Statistica Sinica **14**(2004), 431-448

LINEAR FITTING BY SIMPLICIAL INTERCEPT DEPTH (SID): REFLECTION INVARIANCE AND ROBUSTNESS

Regina Y. Liu, Kesar Singh and Julie H. Teng

Rutgers University

Abstract: This paper introduces the method of *simplicial intercept depth* (SID) for linear fitting. The SID method is invariant under rotations and reflections. It is also robust against outliers, with its breakdown point bounded away from zero. The paper also presents some comparisons, in terms of robustness, efficiency and invariance, between the SID method and other regression methods such as regression depth, orthogonal regression, L_1 regression and least squares. Finally, the paper introduces the *simplicial fit plot* as a new graphical tool for a visual assessment of the goodness of a given linear fit. The area under the simplicial fit plot of the proposed linear fit corresponds to its SID value. Hence the SID value provides a goodness-of-fit measure for any given linear fit, and can be viewed as a robust analog of the usual *coefficient of determination* arising from the least squares method.

Key words and phrases: Breakdown, linear fit, reflection invariance, rotation invariance, simplicial intercept depth, simplicial linear fit.

1. Introduction

Linear fitting is one of the most commonly used statistical tools for studying related variables. The goal of this paper is to introduce the *simplicial intercept depth* (SID) for linear fitting, and the *simplicial fit* (*S-fit*) *plot* for visually assessing the goodness-of-fit of a given linear fit. The SID is presented as a new notion of depth for lines (or hyperplanes in general) as well as a new robust linear fitting method which is invariant under *rotations* and *reflections*. These desirable invariance properties are not generally shared by other regression methods. The reflection invariance property implies treating the input variable and the output variable symmetrically, and it is particularly desirable in linear fitting in case the choice of the input variable, between the variables under consideration, is not evident. Simple examples are the crime rate and the median income of a region, and the test scores in physics and mathematics of a student. Reflection invariance keeps the linear relationship between the two variables unaffected when their roles as input and output variables are reversed; its desirability can be viewed as one of the distinguishing features between linear fitting and regression.

This point will be illustrated further in Section 3.1 which discusses the modeling of verbal and mathematics test scores in the Scholastic Assessment Test (SAT). The invariance of SID under orthonormal transformations (which include rotations and reflections) is discussed in detail in Section 2.1. Comparisons of the SID with the orthogonal regression and its L_1 version, which are two obvious reflection invariant methods, are presented throughout the paper.

The SID may be viewed as a generalization of the simplicial depth introduced in Liu (1990), and can be described in the basic two-dimensional setting as follows. In $\mathbb{R}^2$, the depth of a line is simply the averaged length-ratio of the line intercept within a triangle to its longest side, where the average runs over all the triangles generated from the data. In essence, the depth here measures *how deep a line cuts through all the triangles formed by the data.* The SID value is 1 if all data points fall on a line, but it decreases as data points become more scattered. The line with the highest SID value is considered the best fit. Note that the SID value of this fitted line can actually serve as a robust analog of the coefficient of determination defined in the least squares approach, since it provides a measure of goodness-of-fit of the resulting fitted line. This measure is further illustrated graphically as the area under the *simplicial fit plot* which is to be discussed in Section 4.

Among existing linear fitting methods, the *least squares method* has been used most extensively, for mathematical convenience as well as for its certain optimality properties under normally distributed errors. However, it is less satisfactory when error distributions are heavy-tailed or when outliers are present. Many robust alternatives have been proposed for linear fitting of data and for identifying possible outliers. For example, L_1-regression is known to be robust with respect to outliers along the direction of the output variable, but not with respect to outliers along the direction of the input variable (also referred to as *leverage points*). Many other robust regression methods exist in the literature. Examples include the median-type estimators in : the pairwise slopes method in Theil (1950) and Sen (1968), the least median of squares regression in Rousseeuw (1984), the remedian in Rousseeuw and Bassett (1990), the resistant line method in Tukey (1970) and Johnstone and Velleman (1985); the M-method in Maronna and Yohai (2000); the R-estimator method in : Jureckovà (1971), Hettmansperger and McKean (1977) and Hossjer (1994); and the L-estimator method in : Bickel (1973), Koenker and Bassett (1978), Ruppert and Carroll (1980), Carroll and Ruppert (1985), Simpson, Ruppert and Carrol (1992) and Stromberg, Hossjer and Hawkins (2000). Recently, there has been significant progress made by Rousseeuw and Hubert (1999) in developing the concept of depth for regression. They introduced the so-called *regression depth* and the corresponding regression depth method. The deepest fit is shown to be robust against outliers, with a high breakdown point. Unlike the usual concept of depth defined with respect to a multivariate data cloud, the regression depth requires a new geometric perspective, since it measures the depth of hyperplanes rather than that of finite-dimensional vectors. All the above regression methods offer some robust and computationally feasible solutions, and they have all generated much follow-up.

In Section 2, we describe in detail the SID method, and study its invariance and breakdown properties. We present in Section 3 some applications of the SID method to both real and simulated datasets. The results show clearly that the SID method is highly robust against outliers. (See Figure 4 for a visual display of breakdown properties in linear fitting.) Moreover, we also present some comparisons of the SID method in the aspects of robustness, efficiency, and invariance to the methods of least squares, regression depth, L_1 regression, and orthogonal regression and its L_1 version. In Section 4, the *simplicial fit plot* is proposed as a new graphical tool for visually assessing the goodness of a given linear fit. This assessment can be further summed up by the area under the simplicial fit plot which is exactly the SID value achieved by the given linear fit. Section 5 contains some concluding remarks and open problems.

2. The Simplicial Intercept Depth (SID) Method for Linear Fitting

Although the SID method applies to data of any dimension, for simplicity we focus here mainly on the following two-dimensional linear model:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad \text{for} \qquad \mathrm{i} = 1, \ldots, \mathrm{n}, \tag{2.1}$$

where n is the sample size, x_i is the i-th input variable or covariate, and y_i is the i-th response or output variable. The e_i's are independent error terms and are usually assumed to have zero mean and unknown variance σ^2. The parameters β_0 and β_1 are the unknown intercept and slope which are to be estimated from the given dataset $\mathcal{W}_n = \{W_1, \ldots, W_n\}$, where $W_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$.

Let $\Delta(W_i, W_j, W_k)$ denote the triangle whose vertices are the data points W_i, W_j and W_k. Let $\boldsymbol{\Delta}$ denote the collection of all these triangles, $\boldsymbol{\Delta} \equiv \{\Delta_1, \ldots \Delta_K\}$, where $K \equiv \binom{n}{3}$.

Let $y = b_0(i,j) + b_1(i,j)x$ denote the line which passes through the pair of sample points (x_i, y_i) and (x_j, y_j). For simplicity, we denote $y = b_0(i,j)+b_1(i,j)x$ by L_{ij}, and denote the pool of all $\binom{n}{2}$ lines by $\mathcal{P} \equiv \{L_{ij} : i, j = 1, \ldots, n, i < j\}$. Given a line L_{ij} and a triangle Δ_k, we can check if they intersect. If the line L_{ij} passes through Δ_k, then there are two intersection points. Let $\ell_k(i,j)$ denote the distance between these two intersection points. We refer to $\ell_k(i,j)$ as the "intercept of L_{ij} within the triangle Δ_k". Clearly $\ell_k(i,j) = 0$ if the line L_{ij} does not intersect with the triangle Δ_k, or has only one intersection point.

Definition 2.1. For a given line L_{ij} in $\mathcal{P}$, the *simplicial intercept depth* (denoted by SID) is

$$\mathrm{SID}(L_{ij}) = \frac{1}{\binom{n}{3}} \sum_{k=1}^{\binom{n}{3}} \left\{ \frac{\ell_k(i,j)}{m(\Delta_k)} \right\}, \tag{2.2}$$

where $\Delta_k \in \mathbf{\Delta}$, $\{i,\ j = 1, \ldots, n,\ i < j\}$, and $m(\Delta_k)$ is the length of the longest side of Δ_k (it may be realized by more than one side).

In other words, the depth of the line L_{ij} is the average of the ratios of its intercept within each triangle to the length of the longest side of that triangle, over all triangles in $\mathbf{\Delta}$. A larger SID value for L_{ij} implies that the line L_{ij} cuts "deeper" into more triangles in $\mathbf{\Delta}$. Thus it gives a greater number of relatively "*longer*" intercepts.

In degenerate cases, i.e., when two or all three vertices of the triangle coincide, we can define the ratio $\ell_k(i,j)/m(\Delta_k)$, separately, as in the following two cases:

Case 1. Two vertices of the triangle are identical: In this case, the triangle becomes a line. The ratio is defined to be 1 if the line L_{ij} coincides with the line representing the triangle, and 0 otherwise.

Case 2. All three vertices of the triangle are identical: the ratio is defined to be 1 if L_{ij} passes through this common vertex, and 0 otherwise.

The value $\mathrm{SID}(L_{ij})$ defined in (2.2) provides a natural measure of goodness-of-fit of the line L_{ij} with respect to the given data set. Based on the SID, the best linear fit for dataset $\mathcal{W}_n$ is the line in $\mathcal{P}$ with maximum SID value. We denote the best fit by $y = b_0^* + b_1^* x$.

The SID value of the fitted line is 1 if and only if all the data points are on the line, and it decreases as the data points scatter away from the line. Therefore, the value in the SID method plays a role similar to that of the coefficient of determination in the least squares method. In other words, the SID value can be used as an alternative measure of goodness-of-fit of a linear fit for a given dataset. This point is discussed further, highlighted by graphs, at the end of Section 4.

In theory, the pool of lines under consideration in Definition 2.1 can be the set of all lines in the plane. However, it is impractical, if not impossible, to search for the maximum SID among all possible lines. Therefore, we restrict ourselves only to the lines contained in $\mathcal{P}$. Clearly, this restriction may lead to some loss of efficiency. One way to make up for this loss of efficiency is to increase the size of the pool $\mathcal{P}$ by including more viable line candidates. For example, consider also the averages of each pair of the original observations, and include in $\mathcal{P}$ the additional lines which pass through the averages or the combination of an average and an original observation. The increase of efficiency, in this case, may lead to a reduction of the robustness achieved by using only the restricted pool $\mathcal{P}$. However, if the sample size is not too small, the reduced robustness should be insignificant.

2.1. Rotation and reflection invariance properties of the SID method

To show that the SID method is invariant under both rotation and reflection transformations, it suffices to show that it is invariant under any orthonormal transformation. Consider the simple linear model case at (2.1) with the data $\mathcal{W}_n = \{W_1, \ldots, W_n\}$, where $W_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$. Let $\mathbf{H}$ be a 2×2 orthonormal matrix, i.e., $\mathbf{H}^t\mathbf{H} = \mathbf{H}\mathbf{H}^t = \mathbf{I}$, and consider the orthonormal transformation of the original data, $\tilde{\mathcal{W}}_n = \{\tilde{W}_1, \ldots, \tilde{W}_n\}$, where $\tilde{W}_i = \mathbf{H}W_i$. Let L be a given line on the plane, say, $ax + by = c$ for some fixed constants a, b and c. Let $\tilde{L}$ denote the line $\tilde{a}x + \tilde{b}y = c$, where $\begin{pmatrix} \tilde{a} \\ \tilde{b} \end{pmatrix} = \mathbf{H}\begin{pmatrix} a \\ b \end{pmatrix}$.

Proposition 2.1. $\mathrm{SID}_{\mathcal{W}_n}(L) = \mathrm{SID}_{\tilde{\mathcal{W}}_n}(\tilde{L})$, *where* $\mathrm{SID}_{\mathcal{W}_n}(L)$ *denotes the SID value of the line* L *w.r.t. the dataset* $\mathcal{W}_n$*, and* $\mathrm{SID}_{\tilde{\mathcal{W}}_n}(\tilde{L})$ *the SID value of the line* $\tilde{L}$ *w.r.t. the dataset* $\tilde{\mathcal{W}}_n$*.*

Proof. Recall that under a orthonormal transformation, any geometric figure is transformed into another figure which is exactly congruent to the original one, i.e., the transformed object is identical to the original one except for a possible orientation difference. The length, area, volume, etc. are unchanged. Thus the ratio of the intercept to the longest side in each triangle used to define the SID is also preserved.

As a corollary of the above proposition, if $ax + by = c$ is the best fitting line by the SID method for the dataset $\mathcal{W}_n$, then the best fitting line by the SID method for the transformed dataset $\tilde{\mathcal{W}}_n$ is simply the corresponding $\tilde{a}x + \tilde{b}y = c$.

Consider the special case of the reflection transformation, i.e., $\mathbf{H} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. Then $\tilde{W}_i = \begin{pmatrix} y_i \\ x_i \end{pmatrix}$ and $\begin{pmatrix} \tilde{a} \\ \tilde{b} \end{pmatrix} = \begin{pmatrix} b \\ a \end{pmatrix}$. This implies that if the (x, y) variables reverse their roles, then the SID method preserves the original linear relationship between the two variables. This is a desirable property since, in many practical situations, it is often not clear which variable is more suitable as the input variable, see the examples mentioned in the introduction and Section 3.1.

There is an intrinsic difference between the goals of regression and those of linear fitting. Specifically, a regression method is generally designed to predict the output variable on the basis of the input variable and hence it is crucial to identify the input variable. On the other hand, the goal of the linear fitting is simply to find the best linear relationship that exits between the variables. In particular, a vertical line is always considered a nonfit in linear regression, but it may very well be a fit in the linear fitting case.

We conclude this section with some remarks on invariance properties. The regression depth described in Rousseeuw and Hubert (1999) is not invariant under rotations and reflections, but it is invariant under monotone transformations of the output variable. This last invariance is advantageous for eventual generalizations to non-linear regression settings. Although the SID is invariant under rotations and reflections, it is not invariant under monotone transformations of the output variable.

2.2. Breakdown properties of the SID method

Let $y = b_0 + b_1 x$ be a fitted line for the given dataset of paired observations $\mathcal{W}_n = \{(x_i, y_i)^t, i = 1, \ldots, n\}$. The parameters b_0 and b_1 are allowed to be infinity, in which case the above equation conventionally denotes a vertical line with the form $x = c$, for some constant c.

We adopt the following definition of *breakdown* for a linear fit. It defines the breakdown of a fitted line as the minimum of the breakdowns of it's estimators of the coefficient parameters, as shown in (2.5). This definition is, in essence, equivalent to the one given in Chapter 1 of Rousseeuw and Leroy (1987).

Definition 2.2. Let m be the minimum number of new arbitrary pairs of (x_i, y_i)'s needed to be added to the original dataset to bring at least one of the two estimators for $\{\beta_0, \beta_1\}$ arbitrarily close to any chosen value. Then the linear fit is said to have the *finite sample breakdown* $d_n = m/(m+n)$ and the *limiting breakdown* $d_\infty = \lim_{n\to\infty} d_n$, where n is the size of the original sample $\mathcal{W}_n$.

We establish below a non-zero lower bound for the breakdown of the SID fit, and then show that the breakdown is zero for the least squares method, the L_1 method, and both L_2- and L_1- versions of the orthogonal regression method, denoted, respectively, OR-method and L1OR-method.

For the given dataset $\mathcal{W}_n$, let d be the maximum SID value attained by its SID fit, i.e., $d = \sup_{b_0,b_1} \mathrm{SID}(b_0, b_1)$, where $\mathrm{SID}(b_0, b_1)$ is the SID value for the line $y = b_0 + b_1 x$ w.r.t. the given dataset. To move the slope b_1 to an arbitrary value, we would need to have sufficient contamination in order to move the fitted line to a line whose original SID value is d_1, where $d_1 = \inf_{b_1} \sup_{b_0} \mathrm{SID}(b_0, b_1)$. Note that $d - d_1 > 0$ if there exists a unique maximizer for the SID value for the original data (cf. Remark 2.1).

Now, assume that we contaminate the original dataset by adding m points. These m points generate $T(m) \equiv \binom{m}{3} + \binom{m}{2}n + m\binom{n}{2}$ new triangles. The new SID value for the original best SID fit will be at least $d\binom{n}{3}/\binom{n+m}{3}$, and the new SID value for the line whose original SID value is d_1 will be at most $[d_1\binom{n}{3} + T(m)]/\binom{n+m}{3}$. If, after the contamination, the line with original SID value d_1 amasses higher new SID than the original best fit, then $d\binom{n}{3} \leq d_1\binom{n}{3} + T(m)$ or, equivalently, $(d - d_1)\binom{n}{3} \leq T(m)$. If $m = \lambda n$, then the above inequality can be expressed as $d - d_1 \leq \lambda^3 + 3\lambda^2 + 3\lambda + O(n^{-1})$.

Denote by $d_n(\beta_1)$ the finite sample breakdown for the estimators of β_1, i.e., $d_n(\beta_1) = m/(m + n) = \lambda/(1 + \lambda)$. We observe the bound

$$d - d_1 \leq S^3 + 3S^2 + 3S + O(n^{-1}) \equiv g(d_n(\beta_1)) + O(n^{-1}),$$

where $S = d_n(\beta_1)/(1 - d_n(\beta_1))$. Since g is an increasing function of $d_n(\beta_1)$, we arrive at

$$d_n(\beta_1) \geq g^{-1}(d - d_1), \tag{2.3}$$

where $g^{-1}(a) > 0$ if $a > 0$. Thus $d_\infty(\beta_1) > 0$ as long as $(d-d_1)$ does not converge to zero as $n \to \infty$.

Similar arguments can be used to establish that

$$d_n(\beta_0) \geq g^{-1}(d - d_0), \tag{2.4}$$

where $d_0 = \inf_{b_0} \sup_{b_1} \mathrm{SID}(b_0, b_1)$. Assume that $(d - d_0) > 0$ for the original data (cf. Remark 2.1). Then the outcomes in (2.3) and (2.4) lead to the nonzero limiting lower bound for the SID fit

$$d_n = \min(d_n(\beta_0), d_n(\beta_1)) \geq \min\{g^{-1}(d - d_0), g^{-1}(d - d_1)\}. \tag{2.5}$$

Remark 2.1. Both $(d - d_0)$ and $(d - d_1)$ are positive if there exists a unique maximizer of the SID. The latter holds except for some pathological examples, such as the case of spherically symmetric data.

We now proceed to show the zero breakdown for the linear fit obtained by orthogonal regression (OR). Similar arguments holds for its L_1 version (L1OR), and for the usual L_1 and least squares methods.

Recall that the best fitted line with the OR method is obtained by minimizing the sum of the squared orthogonal distances of the sample points to the candidate line. Replacing the squared distance above with the absolute distance (i.e., the squared root of the squared distance) yields the best fitted line for the L1OR method. It can be shown that the sample OR best fitted line for the given model in (2.1) is $y = \beta_0^* + \beta_1^* x$ with $\beta_0^* = \bar{y} - \beta_1^* \bar{x}$, when β_1^* is the minimizer of

$$\sum (y_i - \bar{y} + \beta_1^* \bar{x} - \beta_1^* x_i)^2 / [1 + (\beta_1^*)^2].$$

Suppose we want to move the slope β_1^* of the best OR fitted line $y = \beta_0^* + \beta_1^* x$ to be arbitrarily close to a chosen value $\tilde{\beta}_1$. Choose a point (x^*, y^*) on the line $y = \beta_0^* + \tilde{\beta}_1 x$. Consider contaminating the given dataset by adding (x^*, y^*), and let $(x^*, y^*) \to (\infty, \infty)$ along the line $y = \beta_0^* + \tilde{\beta}_1 x$. If $\beta_1^* \neq \tilde{\beta}_1$, as $(x^*, y^*) \to (\infty, \infty)$, the orthogonal distance from (x^*, y^*) to $y = \beta_0^* + \beta_1^* x$ grows to ∞. Therefore, the total squared orthogonal distance of the new dataset for the line $y = \beta_0^* + \beta_1^* x$ also grows to ∞. Since one can keep the total orthogonal distance bounded by choosing a line passing through (x^*, y^*), the slope of the fitted line for the contaminated dataset is forced to converge to $\tilde{\beta}_1$. This shows that a single point contamination is sufficient to cause the breakdown of the slope of the OR fit, and thus the breakdown for the OR method is zero.

Remark 2.2. It is worth noting that the breakdown of the intercept is 1/2 for the usual L_1 regression and the L1OR method. The L1OR is clearly more robust than the OR method, just as the L_1 regression is more robust than the LS method.

3. Examples and Empirical Comparisons

The SID method is applied to three datasets. The first dataset consists of 1994 SAT averages for verbal and math scores for the 50 states in the USA (Kitchens, (1998, pp.346-347)); the second is the Hertzprung-Russell diagram of a star cluster in the direction of Cygnus (from Rousseeuw and Leroy (1987, p.27)); the third (from Rousseeuw and Leroy (1987, p.26)) records the number of international telephone calls from Belgium in the years 1950-1973. The last two datasets have been studied extensively in Rousseeuw and Hubert (1999). We apply to these three datasets the following linear fitting methods: least squares (LS), L_1, OR, L1OR, SID and regression depth (RD).

3.1. Invariance comparison

Based on the scatterplot (in Figure 1) and the physical meanings of the SAT scores, it is unclear whether the verbal or the math score should be the input variable. Since SID, OR and $L1$OR methods are reflection invariant, we present only their linear fits with math score as the output variable. The equations listed next to LS and LS' are the LS fits obtained by using, respectively, verbal and math score as the input variable. The last LS' simply inverts the two variables in the previous LS. Clearly, LS' and LS are different. This difference illustrates that the LS method is not reflection invariant.

$$\begin{aligned} \text{SID}: &\quad \text{Math} = -8.2828 + 1.1414\text{Verbal},\\ \text{OR}: &\quad \text{Math} = 8.2377 + 1.1029\text{Verbal},\\ L1OR: &\quad \text{Math} = 3.7778 + 1.1111\text{Verbal},\\ \text{LS}: &\quad \text{Math} = 34.7676 + 1.0436\text{Verbal},\\ \text{LS}': &\quad \text{Verbal} = 12.8299 + .8662\text{Math},\\ \text{LS}': &\quad \text{Math} = -19.8117 + 1.1545\text{Verbal}. \end{aligned}$$

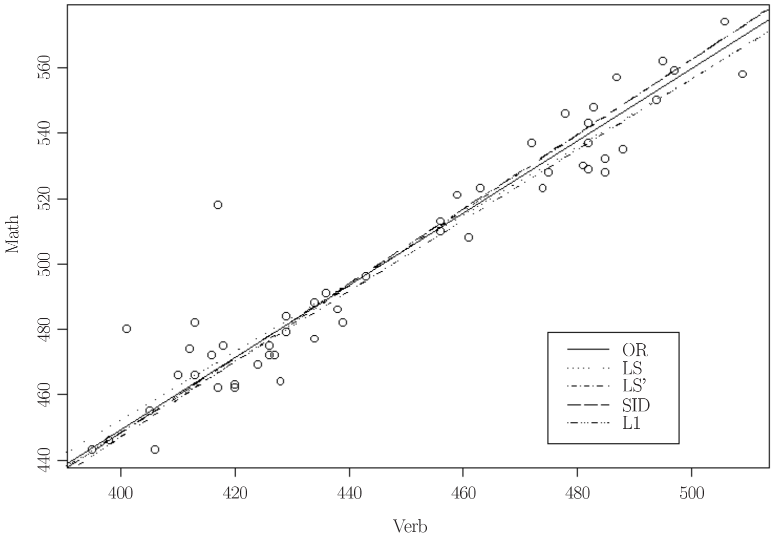

Figure 1. SAT scores data.

3.2. Robustness comparison

Figure 2 contains the fitted lines for the star data from five methods. To avoid crowding, the fitted lines are presented in two separate plots. The respective equations for the fitted lines are:

$$\begin{aligned} \text{LS}: &\quad \hat{y} = 6.7935 - 0.4133x,\\ L_1: &\quad \hat{\hat{y}} = 8.1492 - 0.6922x, \end{aligned}$$

$$\begin{aligned} \text{SID}: &\quad y^* = -15.8164 + 4.7273x, \\ \text{RD}: &\quad y^{**} = -7.3258 + 2.7875x, \\ \text{OR}: &\quad \tilde{y} = 35.42935 + -7.05736x. \end{aligned}$$

Note that the small cluster of four giant stars on the upper left corner in Figure 2 are quite far away from the rest of the data. They are generally viewed as outliers. Clearly, the LS and the L_1 lines are unduly influenced by the outliers. The OR line provides a slight improvement, but it still does not capture the linear structure of the majority of the data. The SID and RD lines practically ignore the cluster of outliers, and go through the bulk of the major cluster. Furthermore, the SID line appears to be insensitive even to the points sitting between the major cluster and the cluster of the four outliers.

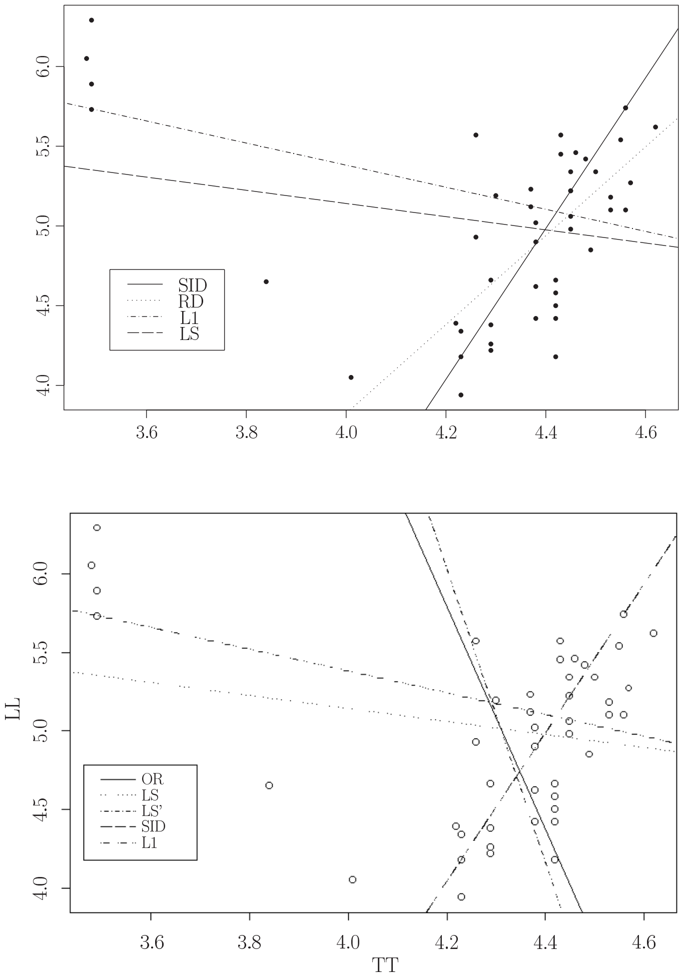

Figure 2. Stars data.

Figure 3, again in two separate plots, contains the five fitted lines for the telephone calls data. The respective equations for the fitted lines are:

$$\begin{aligned}
\text{LS}: &\quad \hat{y} = -26.0059 + 0.5041x,\\
L_1: &\quad \hat{\hat{y}} = -7.519 + 0.153x,\\
\text{SID}: &\quad y^* = -5.1632 + 0.1105x,\\
\text{RD}: &\quad y^{**} = -7.8623 + 0.1575x,\\
\text{OR}: &\quad \tilde{y} = -48.4934 + 0.8698x.
\end{aligned}$$

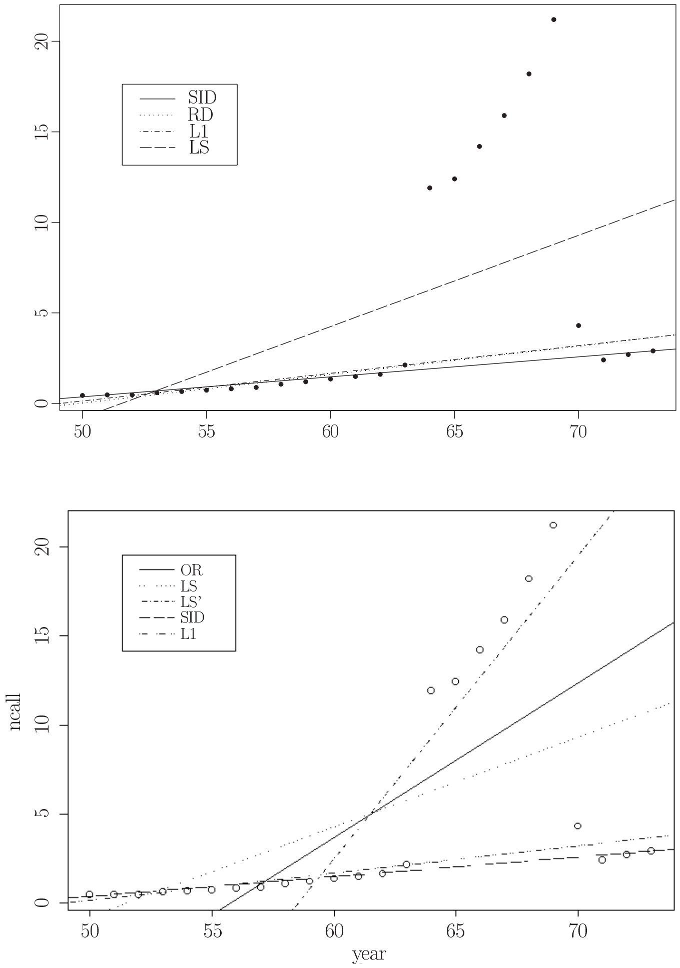

Figure 3. Telephone calls data.

In this case, the six points which curve up in the middle of the scatter plot are viewed as outliers. Here, the LS line is affected the most by the outliers.

The L_1 line and the RD line almost coincide with each other, and they show a substantial improvement over the LS line. The SID line lies at the bottom of the plot, and appears to be completely unaffected by the outliers.

3.3. Efficiency comparison

Asymptotic relative efficiency (ARE) is a standard measure for comparing two estimators. It provides a ratio of the sample sizes needed for the two estimators to achieve the same level of mean square error (mse). It is usually the limiting ratio of the mse's achieved by the two estimators. We present here a comparison in terms of the relative efficiency of the SID method to the LS, the OR, and the L1OR methods, in three simulated examples. All three examples are simulated from a normal model, with the last being contaminated by one standard Cauchy error variable. Note that, since the LS method is known to be optimal in the normal model, it should be expected that the relative efficiency of the SID to the LS method be less than 1. Recall that the relative efficiency is defined as $\sigma^2_{n,LS}/\sigma^2_{n,SID}$, where $\sigma^2_{n,LS}$ and $\sigma^2_{n,SID}$ are respectively the mse's of the estimators derived from the LS method and the SID method with the sample size n. We set $n = 20$ in our simulation examples, and omit n in all notations. This simulation study is repeated for the OR and the L1OR methods.

Example 3.1. Consider the model $y = 2 + \beta x + \epsilon$, where ϵ is $N(0, \sigma^2)$. The value of σ is chosen to be 0.01, 0.1, 1, or 2. The sample size n is 20. The independent variable x is assigned values which are equally spaced from 0 to 19. For each sample, the estimates for the intercept and the slope are calculated. This simulation is run 1,000 times for each value of σ.

Respectively for $\beta = 3$ and $\beta = 100$, Tables 3.1 and 3.2 list the relative efficiency $\sigma^2_{LS}/\sigma^2_{SID}$ of the SID method compared to the LS method in estimating the intercept and the slope under each value of σ. The relative efficiency is about 40% to 55%, comparable to the L1OR and L_1 methods in the same setting. The OR method, having a squared objective function similar to that of the regular LS method, retains high efficiency. Drawing from the efficiency study between regression depth and L_1 methods in Rousseeuw and Hubert (1999), we conclude that the SID and the regression depth methods have similar levels of efficiency loss when they are compared with the LS method. Note that the efficiency comparison here tends to be conservative for the SID fit, and as discussed in Section 5.

Example 3.2. Consider the model $y = 2 + \beta x + \epsilon$, where ϵ is $N(0, 1)$. The same simulation procedure as in Example 3.1 is repeated, except for that each sample now contains 19 observations generated from the model, and one contaminated observation generated from the standard Cauchy error variable (with center 0 and scale 1).

Table 3.1. Model: $y = 2 + 3x + \epsilon$.

	$\sigma^2_{LS}/\sigma^2_{SID}$		$\sigma^2_{LS}/\sigma^2_{OR}$		$\sigma^2_{LS}/\sigma^2_{L1OR}$	
σ	Intercept	Slope	Intercept	Slope	Intercept	Slope
0.01	0.4943	0.476232	0.999922	0.999874	0.650103	0.627649
0.1	0.523508	0.510057	0.998869	0.999137	0.634658	0.613791
1	0.445502	0.437643	0.989462	0.990643	0.637411	0.638469
2	0.447891	0.409257	0.991620	0.982825	0.663385	0.649616

Table 3.2. Model: $y = 2 + 100x + \epsilon$.

	$\sigma^2_{LS}/\sigma^2_{SID}$		$\sigma^2_{LS}/\sigma^2_{OR}$		$\sigma^2_{LS}/\sigma^2_{L1OR}$	
σ	Intercept	Slope	Intercept	Slope	Intercept	Slope
0.01	0.552381	0.502035	0.9999996	0.9999996	0.6630726	0.6560074
0.1	0.456335	0.448066	0.9999818	0.9999820	0.6826832	0.6741967
1	0.491115	0.460199	0.9997263	0.9997879	0.6412567	0.6318704
2	0.528545	0.512687	0.9996104	0.9996723	0.6466665	0.6422692

Table 3.3. Contamination with one Cauchy error variable.

	$\sigma^2_{LS}/\sigma^2_{SID}$	
model	Intercept	Slope
$y = 2 + 3x + \epsilon$	3079.849	7618.124
$y = 2 + 100x + \epsilon$	227.4719	613.0168

Even with only one Cauchy outlier (an equivalence of 5% contamination), it is evident from Table 3.3 that the SID method is of much greater efficiency than the LS method. Scanning through the results from our 1,000 runs, we notice that the SID estimates are generally quite stable and close to the true values, while the LS estimates often fluctuate widely. We show some of the extreme cases in the following table.

Table 3.4. Some extreme cases from the LS method.

run	Intercept_{SID}	Slope_{SID}	Intercept_{LS}	Slope_{LS}
386	1.29955061	3.07375115	64.45405029	-5.923531292
396	1.57862889	3.05072018	1078.798626	-150.812325
496	1.99894466	2.99129005	-28.77564588	7.38242724
623	1.36140617	3.05386946	25.50574458	-0.409166007
854	2.37564413	2.95549022	16.39880004	0.977085454
878	3.10025679	2.93226526	13.58867624	1.451392556
978	1.95624724	3.03040365	123.9870806	-14.41739117

3.4. A Visual display of breakdown in linear fitting

Figure 4 contains 1,000 lines obtained by the methods of LS, SID, L1OR and OR for the 1,000 simulated samples from the setting in Example 3.2 for the target line $y = 2+3x+\epsilon$. The 1,000 lines under SID are tightly bundled together which shows the relatively small effect from the Cauchy contamination. The plot under L1OR is the next best, with several lines turning vertical. The lines under OR are more spread out than those under L1OR but they lie within the first and third quadrants. Finally, the lines under LS are spread out in all directions. These four graphs suggest that SID is far more robust than the others.

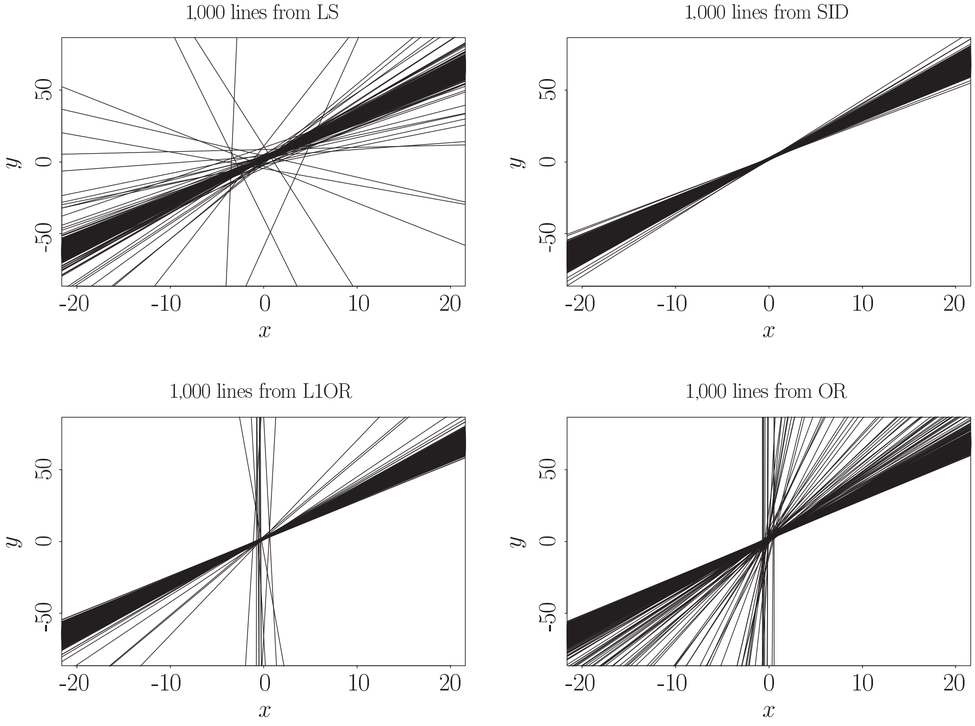

Figure 4. 1,000 simulated lines from Example 3.2.

4. Simplicial Fit Plot

In most regression methods, the residual plot is used as a tool to visually determine the goodness of the proposed fit. However, since the residuals do not have a standardized scale it can be difficult to comprehend the magnitude of the residuals. In this section, we propose a goodness-of-fit plot in a standardized scale, which provides the "depth" of "cut" through each triangle by the fitted line. We shall call this plot the *simplicial fit plot* (S-fit plot).

Recall that $\mathbf{\Delta} \equiv \{\Delta_1, \dots \Delta_{\binom{n}{3}}\}$ denotes the collection of all $\binom{n}{3}$ triangles generated from the sample of size n, and that the SID is the average of intercept ratios, $\{r_k = \ell_k / m(\Delta_k),\ k = 1, \dots, \binom{n}{3}\}$, (see (2.2)). The S-fit plot of a line L is the quantile plot of $\{r_k : k = 1, \dots, \binom{n}{3}\}$ As t grows from 0% to 100%, $q(t)$ indicates the corresponding tth quantile of $\{r_k : k = 1, \dots, \binom{n}{3}\}$. If the line cuts deeply into more triangles, $q(t)$ assumes higher values early on and tends to stay higher throughout. Figure 5 shows four S-fit plots for the fitted lines obtained in Section 3.2 for the stars data. There, the S-fit plots of the RD and the SID lines rise above zero sooner than those of the LS and the L_1 lines. The one for the SID line stays considerably higher compared to the others.

For a given fit, it is worth noting that the area under its S-fit plot is exactly its SID value. The perfect fit has the SID value 1, which corresponds to the area under its S-fit plot with $q(t) = 1$, $\forall\, t \in [0, 1]$. Smaller SID values correspond to slower rising S-fit curves. The S-fit plot, together with its corresponding SID value, can be a convenient graphical tool for a quick visual assessment of a given linear fit.

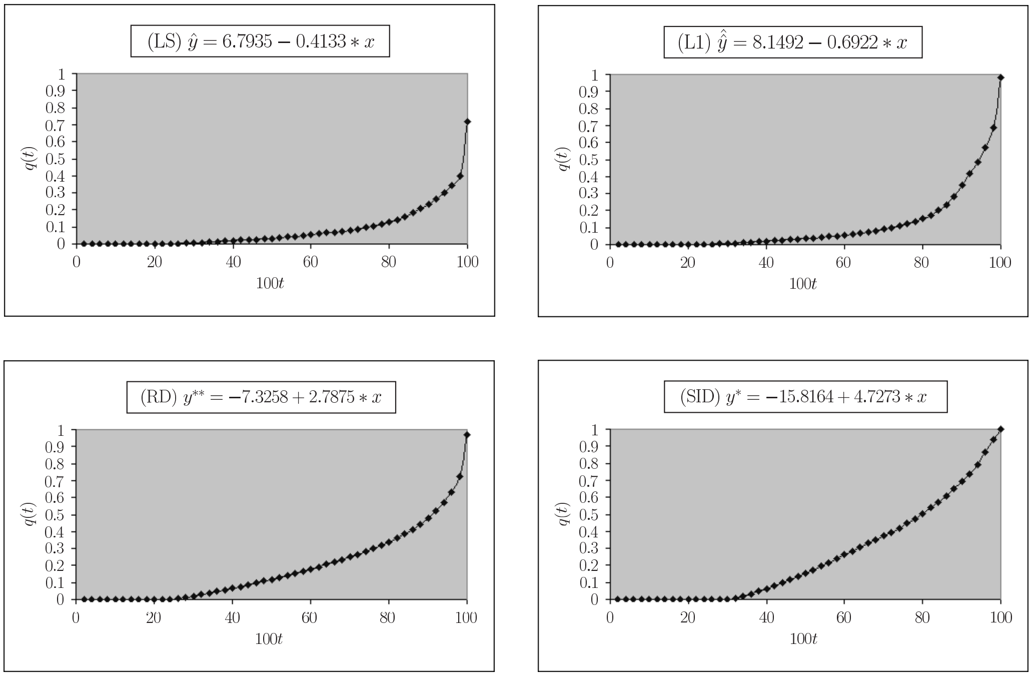

Figure 5. Simplicial fit plots for star data.

5. Concluding Remarks and Open Problems

In this article we have introduced the SID method for linear fitting by extending the ideas behind the simplicial depth described in Liu (1990). We show that the SID method is both rotation and reflection invariant. We also provide a nonzero lower bound for the breakdown of the SID fit. Some comparisons in reflection invariance, robustness and efficiency to other regression methods are presented. The SID fit appears to be quite robust on the real and simulated data.

The SID value quantifies the goodness-of-fit of a proposed linear fit, and it can be viewed as a robust analog of the coefficient of determination in the framework of the least squares method. A linear fit with a higher SID value is considered a better fit. The SID is 1 if and only if all the sample points are on the line, and it becomes smaller as the sample points are farther away from the line. It decreases to zero if the line falls completely outside of the data cloud. To illustrate how the SID characterizes the goodness-of-fit, the *simplicial fit plot* (S-fit plot) is proposed as a companion graph showing how the linear fit performs with respect to each triplet of the given sample. For a perfect fit, namely the SID is 1, the S-fit plot attains the value 1 throughout. The slower the S-fit plot curves up, the worse fit of the line is. The area under the curve is exactly the SID value. The SID value and the S-fit plot together provide a quick visual assessment of the goodness of a linear fit for a give dataset.

It seems natural to conjecture that if the underlying distribution of the sample points is elliptical, namely x and y jointly follow an elliptical distribution, then the population version of the SID is maximized along the major axis (or the principle component). We have established this claim under the assumption that the maximizer of the SID is unique. Although the general proof has eluded us, preliminary empirical simulations appear to support the claim. We also believe that as long as the distribution has a mirror symmetry and is elongated along a specific line, that line is the population version of the SID fit. This should hold for the orthogonal regression method, both the L_1 and the L_2 versions. Needless to say, the population version of the LS regression line does not necessarily coincide with the major axis. For example, if x and y follow a bivariate normal, then the LS line is $E(y|x) = \mu_y + (\sigma_{x,y}/\sigma_x^2)(x - \mu_x)$. This line is different from the major axis.

It is worth pointing out that the efficiency comparison presented in Section 3.3 is a conservative one for the SID method. The assumed models in Examples 3.1 and 3.2 are regression models. The lines being estimated are the target lines for LS and L_1 methods. This may not be the case for the SID fit, in light of the discussion in the paragraph above. Consequently, in addition to the increase in variance, there is a bias factor in the σ^2_{SID} presented in Tables 3.1 and 3.2. This bias factor contributes to some of the loss of efficiency in the SID fit. This explains why the σ^2_{SID} in both Tables should be viewed as the mean squared error and not just as the variance alone.

The asymptotics of the SID method (such as consistency and asymptotic distribution) are yet to be investigated. They are needed for making inferences with the SID method. They are also needed for a comparison with RD using the asymptotics of the RD established in Bai and He (1999) and He and Portnoy (1998). As for the breakdown properties, what we have provided in Section 2 is only a lower bound. It would be valuable to determine an exact, or a more precise, breakdown value.

Although it is conceptually straightforward to generalize the SID method to the higher dimensional case, developing a usable algorithm seems nontrivial. For example, in the linear fitting of the three dimensional case, the fit under consideration is a plane, and the SID measures how "deep" the plane "cuts" through the simplices (tetrahedrons in this case) generated by the sample points. Specifically, the SID here is the averaged ratio of the area of the intercept (the intersection of the plane inside the simplex) to the area of the largest face of the simplex formed by any four observations. The search for efficient algorithms for computing various notions of depth has generated much research interest in the computer science community (see, for example, Langerman and Steiger (2000)), and the prospects for fast computing algorithms for higher dimensional depth appear quite real.

Acknowledgement

The research was in part supported by the National Science Foundation, the Federal Aviation Administration, and the National Security Agency. The authors thank Ms. Lixia Pei for her assistance in computing.

References

Bai, Z. and He, X. (1999). Asymptotic distributions of the maximal depth estimators for regression and multivariate location. *Ann. Statist.* **27**, 1616-1637.

Bickel, P. (1973). On some analogues to linear combination of order statistics in the linear model. *Ann. Statist.* **1**, 597-616.

Carroll, R. and Ruppert, D. (1985). Transformations in regression: a robust analysis. *Technometrics* **27**, 1-12.

He, X. and Portnoy, S. (1998). Asymptotics of the deepest fit. *Statistical Inference and Related Topics: A Festschrift in Honor of A. K. Md. Saleh. Nova Science.* New York.

Hettmansperger, T. and McKean, J. (1977). A robust alternative based on ranks to least squares in analyzing linear models. *Technometrics* **19**, 275-284.

Hossjer, O. (1994). Rank-based estimates in the linear model with high breakdown point. *J. Amer. Statist. Assoc.* **89**, 149-158.

Johnstone, I. and Velleman, P. (1985). The resistant line and related regression methods. *J. Amer. Statist. Assoc.* **80**, 1041-1059.

Jurecková, J.(1971). Nonparametric estimate of regression coefficients. *Ann. Math. Statist.* **42**, 1328-1338.

Kitchens, L. (1998). *Exploring Statistics: A Modern Introduction to Data Analysis and Inference.* 2nd edition. Duxbury.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrics* **46**, 33-50.

Langerman, S. and Steiger, W. (2000). An optimal algorithm for hyperplane depth in the plane. *Proceedings 11th Annual SIAM Symposium on Discrete Algorithms*, 54-59.

Liu, R. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18**, 405-414.

Liu, R., Parelius, J. and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and Inference (with discussion). *Ann. Statist.* **27**, 783-858.

Liu, R. and Singh, K. (1999). Invited discussion on "Regression depth" by P. Rousseeuw and M. Hubert. *J. Amer. Statist. Assoc.* **94**, 407-409.

Maronna, R. and Yohai, V. (2000). Robust regression with both continuous and categorical predictors. *J. Plann. Inference* **89**, 197-214.

Rousseeuw, P. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79**, 871-880.

Rousseeuw, P. and Bassett, G. (1990). The remedian: a robust averaging method for large data sets. *J. Amer. Statist. Assoc.* **85**, 97-104.

Rousseeuw, P. and Hubert, M. (1999). Regression depth (with discussion). *J. Amer. Statist. Assoc.* **94**, 388-433.

Rousseeuw, P. and Leroy, A. (1987). *Robust Regression and Outlier Detection.* Wiley, New York.

Ruppert, D. and Carroll, R. (1980). Trimmed least squares estimation in the linear regression. *J. Amer. Statist. Assoc.* **75**, 828-838.

Sen, P. (1968). Estimates of the regression coefficient based on Kendall's tau. *J. Amer. Statist. Assoc.* **63**, 1379-1389.

Simpson, D., Ruppert, D. and Carrol, R. (1992). On one-step GM-estimates and stability of inferences in regression. *J. Amer. Statist. Assoc.* **87**, 439-450.

Stromberg, A., Hossjer, O. and Hawkins, D. (2000). The least trimmed difference regression estimator and alternatives. *J. Amer. Statist. Assoc.* **95**, 853-864.

Theil, J. (1950). A rank-invariant method of linear and polynomial regression analysis (parts 1-3). *Koninklijke Nederlandse Akademie var Wetenschappen Proceedings* **53**, 386-392, 521-525, 1397-1412.

Tukey, J. (1970). *Exploratory Data Analysis*, (limited Preliminary Edition). Addison-Wesley, Reading, Massachusetts.

Department of Statistics, Rutgers University, Hill Center, Piscataway, NJ 08854-8019, U.S.A.
E-mail: rliu@stat.rutgers.edu

Department of Statistics, Rutgers University, Hill Center, Piscataway, NJ 08854-8019, U.S.A.
E-mail: kesar@stat.rutgers.edu

Department of Statistics, Rutgers University, Hill Center, Piscataway, NJ 08854-8019, U.S.A.
E-mail: hteng@stat.rutgers.edu

(Received March 2001; accepted October 2003)