that are in the ellipsoid

$$\{\boldsymbol{x} : (\boldsymbol{x} - T_R(\boldsymbol{X}))^T C_R^{-1} (\boldsymbol{x} - T_R(\boldsymbol{X})) < RD_{(h)}^2\} \tag{3.1}$$

while points to the left of $MD_{(h)}$ are in an ellipsoid determined by the classical estimators.

Two robust estimators of location/dispersion that have the desired properties are the minimum covariance determinant (MCD(c)) estimator and the minimum volume ellipsoid (MVE(c)) estimator. The MCD finds the subset of $c \approx n/2$ observations whose classical covariance matrix has the lowest determinant. Then (T_{MCD}, C_{MCD}) is the classical sample mean and covariance matrix of these c observations. The MCD estimator produces an ellipsoid that covers c cases and has small volume (recall that the volume of the ellipsoid given by equation (3.1) is proportional to the determinant of the covariance matrix C_R), but the MVE finds the ellipsoid with the smallest volume that covers the c cases. See Rousseeuw and Leroy (1987, pp. 262-263) and Rousseeuw (1984).

Computing these estimators is very expensive, so approximations based on iterative

algorithms are used. The basic idea begins with the classical estimator computed from an initial randomly selected subset of $p+1$ points called a "start," from which the Mahalanobis distances $\{D_i : i = 1, ..., n\}$ are computed for all n points. At the next iteration, the classical estimator is computed on the c cases corresponding to the smallest distances. This iteration continues until convergence. We call the final subset of c cases the "attractor" of the start. We use K starts and compute K estimators from the resulting attractors. The algorithm estimator is the one that minimizes the MCD criteria. The FMCD algorithm of Rousseeuw and Van Driessen (1999) is available from the web site (http://win-www.uia.ac.be/u/statis/) while Hawkins and Olive (1999) describe a similar algorithm that is available from (http://www.stat.umn.edu).

The DD plot will tend to be very linear if the algorithm produces consistent estimators, but the algorithms described above are inconsistent if K is fixed (Lopuhaä 1999). We use the estimator (T_{FMCD}, C_{FMCD}) to compute the $\mathrm{RD}_i(A)$ provided that $K \geq \max(500, n/100)$ starts are used. (The default for the *Splus* function `cov.mcd` is $K = 500$ starts.) This estimator is seeking the most concentrated ellipsoid that contains $c \approx n/2$ cases. If the data distribution is not EC, then the distribution of the c cases is probably much closer to being EC. The DD plot will follow the identity line closely only if $\mathrm{med}(\mathrm{MD}_i) \approx \mathrm{MED}$, and $\mathrm{RD}_i^2 =$

$$(\boldsymbol{x}_i - T_{FMCD})^T [(\frac{\mathrm{MED}}{\mathrm{med}(\mathrm{RD}_{A,i})})^2 C_{FMCD}^{-1}](\boldsymbol{x}_i - T_{FMCD}) \approx (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T S^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) = \mathrm{MD}_i^2$$

for $i = 1, ..., n$. When the distribution is not EC, (T_{FMCD}, C_{FMCD}) and $(\bar{\boldsymbol{x}}, S)$ will often produce ellipsoids that are far from concentric.

This choice is certainly not perfect. There exist data sets with outliers or two groups

such that both the classical and robust estimators produce ellipsoids that are nearly concentric. We suspect that the situation worsens as p increases. In a simulation study, $N_p(\mathbf{0}, \boldsymbol{I}_p)$ data were generated and the affine equivariant estimator `cov.mcd` was used to compute first the $\mathrm{RD}_i(A)$, and then the RD_i using equation (2.2). The results are shown in Table 1. Each choice of n and p used 100 runs, and the 100 correlations between the RD_i and the MD_i were computed. The mean and minimum of these correlations are reported along with the percentage of correlations that were less than 0.95 and 0.80. The simulation shows that small data sets (of roughly size $n < 8p+20$) yield plot distances that may not cluster tightly about the identity line even if the data distribution is Gaussian.

Figure 1 shows the DD plots for 3 artificial data sets. The DD plot for 200 $N_3(\mathbf{0}, \boldsymbol{I}_3)$ points shown in Figure 1a resembles the identity line. The DD plot for 200 points from the elliptically contoured distribution $0.6N_3(\mathbf{0}, \boldsymbol{I}_3) + 0.4N_3(\mathbf{0}, 25\ \boldsymbol{I}_3)$ in Figure 1b clusters about a line through the origin with a slope close to 2.0.

A weighted DD plot uses only the cases with $\mathrm{RD}_i < \sqrt{\chi^2_{p,.975}}$. This emphasis on the lower left corner of the DD plot can magnify features that are obscured when large RD_i's are present. If the distribution of $\boldsymbol{x}$ is EC, proposition 1 implies that the correlation of the points in the weighted DD plot will tend to one and that the points will cluster about a line passing through the origin. For example, the plotted points in the weighted DD plot (not shown) for the non-Gaussian EC data of Figure 1b are highly correlated and still follow a line through the origin with a slope close to 2.0.

Figures 1c and 1d illustrate how to use the weighted DD plot. The ith case in Figure 1c is $(\exp(x_{i,1}), \exp(x_{i,2}), \exp(x_{i,3}))^T$ where $\boldsymbol{x}_i$ is the ith case in Figure 1a; i.e.,

the marginals follow a lognormal distribution. The plot does not resemble the identity line, correctly suggesting that the distribution of the data is not Gaussian; however, the correlation of the plotted points is rather high. Figure 1d is the weighted DD plot where cases with $\mathrm{RD}_i \geq \sqrt{\chi^2_{3,.975}} \approx 3.06$ are given zero weight. Notice that the correlation of the plotted points is not close to one and that the best fitting line in Figure 1d may not pass through the origin. These results suggest that the distribution of $\boldsymbol{x}$ is not EC.

It is easier to use the DD plot as a diagnostic for a target distribution than as a diagnostic for elliptical symmetry. If the data arise from the target distribution, then the DD plot will tend to be a useful diagnostic when the sample size n is such that the sample correlation coefficient in the DD plot is at least 0.80 with high probability. As a diagnostic for elliptical symmetry, it may be useful to add the OLS line to the DD plot and weighted DD plot as a visual aid, along with numerical quantities such as the OLS slope and the correlation of the plotted points.

4 APPLICATIONS

The DD plot can be used to diagnose elliptical symmetry, to detect outliers, and to assess the success of numerical methods for transforming data towards an elliptically contoured distribution. Since many statistical methods assume that the underlying data distribution is Gaussian or EC, there is an enormous literature on numerical tests for elliptical symmetry. Bogdan (1999) and Czörgö (1986) provide references for tests for multivariate normality while Koltchinskii and Li (1998) have references for tests for elliptically contoured distributions. The DD plot can be used simultaneously as a diagnostic for

whether the data arise from a Gaussian distribution or from another EC distribution. EC data will cluster about a straight line; Gaussian data in particular will cluster about the identity line.

Numerical methods for transforming data towards a target EC distribution have been developed. Generalizations of the Box-Cox transformation towards a multivariate normal distribution are described in Velilla (1993). Alternatively, Cook and Nachtsheim (1994) offer a two-step numerical procedure for transforming data towards a target EC distribution. The first step simply gives zero weight to a fixed percentage of cases that have the largest robust Mahalanobis distances, and the second step uses Monte Carlo case reweighting with Voronoi weights.

Example. Buxton (1920, pp. 232-5) gives 20 measurements of 88 men. We will examine whether the multivariate normal distribution is a plausible model for the measurements *head length, nasal height, bigonal breadth,* and *cephalic index* where one case has been deleted due to missing values. This data set can be downloaded from the web site (http://www.stat.umn.edu/hawkins). Figure 2a shows the DD plot. Five head lengths were recorded to be around 5 feet and are massive outliers. Figure 2b is the DD plot computed after deleting these points and suggests that the normal distribution is plausible.

The DD plot complements rather than replaces the numerical procedures. For example, if the goal of the transformation is to achieve a multivariate normal distribution and if the data points cluster tightly about the identity line, as in Figure 1a, then perhaps no transformation is needed. For the data in Figure 1c, a good numerical procedure should

suggest coordinatewise log transforms. Following this transformation, the resulting plot shown in Figure 1a indicates that the transformation to normality was successful.

Robust distances can also be used to estimate h and $c\boldsymbol{\beta}$ in models of the form

$$y = h(\boldsymbol{\beta}^T\boldsymbol{x}) + e = m_{a,c}(a + c\boldsymbol{\beta}^T\boldsymbol{x}) + e \tag{4.1}$$

where both h and $\boldsymbol{\beta}$ may be unknown and

$$m_{a,c}(u) = h(\frac{u-a}{c})$$

for some constants a and $c \neq 0$. Notice that if the signal to noise ratio is high, the plot of $a + c\boldsymbol{\beta}^T\boldsymbol{x}$ vs y will suggest a functional form for h.

Let the OLS estimator $(\hat{\alpha}, \hat{\boldsymbol{\beta}}^T)^T$ be computed from the regression of y on the predictors $\boldsymbol{x}$ plus a constant. Li and Duan (1989) and Aldrin, Bϕlviken, and Schweder (1993) show that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $k\boldsymbol{\beta}$ when the predictor distribution is EC. Without loss of generality, assume that $\boldsymbol{\beta}^T\Sigma_{\boldsymbol{x}}\boldsymbol{\beta} = 1$. Then $\hat{\boldsymbol{\beta}}$ estimates the population parameter

$$\boldsymbol{\beta}_{OLS} = \Sigma_{\boldsymbol{x}}^{-1}\Sigma_{\boldsymbol{x},y} = k(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{B}(\boldsymbol{x})$$

where $\Sigma_{\boldsymbol{x}} = \mathrm{Cov}(\boldsymbol{x})$, $\Sigma_{\boldsymbol{x},y} = \mathrm{Cov}(\boldsymbol{x}, y)$, $k(\boldsymbol{x}) = E[\boldsymbol{\beta}^T(\boldsymbol{x} - E(\boldsymbol{x}))h(\boldsymbol{\beta}^T\boldsymbol{x})]$, and $\boldsymbol{B}(\boldsymbol{x})$ is the bias vector defined by $\boldsymbol{B}(\boldsymbol{x}) = \Sigma_{\boldsymbol{x}}^{-1}E[h(\boldsymbol{\beta}^T\boldsymbol{x})\boldsymbol{u}]$ where $\boldsymbol{u} = \boldsymbol{x} - E(\boldsymbol{x}) - (\boldsymbol{\Sigma_x\beta})\boldsymbol{\beta}^T(\boldsymbol{x} - E(\boldsymbol{x}))$. If the predictor distribution is EC then $\boldsymbol{B} = \boldsymbol{0}$ and the bias can also be small if no strong nonlinearities are present in the predictors. Hence $\hat{\boldsymbol{\beta}}$ estimates $k\boldsymbol{\beta}$, and h can be visualized with a graph of $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ vs y if the predictor distribution is EC. With two predictors and a non-EC distribution, Cook and Weisberg (1999, ch. 8) demonstrate that h can be visualized using a three-dimensional plot with y on the vertical axis and the two predictors on the horizontal and out of page axes. When we rotate the plot about the

vertical axis, each combination of the predictors gives a two dimensional "view." We then search for the view that has a smooth mean function and the smallest possible variance function.

For higher dimensions, the bias $\boldsymbol{B}$ can often be made small by trimming $K\%$ of the cases with the largest robust distances (recall Winsor's principle: "all data are roughly Gaussian in the middle," see Hoaglin, Mosteller and Tukey 1983, p. 363), and then computing the OLS estimator $\hat{\boldsymbol{\beta}}_K$ from the retained cases. Use $K = 0$, 10, 20, 30, 40, 50, 60, 70, 80, and 90 to generate ten plots of $\hat{\boldsymbol{\beta}}_K^T \boldsymbol{x}$ vs y using all n cases. In analogy with the Cook and Weisberg procedure for visualizing h with two predictors, the plot with a smooth mean function and the smallest variance function will be called the "best trimmed view."

As an example, suppose that the predictors are the lognormal data from Figure 1c and that $y = (x_1 + 2x_2 + 3x_3)^3 + e$ where e is a $N(0, 1)$ random variable; i.e., nonlinearities are present in the predictors and $\boldsymbol{\beta} = (1, 2, 3)^T$. Figure 3a shows the plot of $\boldsymbol{\beta}^T \boldsymbol{x}$ vs y, called the "true view." The OLS estimate $\hat{\boldsymbol{\beta}} = (641.427, 2977.751, 2864.351)^T$, and the corresponding vector of OLS standard errors is $(167.38, 138.75, 189.02)^T$. Figure 3b shows that the OLS view has considerable bias. The 70% trim gives the "best trimmed view" and $\hat{\boldsymbol{\beta}}_K = (94.715, 203.507, 301.730)^T \approx 100\boldsymbol{\beta}$. The best trimmed view, shown in Figure 3c (where the fitted values $\hat{\boldsymbol{\beta}}_K^T \boldsymbol{x}$ are denoted by the label "BESTFIT"), is almost the same as the true view. Trimming was effective in reducing the bias for the estimation of $c\boldsymbol{\beta}$. If the same function were generated with the Gaussian data of Figure 1a, then $\hat{\boldsymbol{\beta}} = (41.548, 87.465, 120.671)^T \approx 42\boldsymbol{\beta}$, the corresponding vector of OLS standard

errors is $(7.21, 7.33, 7.34)^T$, and $\hat{\boldsymbol{\beta}}_K = (12.386, 25.095, 37.414)^T \approx 12.5\boldsymbol{\beta}$. In this case 50% trimming gives the best trimmed view but all ten views were close to the true view.

The trimmed view has many applications. The response y can be predicted for a given $\boldsymbol{x}$ by finding the value of $\hat{\boldsymbol{\beta}}_K^T\boldsymbol{x}$ on the horizontal axis and the corresponding y value on the vertical axis. The trimmed view can also be used as a graphical diagnostic for linearity or monotonicity of h. See Heckman and Zamar (2000) for the importance of detecting monotonicity in regression. The plot can also suggest parametric forms for h and starting values for nonlinear regression. If it is assumed that $y = t^{-1}(\boldsymbol{\beta}^T\boldsymbol{x} + e)$ where t^{-1} is monotone, then the inverse response plot of y vs $\hat{\boldsymbol{\beta}}_T^T\boldsymbol{x}$ will suggest a functional form for t. Hence the trimmed view can make the Cook and Weisberg (1994) procedure for response transformations resistant to nonlinearities in the predictors.

It should be noted that the OLS view and best trimmed view can fail if h is symmetric about $E(\boldsymbol{\beta}_{OLS}^T\boldsymbol{x})$ so that $\mathrm{Cov}(\boldsymbol{x}, y) = 0$. A useful view for visualizing h can sometimes be found if a subset of the predictors can be extracted for which the correlation between the OLS fitted values and the response y is nonzero. As an example, let $y = (x_1 + 2x_2 + 3x_3)^2 + e$ where e is $N(0, 1)$ and the predictors are the same as those used to construct Figure 1a. Figure 4a shows the true view while Figure 4b shows the OLS view. The OLS estimate suggests that the order of importance of the predictors is reversed from the true order of importance since $\hat{\boldsymbol{\beta}} = (2.329, 2.275, 0.990)^T$. The correlation between the response y and the OLS fitted values $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ is nearly zero, but the correlation between y and $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ is positive if the cases that have $\hat{\boldsymbol{\beta}}^T\boldsymbol{x} < \mathrm{med}(\hat{\boldsymbol{\beta}}^T\boldsymbol{x}) \approx 0$ are given zero weight. With this weighting, ten trimmed views were generated, the best of which trimmed 60%

of the retained cases yielding $\hat{\boldsymbol{\beta}}_K = (2.648, 6.075, 8.085)^T$ (see Figure 4c).

5 Discussion

The proposed applications of the robust distances RD_i are simple to construct and interpret. Programs for the DD plots and trimmed views can be written in a few lines using *Splus* (MathSoft 1999) or *R* (http://www.r-project.org/). If the data distribution is multivariate normal, then the points in the DD plot will cluster tightly about the identity line; if the distribution is non-Gaussian but EC, the points will still cluster tightly about a line but with non-unit slope.

The ten trimmed views for a smooth mean function and a small variance function are especially informative if the signal to noise ratio is high. Even with high noise levels, views similar to the true view can be obtained.

The construction of the views is not limited to OLS. Li and Duan (1989) show that maximum likelihood type estimators such as those used to estimate GLM's will also produce consistent estimators of $c\boldsymbol{\beta}$ in models of the form $y = g(\boldsymbol{\beta}^T\boldsymbol{x}, e)$.

Making other regression methods such as SAVE resistant to the presence of strong nonlinearities in the predictors will require further research. If the points in the DD plot do not cluster about a line, then nonlinearity may be present. Marginal or multivariate transformations (e.g. Box-Cox) can be very effective for eliminating gross nonlinearities as can the transformation of Cook and Nachtsheim (1994). Using robust distances to select a subset of data can also be effective, but specific recommendations will depend on the regression procedure.

ACKNOWLEDGMENTS

The author thanks R. Dennis Cook, Douglas M. Hawkins, Editor Karen Kafadar, the Associate Editor, and a referee for many comments that led to improvements in this article. The author thanks Peter J. Rousseeuw and Katrien Van Driessen for making the preprint of their 1999 paper available on the web.

6 References

Aldrin, M., Bϕlviken, E., and Schweder, T. (1993), "Projection Pursuit Regression for Moderate Non-linearities," *Computational Statistics and Data Analysis,* 16, 379-403.

Ansley, C.F., and Kohn, R. (1994), "Convergence of the Backfitting Algorithm for Additive Models," *Journal of the Australian Mathematical Society, Series A,* 57, 316-329.

Bogdan, M. (1999), "Data Driven Smooth Tests for Bivariate Normality," *Journal of Multivariate Analysis,* 68, 26-53.

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," *The Annals of Statistics,* 17, 453-555.

Buxton, L.H.D. (1920), "The Anthropology of Cyprus," *The Journal of the Royal Anthropological Institute of Great Britain and Ireland,* 50, 183-235.

Cook, R.D. (1993), "Exploring Partial Residual Plots," *Technometrics,* 35, 351-362.

Cook, R.D. (1996), "Graphics for Regressions with Binary Response," *Journal of the American Statistical Association,* 91, 983-992.

Cook, R.D. (1998a), "Principal Hessian Directions Revisited," *Journal of the American Statistical Association,* 93, 84-100.

Cook, R.D. (1998b), *Regression Graphics: Ideas for Studying Regression Through Graphics,* John Wiley and Sons, Inc., NY.

Cook, R.D., and Critchley, F. (2000), "Identifying Outliers and Regression Mixtures Graphically," *Journal of the American Statistical Association,* 95, 781-794.

Cook, R.D., and Croos-Dabrera, R. (1998), "Partial Residual Plots in Generalized Linear Models," *Journal of the American Statistical Association,* 93, 730-739.

Cook, R.D., and Lee, H. (1999), "Dimension Reduction in Binary Response Regression," *Journal of the American Statistical Association,* 94, 1187-1200.

Cook, R.D., and Nachtsheim, C.J. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *Journal of the American Statistical Association,* 89, 592-599.

Cook, R.D., and Weisberg, S. (1994), "Transforming a Response Variable for Linearity," *Biometrika,* 81, 731-737.

Cook, R.D., and Weisberg, S. (1999), *Applied Regression Including Computing and Graphics,* John Wiley and Sons, Inc., NY.

Czörgö, S. (1986), "Testing for Normality in Arbitrary Dimension," *The Annals of Statistics,* 14, 708-723.

Easton, G.S., and McCulloch, R.E. (1990), "A Multivariate Generalization of Quantile Quantile Plots," *Journal of the American Statistical Association,* 85, 376-386.

Eaton, M.L. (1986), "A Characterization of Spherical Distributions," *Journal of Multi-*

variate Analysis, 20, 272-276.

Hardin, J., and Rocke, D.M. (1999), "The Distribution of Robust Distances," Preprint.

Hawkins, D.M., and Olive, D.J. (1999), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics and Data Analysis,* 30, 1-11.

Heckman, N.E., and Zamar, N.H. (2000), "Comparing the Shapes of Regression Functions," *Biometrika,* 87, 135-144.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1983), *Understanding Robust and Exploratory Data Analysis,* John Wiley and Sons, Inc., NY.

Johnson, M.E. (1987), *Multivariate Statistical Simulation,* John Wiley and Sons, Inc., NY.

Koltchinskii, V.I., and Li, L. (1998), "Testing for Spherical Symmetry of a Multivariate Distribution," *Journal of Multivariate Analysis,* 65, 228-244.

Li, K.C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association,* 86, 316-342.

Li, K.C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics,* 17, 1009-1052.

Li, R., Fang, K., and Zhu, L. (1997), "Some Q-Q Probability Plots to Test Spherical and Elliptical Symmetry," *Journal of Computational and Graphical Statistics,* 6, 435-450.

Liu, R.Y., Parelius, J.M., and Singh, K. (1999), "Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics, and Inference," *The Annals of Statistics,* 27, 783-858.

Lopuhaä, H.P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics,* 27, 1638-1665.

Luo, Z. (1998), "Backfitting in Smoothing Spline Anova," *The Annals of Statistics,* 26, 1733-1759.

MathSoft (1999), *S-Plus 2000 User's Guide,* Data Analysis Products Division, MathSoft, Seattle, WA.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association,* 79, 871-880.

Rousseeuw, P.J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection,* John Wiley and Sons, Inc., NY.

Rousseeuw, P.J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics,* 41, 212-223.

Rousseeuw, P.J., and van Zomeren, B.C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association,* 85, 633-651.

Velilla, S. (1993), "A Note on the Multivariate Box-Cox Transformation to Normality," *Statistics and Probability Letters,* 17, 259-263.

Wand, M.P. (1999), "A Central Limit Theorem for Local Polynomial Backfitting Estimators," *Journal of Multivariate Analysis,* 70, 57-65.

Table 1: Corr(RD_i, MD_i) for $N_p(\mathbf{0}, \boldsymbol{I}_p)$ Data, 100 Runs

p	n	mean	min	% < 0.95	% < 0.8
3	44	0.866	0.541	81	20
3	100	0.967	0.908	24	0
7	76	0.843	0.622	97	26
10	100	0.866	0.481	98	12
15	140	0.874	0.675	100	6
15	200	0.945	0.870	41	0
20	180	0.889	0.777	100	2
20	1000	0.998	0.996	0	0
50	420	0.894	0.846	100	0

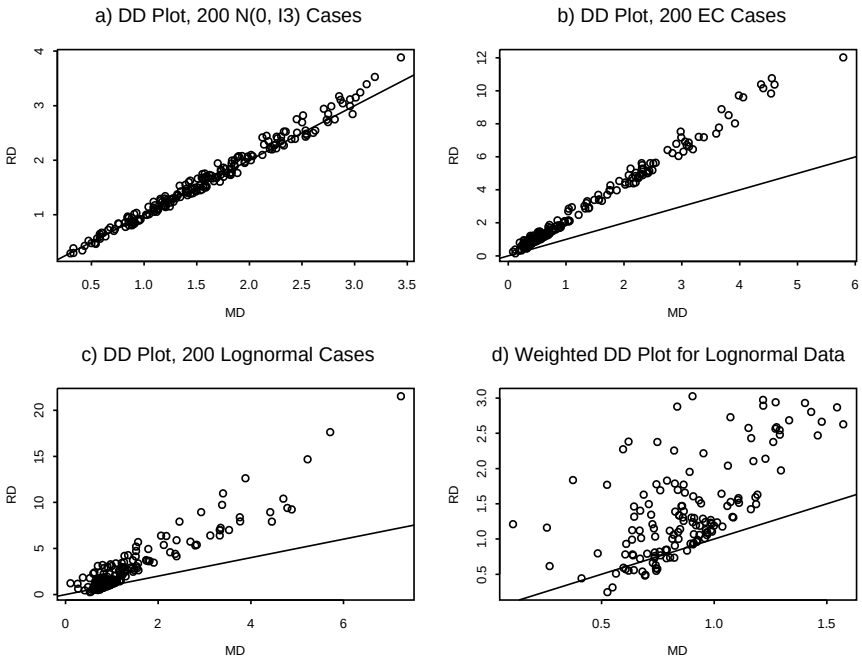

Figure 1: 4 DD Plots

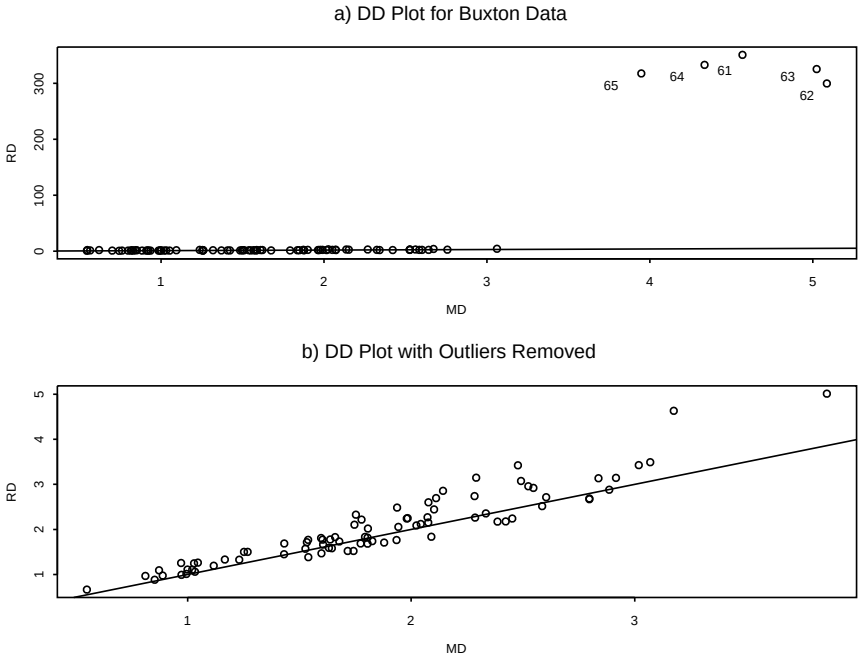

Figure 2: DD Plots for the Buxton Data

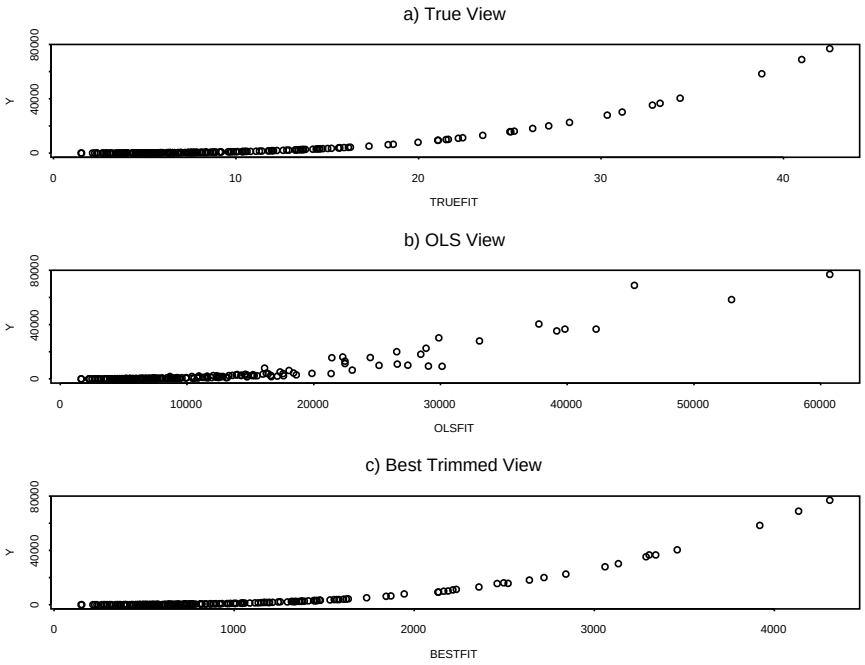

Figure 3: The TRUEFIT is $\boldsymbol{\beta}^T\boldsymbol{x}$ and $h(u) = u^3$.

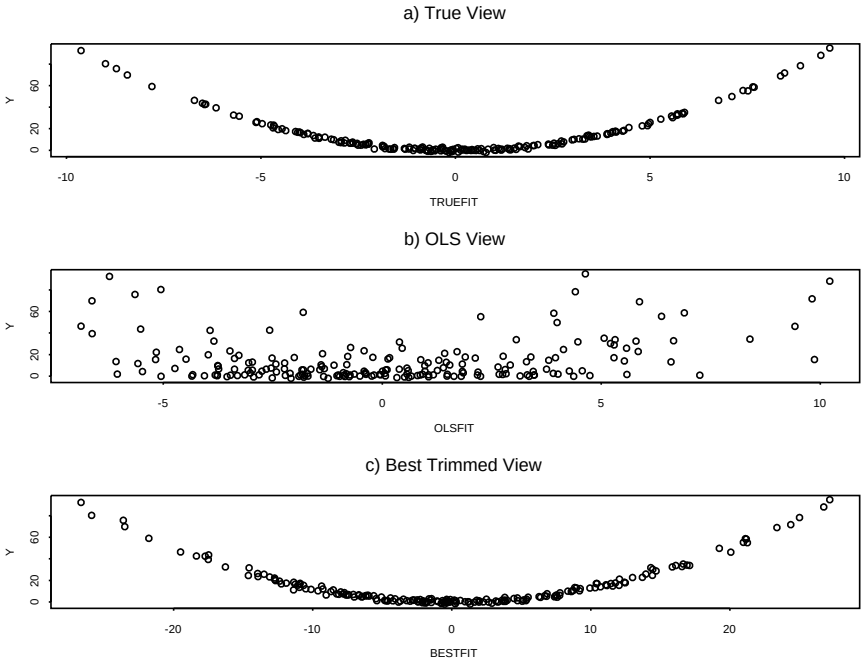

Figure 4: Views for Estimating $h(u) = u^2$