The Annals of Statistics
1995, Vol. 23, No. 2, 490–501

DIMENSION OF THE SINGULAR SETS OF PLANE-FITTERS

By Steven P. Ellis

University of Rochester and University of Medicine and Dentistry of New Jersey

Let $n > p > k > 0$ be integers. Let δ be any technique for fitting k-planes to p-variate data sets of size n, for example, linear regression, principal components or projection pursuit. Let $\mathscr{S}$ be the set of data sets which are (1) singularities of δ, that is, near them δ is unstable (for example, collinear data sets are singularities of least squares regression) and (2) nondegenerate, that is, their rank, after centering, is at least k. It is shown that the Hausdorff dimension, $\dim_H(\mathscr{S})$, of $\mathscr{S}$ is at least $nk + (k+1)(p-k) - 1$. This bound is tight.

Under hypotheses satisfied by some projection pursuits (including principal components), $\dim_H(\mathscr{S}) \geq np - 2$, that is, once singularity is taken into account, only two degrees of freedom remain in the problem!

These results have implications for multivariate data description, resistant plane-fitting and jackknifing and bootstrapping plane-fitting.

1. Introduction. Ellis (1991a) investigates plane-fitting, that is, the general class of multivariate procedures including linear regression and (some) projection pursuits (including principal components). It is shown that under very mild hypotheses, a plane-fitting technique must have singularities, that is, data sets near which the plane-fitting technique is unstable (see Section 2 for precise definitions).

Others have investigated the singularity of particular plane-fitting techniques. The singularities of least squares regression are precisely the (multi)collinear data sets [Belsley (1991)]. Hettmansperger and Sheather (1992) investigated the singularities of the least median of squares regression.

Singularity is a very basic issue in plane-fitting, or data analysis generally. It is undesirable, or at least creates a need for caution. For example, if one wishes to make an inference about a population plane from a multivariate sample, a confidence set for the plane ordinarily will be large if the data lie near a singularity [see Examples 2.3 and 2.4 in Ellis (1991a)].

Even if one wishes to fit a plane to data, not to make an inference, but merely to describe the linear tendency in the data, doing so near a singularity requires caution. Near a singularity, the sort of linear pattern to which the plane-fitter is sensitive will only be weakly or ambiguously present and so the fitted plane will be a poor data summary.

Received July 1992; revised June 1994.

AMS 1991 *subject classifications*. Primary 62H99; secondary 62J99.

Key words and phrases. Bootstrap, collinearity, Hausdorff dimension, jackknife, principal components, projection pursuit, regression.

Ellis (1991a) concerns existence and severity of singularities. A practical issue is how often will they be a problem? Investigating this in a general way means asking how large is the collection of data sets near the set of singularities. The dimension ("degrees of freedom") of the set of singularities gives insight into this question. (This is illustrated in Section 2.) The hard part is assessing the dimension of the "singular set" of the plane-fitter, that is, the collection of its "nondegenerate" singularities.

Ellis (1993a) measures the size of the singular set using a notion analogous to covering dimension. Here, I give a tight lower bound on the Hausdorff dimension of the singular set (Theorem 2.6). The lower bound is surprisingly large. More surprising is Theorem 2.2. Under its mild hypotheses, which are satisfied by important projection pursuit methods (like principal components plane-fitting), singularity uses up all but at most 2 degrees of freedom.

Singularity in plane-fitting is due to topological aspects of the global geometry of the plane-fitting problem. So to prove general results, topological methods are used in Ellis (1991a, 1993a) and here. In the present paper I also use a few facts about Hausdorff dimension.

By using such tools one can get nontrivial results with almost vacuous hypotheses. In particular, unlike most results in mathematical statistics, the theorems in this paper are neither asymptotic nor do they include any distributional assumptions.

In fact, the results are so general that they can be applied to plane-fitting behavior, not just plane-fitting algorithms. Let a data analyst be free to fit a plane after unlimited data snooping and subjective judgment. If he/she recognizes a perfect fit when presented with one, the results of Ellis (1991a, 1993a) and the present paper apply to him/her regarded as a plane-fitter. (The fine print: The data analyst must behave like a function, that is, when confronted with the same data set under the same circumstances, he/she must always fit the same plane. To apply Theorem 2.2 below, something must be known about the analyst's handling of data sets which are nearly a perfect fit.)

There is, of course, a cost to such great generality. The results give only qualitative information. The techniques used here need to be developed to give more information under stronger assumptions.

I expect that the topological approach could be fruitfully applied to investigate the singularities of many multivariate techniques [e.g., Ellis (1991b)].

The results are presented in the next section. They raise some issues concerning multivariate data description, resistant plane-fitting and jackknifing and bootstrapping plane-fitting. These are briefly discussed in Section 3. Proofs are sketched in Section 4.

2. Main results.

Let

$$n > p > k > 0$$

be fixed integers. A "data set" is a point in the space, $\mathscr{Y}$, of all $n \times p$ matrices with the topology of $\mathbb{R}^{np}$. (So a data set consists of n p-dimensional observations; $\mathbb{R}$ = reals.) By a "plane" I mean a linear manifold, that is, an affine subspace of Euclidean space.

"Fitting" a plane to $Y \in \mathscr{Y}$ means assigning to it a k-dimensional plane in $\mathbb{R}^p$ describing its linear structure. Operationally, a plane-fitting technique will be any rule assigning k-planes to data sets which gives the right answer at those data sets that lie exactly on a unique k-plane. (This is made precise presently.) Let δ be such a rule. It may not be defined at literally every data set, but assume it is defined on a dense subset $\mathscr{Y}' \subset \mathscr{Y}$. One can study the behavior of δ via the map Φ defined as follows. If $Y \in \mathscr{Y}'$, let $\Phi(Y)$ be the k-dimensional subspace (i.e., plane through the origin) of $\mathbb{R}^p$ parallel to $\delta(Y)$. The range of Φ is the "Grassman manifold," $G(k, p)$, consisting of all k-planes in $\mathbb{R}^p$ passing through the origin [Boothby (1975), page 63]. The map Φ is a *plane-fitter*.

I now formalize this. If $Y \in \mathscr{Y}$, let γ denote the lowest dimensional plane in $\mathbb{R}^p$ containing all the rows of Y. Let $\Delta(Y)$ be the linear subspace of $\mathbb{R}^p$ (i.e., plane through the origin) parallel to γ and having the same dimension. [Thus, if $y \in \gamma$, then $\Delta(Y) = \gamma - y$.] Y is *degenerate* if $\Delta(Y)$ has dimension less than k. Let $\mathscr{D} \subset \mathscr{Y}$ be the set of all degenerate data sets. Let $\mathscr{P}_k = \{Y \in \mathscr{Y}$: dimension of $\Delta(Y) = k\}$. $\mathscr{P}_k$ consists of those data sets which are perfect fits, that is, $Y \in \mathscr{P}_k$ if and only if the rows of Y lie exactly on a unique k-plane. Assume $\mathscr{P}_k \cap \mathscr{Y}'$ is dense in $\mathscr{P}_k$. A *plane-fitter* is a map, Φ, from $\mathscr{Y}'$ into $G(k, p)$ satisfying the following:

$$\text{if } Y \in \mathscr{P}_k \cap \mathscr{Y}', \text{ then } \Phi(Y) = \Delta(Y). \tag{2.1}$$

From now on Φ will denote a plane-fitter.

EXAMPLE 2.1 (Notation) Suppose $p = 2$ and $k = 1$. In this case, a data set (i.e., an $n \times 2$ matrix) Y is in $\mathscr{P}_k$ if and only if its rows lie exactly on a unique line in $\mathbb{R}^2$. The line need not pass through the origin. $\Delta(Y)$ is the line through the origin parallel to this line. Let 1_n be the $n \times 1$ column vector of 1's. Then Y is in $\mathscr{P}_k$ if and only if the rows of Y are not all the same and, if Y^1, Y^2 are the columns of Y, there exist $a, b \in \mathbb{R}$ s.t. (such that) $Y^2 = a1_n + bY^1$ [in which case $\Delta(Y) = \{(x, bx) \in \mathbb{R}^p = \mathbb{R}^2, x \in \mathbb{R}\}$] or $Y^1 = a1_n + bY^2$ [in which case $\Delta(Y) = \{(by, y) \in \mathbb{R}^2, y \in \mathbb{R}\}$]. Y is degenerate if and only if all its rows coincide, that is, lie exactly on a 0-plane. Then $\Delta(Y)$ is the origin $\{(0, 0)\}$. [If $Y \notin \mathscr{P}_k$ and $Y \notin \mathscr{D}$, then $\Delta(Y) = \mathbb{R}^2 = \mathbb{R}^p$.]

Now suppose δ is (simple) least squares linear regression. If a and b are the estimated intercept and slope, respectively, from a data set Y, then $\delta(Y)$ is the line (1-plane) $\{(x, a + bx) \in \mathbb{R}^p, x \in \mathbb{R}\}$. If Φ is the plane-fitter corresponding to δ, then $\Phi(Y) = \{(x, bx) \in \mathbb{R}^p, x \in \mathbb{R}\}$. If Y is nondegenerate, but all the entries in Y^1 are equal, then Y is collinear and $\Phi(Y)$ is not defined. We may choose to define $\Phi(Y) = y$-axis, but even if we do not, Φ is still defined on a dense subset of Y and on a dense subset of $\mathscr{P}_k$. Moreover, (2.1) holds.

A *singularity* of Φ is a data set $Y_0 \in \mathscr{Y}$ s.t. $\lim_{Y \to Y_0} \Phi(Y)$ does not exist. For example, the collinear data sets are the singularities of least squares regression. The *singular set*, $\mathscr{S}_\Phi$, of Φ is the set of all its nondegenerate singularities.

Recall the definition of Hausdorff dimension [Falconer (1990), Chapter 2 and Morgan (1988), page 9. See Falconer (1990), Chapter 3 for alternative definitions of dimension]. If m is a positive integer and $A \subset \mathbb{R}^m$, the Hausdorff dimension of A, denoted by $\dim_H(A)$, is the supremum of the set of $r \geq 0$ for which the following quantity is positive:

$$\lim_{\varepsilon \downarrow 0} \inf_{\substack{A \subset \bigcup S_j \\ \operatorname{diam}(Sj) \leq \varepsilon}} \sum \operatorname{diam}(S_j)^r.$$

Here, the infimum is taken over all countable coverings $\{S_j\}$ of A, each element of which has diameter no larger than $\varepsilon > 0$. For calibration, note that if A is a smooth q-dimensional manifold, then $\dim_H(A) = q$.

Typically, $\mathscr{S}_\Phi$ will have Lebesgue measure 0 so the possibility of getting a data set in $\mathscr{S}_\Phi$ can be ignored. However, by definition, Φ is unstable near $\mathscr{S}_\Phi$ and $\dim_H(\mathscr{S}_\Phi)$ gives information about the volume of the set of points near $\mathscr{S}_\Phi$. In fact, we have the following. If $Y \in \mathscr{Y}$, let $|Y| = \sqrt{\operatorname{trace} Y^T Y}$, where the superscript T indicates transposition. If $R > 0$, let

$$\mathscr{S}_\Phi(R) = \{Y \in \mathscr{S}_\Phi : |Y| < R\}$$

and if $\varepsilon > 0$, let $\mathscr{T}_\varepsilon(R) = \{Y \in \mathscr{Y}: |Y - Y'| \leq \varepsilon \text{ for some } Y' \in \mathscr{S}_\Phi(R)\}$. So $\mathscr{T}_\varepsilon(R)$ is the set of points within ε of $S_\Phi(R)$. Let $\mathscr{L}$ be Lebesgue measure on $\mathscr{Y}$. Suppose $0 < \eta < \dim_H[\mathscr{S}_\Phi(R)] < np$. Then as $\varepsilon \downarrow 0$, $\mathscr{L}[\mathscr{T}_\varepsilon(R)] \to 0$, but more slowly than $\varepsilon^{np-\eta}$, that is,

$$\lim_{\varepsilon \to 0} \varepsilon^{\eta - np} \mathscr{L}[\mathscr{T}_\varepsilon(R)] = \infty$$

[Falconer (1990), pages 42–43 or Proposition 1.1 in Ellis (1993b)]. In the interest of brevity, most technical details are omitted from the present paper. See the last reference for a thorough treatment of them. It follows that if Y is a random element of $\mathscr{Y}$ with a continuous nowhere vanishing density, then $\Pr\{Y \in \mathscr{T}_\varepsilon(R)\}$ goes to 0 as $\varepsilon \to 0$ more slowly than $\varepsilon^{np-\eta}$.

Let

$$d = nk + (k+1)(p-k) - 1.$$

THEOREM 2.2 *Let* $R > 0$. *Suppose* $\dim_H[\mathscr{P}_k \cap \mathscr{S}_\Phi(R)] < d$. *Then* $\dim_H[\mathscr{S}_\Phi(R)] \geq np - 2$.

Under the hypotheses of the theorem, singularity accounts for all but at most two of the available degrees of freedom.

The idea of the proof is as follows. First, one proves the theorem under the assumption that Φ has no singularities in an open neighborhood of $\mathscr{P}_k$. A subset of $\mathscr{Y}$ of positive volume is filled with disjoint two-dimensional cones, each of which, for topological reasons, contains at least one singularity. [In Ellis (1991a, 1993a), cones are also used, but they have higher dimension.

The use of two-dimensional cones leads to sharp bounds.] The lower bound $np - 2$ is then intuitive. (Recall that $\mathscr{Y}$ has dimension np.) Next, suppose $\dim_H[\mathscr{P}_k \cap \mathscr{S}_\Phi(R)] < \dim_H(\mathscr{P}_k) - 1$ and $\dim_H[\mathscr{S}_\Phi(R)] < np - 2$. This can be reduced to the first case by pulling the singularities on the manifold $\mathscr{P}_k$ away from it along directions normal to $\mathscr{P}_k$. The result is a plane-fitter Θ covered by the first case. However, each point of $\mathscr{P}_k \cap \mathscr{S}_\Phi(R)$ has been stretched out into a set of dimension $\dim_H(\mathscr{Y}) - \dim_H(\mathscr{P}_k) - 1$ of possible singularities of Θ. Thus, the set of singularities of Θ which arose from points of $\mathscr{P}_k \cap \mathscr{S}_\Phi(R)$ has dimension less than $[\dim_H(\mathscr{P}_k) - 1] + [\dim_H(\mathscr{Y}) - \dim_H(\mathscr{P}_k) - 1] = np - 2$. The remainder of $\mathscr{S}_\Theta(R)$ was already assumed to be of dimension less than $np - 2$. This contradicts what we found in the first case. Thus, the integer d is just $\dim_H[\mathscr{P}_k] - 1$.

Plane-fitters satisfying the hypotheses of Theorem 2.2 are not uncommon.

EXAMPLE 2.3 (Principal components plane-fitting). Let $Y \in \mathscr{Y}$. The covariance matrix of Y is $S(Y) = n^{-1}(Y - 1_n \bar{y})^T(Y - 1_n \bar{y})$, where $\bar{y}$ is the mean of the rows of Y. (Recall that 1_n is the $n \times 1$ column vector of 1's.)

Principal components plane-fitting based on the covariance matrix is the plane-fitter, Ψ, defined as follows. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be the eigenvalues of $S(Y)$ and if $\lambda_k > \lambda_{k+1}$, let $v_1, \dots, v_k$ be (row) eigenvectors of $S(Y)$ corresponding to $\lambda_1, \dots, \lambda_k$, respectively. $\Psi(Y)$ is defined to be the plane spanned by $v_1, \dots, v_k$. It is not hard to see that Ψ has no singularities in $\mathscr{P}_k$. Thus, by Theorem 2.2, $\dim_H[\mathscr{S}_\Psi(R)] \geq np - 2$ for every $R \in (0, \infty)$. In fact, $\dim_H[\mathscr{S}_\Psi(R)] = np - 2$.

The principal components plane-fitter, Ψ_r, based on the correlation matrix is defined similarly. Let $D = D(Y)$ be the $p \times p$ diagonal matrix whose diagonal is the same as that of $S(Y)$. The correlation matrix of Y is $R(Y) = D^{-1/2}S(Y)D^{-1/2}$, providing D^{-1} exists. [Otherwise, $R(Y)$ is not defined.] Let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_p \geq 0$ be the eigenvalues of $R(Y)$, providing it exists, and if $\mu_k > \mu_{k+1}$, let $w_1, \dots, w_k$ be (row) eigenvectors of $R(Y)$ corresponding to $\mu_1, \dots, \mu_k$, respectively; $\Psi_r(Y)$ is defined to be the plane spanned by $w_1 D^{1/2}, \dots, w_k D^{1/2}$, so Ψ_r is a plane-fitter.

Ψ_r has some singularities in $\mathscr{P}_k$. Let $\mathscr{Z}$ denote the set of data sets in $\mathscr{P}_k$ whose correlation matrix is not defined. Suppose $k > 1$. Then $\mathscr{P}_k \cap \mathscr{S}_{\Psi_r}(R) \subset \mathscr{Z}$, but $\dim_H(\mathscr{Z}) < d$. Thus, by Theorem 2.2, if $k > 1$, then $\dim_H[\mathscr{S}_{\Psi_r}(R)] \geq np - 2$. (Theorem 2.6 below can be applied if $k = 1$.)

EXAMPLE 2.4 (Projection pursuit plane-fitting). If $\xi \in G(k, p)$ and $Y \in \mathscr{Y}$, let the rows of $\Pi_\xi(Y) \in \mathscr{Y}$ be the orthogonal projections onto ξ of the rows of Y. Let $\mathscr{R} = \mathscr{P}_k \cup \mathscr{D}$, so $\Pi_\xi(Y) \in \mathscr{R}$. Let Q be an $\mathbb{R}$-valued function on $\mathscr{R}$ and let $Q_Y(\xi)$ denote $Q[\Pi_\xi(Y)]$ regarded as a function of ξ with Y held constant. In projection pursuit with *projection index* Q, a plane is chosen which maximizes $Q_Y(\xi)$. $Q[\Pi_\xi(Y)]$ is supposed to quantify how "interesting" the projection $\Pi_\xi(Y)$ is. See Huber (1985) for an overview.

As a referee points out, projection pursuit fits planes for a different purpose than that which I used above for motivation. It is designed to uncover nonlinear structure, not linear. Just the same, singularities are important in projection pursuit. In fact, suppose Q is continuous on $\mathscr{R}$ and $Y_0 \in \mathscr{Y}$ is a singularity of the projection pursuit method with projection index Q. Then it is clear that Q_{Y_0} achieves its maximum at more than one plane. So if Y_0 has one interesting projection, it has at least one other, that is, it has a rich structure. If $Y \in \mathscr{Y}$ is near Y_0, then there are at least two planes which (nearly) maximize Q_Y. This provides a mathematical justification for the common practice of searching for several planes which (nearly) maximize Q_Y.

First, we consider a version of Friedman–Tukey projection pursuit [Friedman and Tukey (1974)]. Here, the projection index takes the form,

$$Q(Z) = s(Z)c(Z), \qquad Z \in \mathscr{R}$$

where $s(Z)$ measures the spread and $c(Z)$ the extent of clustering of the rows of $Z \in \mathscr{R}$.

Given $Y \in \mathscr{Y}$, let $\Xi(Y)$ be the set of $\xi \in G(k,p)$ s.t. $Q_Y(\xi)$ is maximal. Define $\Phi(Y)$ to be an arbitrary element of $\Xi(Y)$. Consider the following assumptions:

(i) c is strictly positive.
(ii) c is continuous on $\mathscr{P}_k$.
(iii) If $Z \in \mathscr{P}_k$, y is a $1 \times p$ row vector and M is any $p \times p$ matrix s.t. $ZM \in \mathscr{P}_k$, then

$$c(ZM + y1_n) = c(Z).$$

(iv) s is nonnegative and continuous.
(v) $s(Z) = 0$ if $Z \in \mathscr{D}$.
(vi) If $Y \in \mathscr{P}_k$, then $s[\Pi_\xi(Y)]$ is uniquely maximized by $\xi = \Delta(Y)$.

The idea behind (iii) is that $c(Z)$ is invariant under multivariate affine changes in scale and location. The idea behind (v) is that s is sensitive to k-dimensional spread, for example, $s(Z)$ could be the product of the k largest eigenvalues of the covariance matrix of $Z \in \mathscr{R}$. It follows from Rao [(1973), Section 1f.2(vii), page 64] that principal component plane-fitting (based on the covariance matrix) is a version of Friedman–Tukey projection pursuit with this s and c constant.

It turns out if c and s satisfy assumptions (i)–(vi), then $\mathscr{S}_\Phi \cap \mathscr{P}_k$ is empty. Therefore, by Theorem 2.2, $\dim_H[\mathscr{S}_\Phi(R)] \geq np - 2$ for every $R \in (0, \infty)$. Moreover, Q turns out to be continuous, so, as discussed above, if $Y \in \mathscr{S}_\Phi$, then Q_Y has nonunique global maxima.

Next, consider the projection pursuit method of Friedman (1987). In this procedure, the projection index is not evaluated until after some preprocessing of the data. Among the preprocessing is possible dimension reduction [Friedman (1987), Section 6.2]. If the data lie near a plane of dimension q, then the data are projected onto that plane prior to further analysis. In Example 7.3 of Friedman (1987) it is stated that the plane is to be chosen using principal components. If $q = k$, after the dimension reduction there is nothing more to do (particularly if $k = 1$ or 2). Thus, in practice, near $\mathscr{P}_k$

Friedman's (1987) method is just principal component plane-fitting. Since the hypotheses of Theorem 2.2 depend only on the behavior of the plane-fitter in the vicinity of $\mathscr{P}_k$, it follows that this projection pursuit method has already been treated in Example 2.3. It is not necessary to consider the projection index.

REMARK 2.5 (Equivariance). Let Φ be an equivariant plane-fitter, by which, abusing nomenclature, I mean the following. If $s > 0$, M is an $n \times n$ permutation matrix (i.e., a matrix obtained by permuting the rows of the identity matrix), $P \in O(p)$ [the orthogonal group; Guillemin and Pollack (1974), page 22], y is $p \times 1$ and $Y \in \mathscr{Y}$, then

$$\Phi(sMYP + y1_n) = \Phi(Y)P.$$

Here, if $\xi \in G(k,p)$, $\xi P = \{yP \in \mathbb{R}^p \colon y \in \xi \ (y \text{ is } 1 \times p)\}$.

Equivariance is often a natural property to require of a plane-fitter, for example, principal components analysis is equivariant. Equivariance of Φ may influence $\dim_H(\mathscr{S}_\Phi)$. For example, if $Y_0 \in \mathscr{S}_\Phi$, then so is every point in the orbit of Y_0, which consists of all matrices of the form $sMY_0P + y1_n$, where s, M, P and y are as above. Since there are only a finite number of permutation matrices, varying M contributes nothing to the dimension of the orbit of Y_0, but varying s, P and y might. Now, $\dim_H[O(p)] = \frac{1}{2}p(p-1)$ [Guillemin and Pollack (1974), page 23], so

$$\dim_H(\text{orbit of } Y_0) \le 1 + \tfrac{1}{2}p(p-1) + p. \tag{2.2}$$

Moreover, let $Y \in \mathscr{Y}$ be *symmetric*, that is, for some $P \in O(p) \setminus \{I_p\}$ ("$\setminus$" indicates set theoretic subtraction; $I_p = p \times p$ identity matrix) and permutation matrix M, $YP = MY$. If there exists a sequence $\{Y_m\} \subset \mathscr{Y}$ s.t. $Y_m \to Y$, $\Phi(Y_m) \to \xi \in G(k,p)$, say, and $\xi P \neq \xi$, then $Y + y1_n \in \mathscr{S}_\Phi$ for every $p \times 1$ row matrix, y. It can be shown that

$$\dim_H(\{Y + 1_n y \colon Y \in \mathscr{Y} \text{ is symmetric}, y \text{ is } 1 \times p\})$$
$$\le n(p-1) + \tfrac{1}{2}p(p-1) + p.$$

This bounds above how much of $\dim_H(\mathscr{S}_\Phi)$ can be explained by equivariance. This bound and (2.2) apply a fortiori to plane-fitters having in/equivariance properties weaker than the ones listed above. Thus, if n is large, equivariance cannot fully explain the large size of the singular sets of plane-fitters to which Theorem 2.2 applies.

The first statement in the following is immediate from Theorem 2.2.

THEOREM 2.6 *If $R \in (0,\infty)$, then $\dim_H[\mathscr{S}_\Phi(R)] \ge d$. There is a plane-fitter which, for every $R \in (0,\infty)$, achieves this lower bound.*

REMARK 2.7 (Degeneracy). Under (2.1) every degenerate data set is a singularity. Now, $\dim_H(\mathscr{D}) < d \le \dim_H(\mathscr{S}_\Phi)$. Thus, most singularities of a plane-fitter are nondegenerate.

EXAMPLE 2.8 ([Multivariate] least squares [multiple] regression [Johnson and Wichern (1992), pages 314–316]). Let k, m and n be positive integers with $n > k + m$. Let Φ correspond to least squares regression of m-dimensional responses on k-dimensional predictors based on samples of size n. So $p = k + m$. It turns out that $\mathscr{S}_\Phi$ consists precisely of those nondegenerate data sets which are collinear, that is, their predictor components lie on a plane in $\mathbb{R}^k$ of dimension less than k. From this one can calculate

$$\dim_H(\mathscr{S}_\Phi) = n(p-1) + k \geq d. \tag{2.3}$$

Equality holds in (2.3) precisely when $m = 1$. Much has been written [e.g., Belsley (1991)] about the singularity ("collinearity") problem in the $m = 1$ case. In terms of Hausdorff dimension this problem is as mild as possible. Other plane-fitting methods (Examples 2.3 and 2.4) apparently have worse singularity problems.

3. Implications. This section briefly treats a few issues raised by the results in Section 2.

Theorems 2.2 and 2.6 show that the singular set of a plane-fitter is large. Thus, diagnostics are important, not just in regression [Belsley (1991)], but in plane-fitting in general (especially when Theorem 2.2 applies). By "diagnostic" I mean a statistic indicating how stable the plane-fitter is at the data.

However, what should one do if the data fall near a singularity of one's plane-fitter? One possibility is to use other plane-fitters, whose singularities are not near the data. [This is intermediate between using a single plane-fitter and the "grand tour" of Asimov (1985).] Another is to employ techniques of multivariate data description which do not rely on plane-fitting [see, e.g., Chapter 12 in Johnson and Wichern (1992) and Wegman (1990)].

Theorem 2.6 has interesting implications for situations in which observations are set aside and a plane is fitted to those which remain. One such situation arises from a simple interpretation of resistance in plane-fitting. Informally, Φ is resistant of order r if whenever r or fewer observations lie far from the rest, Φ fits a plane to the rest.

Formally, let λ and σ be measures of location and spread for p-vectors. If m is a positive integer, let $\mathscr{Y}_m$ be the set of $m \times p$ matrices. If $X \in \mathscr{Y}_m$, with rows $y_1, \dots, y_m$, define $\lambda(X) = \lambda(y_1, \dots, y_m)$. Define $\sigma(X)$ similarly.

Let r be a positive integer s.t. $r < \min\{\frac{1}{2}n, n - p\}$. Let $M \in (1, \infty)$, let $s \in [1, r]$ be an integer and let $Z \in \mathscr{Y}_s$ have rows $z_1, \dots, z_s \in \mathbb{R}^p$. Now let $\mathscr{P}_k(Z; M)$ be the set of $Y \in \mathscr{Y}$ s.t.:

1. For some choice of $1 \leq i_1 < \cdots < i_s \leq n$, the i_jth row of Y is z_j.
2. If $X \in \mathscr{Y}_{n-s}$ is obtained by deleting rows $i_1, \dots, i_s$ from Y, then $|z_j - \lambda(X)| > M\sigma(X)$, for $j = 1, \dots, s$.
3. The rows of X lie exactly on a unique k-plane, $\gamma(X)$, in $\mathbb{R}^p$.

Let Φ be a plane-fitter. Say that Φ is a *resistant plane-fitter of order r* if it is defined on a dense subset $\mathscr{Y}'$ of $\mathscr{Y}$ (as usual) and there exists a constant $M \in (1, \infty)$ s.t. for any $s = 1, \dots, r$ and $z_1, \dots, z_s \in \mathbb{R}^p$:

4. For a dense subset $\mathscr{P}'_k(Z)$ of $\mathscr{P}_k(Z; M)$, $Y \in \mathscr{P}'_k(Z)$ implies $Y \in \mathscr{Y}'$ and $\Phi(Y)$ is parallel to $\gamma(X)$ [i.e., if $y \in \gamma(X)$, then $\Phi(Y) = \gamma(X) - y$]. (Here, X and Z are as above.)

The following argument shows that if Φ is a resistant plane-fitter of order r, then $\dim_H(\mathscr{S}_\Phi)$ will be at least $r(p-k)+d$. If r observations lie far from the bulk of the data, then Φ will fit planes to the remaining data. Thus, for each choice of r outliers, we have a plane-fitter on $\mathscr{Y}_{n-r}$. By Theorem 2.6, the dimension of the singular set of this plane-fitter will be at least $(n-r)k + (k+1)(p-k) - 1$. However, the set of r p-dimensional "outliers" has dimension rp. Adding, we have

$$\dim_H(\mathscr{S}_\Phi) \geq rp + (n-r)k + (k+1)(p-k) - 1 = r(p-k) + d. \tag{3.1}$$

[See Ellis (1993b) for a careful proof.]

Consider the multiple regression case (Example 2.8 with $m = 1$). Then $p - k = 1$ and we know that the Hausdorff dimension of the singular set of least squares regression (LSR) is d. By (3.1), the Hausdorff dimension of the singular set of a regression plane-fitter which is resistant of order r is at least r larger than that of LSR.

Hettmansperger and Sheather (1992) show that the least median of squares (LMS) regression has more singularities than does LSR. In fact, LMS regression is resistant in the sense defined above. The order is the largest integer strictly less than $\min\{\frac{1}{2}n, n-p\}$.

For what other formalizations of resistance will one observe an enlargement of the singular set of resistant regression methods?

Observations are also set aside in the jackknife or bootstrap [Efron (1982)]. As in the case of resistant plane-fitting, this can have the net effect of increasing the size of the singular set [Ellis (1993b)]. Might this effect positively bias jackknife or bootstrap estimates of variability of a plane-fitter?

4. Sketches of proofs. Let $m, m_1, m_2 \geq 0$ be integers. If $r > 0$ and $x \in \mathbb{R}^m$, let $B_m(x, r)$ be the open ball in $\mathbb{R}^m$ with center x and radius r. If $A \subset \mathbb{R}^{m_1}$, a map $g: A \to \mathbb{R}^{m_2}$ is *Lipschitz* if there exists $\kappa < \infty$ s.t. $|g(x) - g(y)| \leq \kappa |x - y|$ for every $x, y \in A$. In this case

$$\dim_H[g(A)] \leq \dim_H(A) \tag{4.1}$$

[Falconer (1990), Corollary 2.4, page 30].

For simplicity, take $R = \infty$, so $\mathscr{S}_\Phi(R) = \mathscr{S}_\Phi$. The following result is crucial to proving Theorem 2.2.

Proposition 4.1. *Suppose there is an open neighborhood $\mathscr{U} \subset \mathscr{Y}$ of $\mathscr{P}_k$ s.t. Φ has no singularities in $\mathscr{U}$. Then* $\dim_H(\mathscr{S}_\Phi) \geq np - 2$.

Proof. For a point $u = (s, t)$ on the unit circle S^1, let $U(u)$ be the 2×2

matrix

$$U(u) = \begin{cases} \begin{pmatrix} s & t \\ 0 & 0 \end{pmatrix}, & \text{if } t \geq 0, \\ \begin{pmatrix} s & 0 \\ t & 0 \end{pmatrix}, & \text{if } t < 0. \end{cases}$$

Note that U is a continuous map of S^1 into the space of 2×2 matrices. Let $W(u)$ and V be the $(k+1) \times (k+1)$ matrices

$$W(u) = \begin{pmatrix} I_{k-1} & 0^{(k-1)\times 2} \\ 0^{2\times(k-1)} & U(u) \end{pmatrix}, \qquad V = \begin{pmatrix} I_{k-1} & 0^{(k-1)\times 2} \\ 0^{2\times(k-1)} & \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \end{pmatrix}.$$

(Here, "0" denotes a matrix, of the indicated or implied dimensions, filled with 0's.) If $u \in S^1$, $\lambda \in [0,1]$, $a, b \in \mathbb{R}$, M is a $2 \times (p-2)$ matrix and N is an $(n-2) \times p$ matrix, define an $n \times p$ matrix by

$$\begin{aligned} &Z(u, \lambda, a, b, M, N) \\ &\quad = \lambda \begin{pmatrix} 0 & 0 \\ 0 & W(u) \end{pmatrix} + (1-\lambda) \begin{pmatrix} & N \\ M & \begin{pmatrix} 0 & a \\ -a & b \end{pmatrix} \end{pmatrix} + (1-\lambda) \begin{pmatrix} 0 & 0 \\ 0 & V \end{pmatrix}. \end{aligned}$$

Notice that Z is a continuous map into $\mathscr{Y}$.

Choose $\varepsilon \in (0,1)$ so small that for $u \in S^1$, $\lambda \in [0,1]$, $a, b \in (-\varepsilon, \varepsilon)$, $M \in B_{2(p-2)}(0, \varepsilon)$ and $N \in B_{(n-2)p}(0, \varepsilon)$ we have $Z(u, \lambda, a, b, M, N) \notin \mathscr{D}$. Since $\mathscr{U} \cap \mathscr{S}_\Phi = \varnothing$, by making ε still smaller, if necessary, we can find some λ_0, λ_1 s.t. $0 < \lambda_0 < \lambda_1 < 1$ and

$$\begin{aligned} &Z\big(S^1, [0, \lambda_0) \cup (\lambda_1, 1], (-\varepsilon, \varepsilon), (-\varepsilon, \varepsilon), B_{2(p-2)}(0, \varepsilon), B_{(n-2)p}(0, \varepsilon)\big) \\ &\qquad \cap \mathscr{S}_\Phi = \varnothing. \end{aligned}$$

Let $A = S^1 \times [\lambda_0, \lambda_1] \times (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \times B_{2(p-2)}(0, \varepsilon) \times B_{(n-2)p}(0, \varepsilon)$ and let $\mathscr{F} = Z(A)$. Z^{-1} is defined and Lipschitz on $\mathscr{F}$. Thus, by (4.1), to show that $\dim_H(\mathscr{S}_\Phi) \geq np - 2$, it suffices to show that $\dim_H(Z^{-1}[\mathscr{S}_\Phi \cap \mathscr{F}]) \geq np - 2$.

Let $a, b \in (-\varepsilon, \varepsilon)$, $M \in B_{2(p-2)}(0, \varepsilon)$ and $N \in B_{(n-2)p}(0, \varepsilon)$ and consider the two-dimensional cone

$$\mathscr{C} = \mathscr{C}_{a,b,M,N} = Z\big(S^1, [0,1], a, b, M, N\big) \subset \mathscr{Y}.$$

A topological argument shows $\mathscr{C} \cap \mathscr{F} \cap \mathscr{S}_\Phi \neq \varnothing$. The basic idea is this. φ wraps the base of $\mathscr{C}$ (the points where $\lambda = 1$) around a void in $G(k, p)$. Thus, only by tearing a hole in $\mathscr{C}$ can φ map all of it into $G(k, p)$. Tearing means singularity.

Hence, if Π: $A \to E = (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \times B_{2(p-2)}(0, \varepsilon) \times B_{(n-2)p}(0, \varepsilon)$ is projection, then $\Pi(Z^{-1}[\mathscr{S}_\Phi \cap \mathscr{F}]) = E$ and so, by (4.1),

$$\dim_H\big(Z^{-1}[\mathscr{S}_\Phi \cap \mathscr{F}]\big) \geq \dim_H(E) = np - 2. \qquad \square$$

PROOF OF THEOREM 2.2. $\mathscr{P}_k$ is an imbedded smooth $(d+1)$-dimensional submanifold of $\mathscr{Y}$. Let $\mathbb{N}$ be the normal vector bundle of $\mathscr{P}_k$ in $\mathscr{Y}$. To define it, first define the inner product of $Y_1, Y_2 \in \mathscr{Y}$ to be trace $Y_1^T Y_2$. Then $\mathbb{N} \subset \mathscr{P}_k \times \mathscr{Y}$ consists of all pairs (Z, X), where $Z \in \mathscr{P}_k$ and X is normal to $\mathscr{P}_k$ at Z. Let Π: $\mathbb{N} \to \mathscr{P}_k$ be the projection $\Pi(Z, X) = Z$.

By the tubular neighborhood theorem [Guillemin and Pollack (1974); Lang (1972)], there is a neighborhood $\mathscr{V}$ of $\mathscr{P}_k$ in $\mathscr{Y}$ and a C^∞-diffeomorphism, T, of $\mathscr{V}$ onto $\mathbb{N}$ s.t. if $Z \in \mathscr{P}_k$, then $T(Z)$ is $(Z, 0)$.

If $(Z, X) \in \mathbb{N}$, let $f(Z, X) \in \mathbb{N}$ be defined by

$$f(Z,X) = \begin{cases} (Z,X), & \text{if } |X| \geq 2, \\ (Z,(|X|-1)X), & \text{if } 1 \leq |X| < 2, \\ (Z,0), & \text{if } |X| < 1. \end{cases}$$

Let

$$\Theta(Y) = \begin{cases} \Phi \circ T^{-1} \circ f \circ T(Y), & \text{if } Y \in \mathscr{V}, \\ \Phi(Y), & \text{otherwise} \end{cases}$$

("$\circ$" indicates composition of functions). Θ is a plane-fitter with no singularities in $T^{-1}(\{(Z, X) \in \mathbb{N}: |X| < 1\})$. Therefore, by Proposition 4.1, $\dim_H(\mathscr{S}_\Theta) \geq np - 2$.

Suppose $\dim_H(\mathscr{P}_k \cap \mathscr{S}_\Phi) < d$, but $\dim_H(\mathscr{S}_\Phi) < np - 2$. I show that $\dim_H(\mathscr{S}_\Theta) < np - 2$, a contradiction. On $\mathscr{Y} \setminus \mathscr{V}$, $\Theta = \Phi$, so $\dim_H[(\mathscr{Y} \setminus \mathscr{V}) \cap \mathscr{S}_\Theta] < np - 2$. By (4.1), $\dim_H(\{T^{-1}(Z, X) \in \mathscr{S}_\Theta: (Z, X) \in \mathbb{N}, 1 < |X|\}) < np - 2$.

The interesting and delicate case is $\mathscr{S}_\Theta^1 = \{T^{-1}(Z, X) \in \mathscr{S}_\Theta: (Z, X) \in \mathbb{N}, |X| = 1\}$. Suppose $(Z, X) \in \mathbb{N}$ with $|X| = 1$. It is easy to see that $T^{-1}(Z, X) \in \mathscr{S}_\Theta$ only if $Z \in \mathscr{S}_\Phi \cap \mathscr{P}_k$. Thus, locally, $\mathscr{S}_\Theta^1$ looks like a subset of $[\mathscr{S}_\Phi \cap \mathscr{P}_k] \times S^{q-1}$, where $q = \dim_H(\mathscr{Y}) - \dim_H(\mathscr{P}_k) = np - d - 1$ and $S^{q-1} \subset \mathbb{R}^q$ is the unit sphere. By Corollary 7.4 in Falconer [(1990), page 95] we may add dimensions:

$$\dim_H(\mathscr{S}_\Theta^1) \leq \dim_H[\mathscr{S}_\Phi \cap \mathscr{P}_k] + q - 1 < d + q - 1 = np - 2. \qquad \square$$

PROOF OF THEOREM 2.6. I only need to prove that the lower bound is tight. Let ζ be a fixed element of $G(k, p)$ and Ω be the set of $\xi \in G(k, p)$ s.t. ξ contains a nonzero vector orthogonal to ζ. Ω is closed in $G(k, p)$. There is a neighborhood $\mathscr{U}$ of $\mathscr{P}_k$ in $\mathscr{Y}$ in which the principal components plane-fitter, Ψ (based on the covariance matrix), is defined and continuous. Let $\mathscr{C} = \Delta^{-1}(\Omega) \subset \mathscr{P}_k$, $\tilde{\mathscr{P}} = \mathscr{P}_k \setminus \mathscr{C}$ and $\tilde{\mathscr{Y}} = \mathscr{Y} \setminus (\mathscr{D} \cup \mathscr{C})$. If $\mathscr{V} = \{Y \in \mathscr{U}: \Psi(Y) \notin \Omega\}$, then $\tilde{\mathscr{P}} \subset \mathscr{V} \subset \tilde{\mathscr{Y}}$ and $\tilde{\mathscr{Y}} \setminus \mathscr{V}$ has an open neighborhood, $\mathscr{W} \subset \tilde{\mathscr{Y}}$, whose closure (in $\tilde{\mathscr{Y}}$), $\overline{\mathscr{W}}$, is disjoint from $\tilde{\mathscr{P}}$. Let μ: $\tilde{\mathscr{Y}} \to [0, 1]$ be a continuous function which is 1 on $\overline{\mathscr{W}}$ and 0 on $\tilde{\mathscr{P}}$.

Let Π be orthogonal projection onto ζ and, for $t \in [0, 1]$ and $\xi \in G(k, p)$, let

$$g(t, \xi) = \{(1-t)x + t\Pi(x) \in \mathbb{R}^p: x \in \xi\}.$$

Then g is a continuous map of $[0, 1] \times \Omega^c$ into $G(k, p)$.

Define a plane-fitter Φ on $\mathscr{Y}$ as follows:

$$\Phi(Y) = \begin{cases} g[\mu(Y), \Psi(Y)], & \text{if } Y \in \mathscr{Y}\setminus\overline{\mathscr{W}}, \\ \Delta(Y), & \text{if } Y \in \mathscr{Q}, \\ \zeta, & \text{otherwise.} \end{cases}$$

Then $\mathscr{S}_\Phi \subset \mathscr{Q}$, but $\dim_H(\mathscr{Q}) \le d$. □

Acknowledgments. Fred Cohen suggested using two-dimensional cones. The presentation benefitted from comments from the Editors, and Example 2.4 benefitted from comments from a referee.

REFERENCES

ASIMOV, D. (1985). The grand tour. *SIAM J. Sci. Statist. Comput.* **6** 128–143.

BELSLEY, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression.* Wiley, New York.

BOOTHBY, W. M. (1975). *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Academic Press, New York.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

ELLIS, S. P. (1991a). The singularities of fitting planes to data. *Ann. Statist.* **19** 1661–1666.

ELLIS, S. P. (1991b). Topological aspects of the location problem for directional and axial data. *Internat. Statist. Rev.* **59** 389–394.

ELLIS, S. P. (1993a). On the size of the singular sets of plane-fitters. Unpublished manuscript.

ELLIS, S. P. (1993b). On the dimension of the singular sets of plane-fitters. Unpublished manuscript.

FALCONER, K. (1990). *Fractal Geometry: Mathematical Foundations and Applications.* Wiley, New York.

FRIEDMAN, J. H. (1987). Exploratory projection pursuit. *J. Amer. Statist. Assoc.* **82** 249–266.

FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–890.

GUILLEMIN, V. and POLLACK, A. (1974). *Differential Topology.* Prentice-Hall, Englewood Cliffs, NJ.

HETTMANSPERGER, T. P. and SHEATHER, S. J. (1992). A cautionary note on the method of least median squares. *Amer. Statist.* **46** 79–83.

HUBER, P. J. (1985). Projection pursuit. *Ann. Statist.* **13** 435–475.

JOHNSON, R. A. and WICHERN, D. W. (1992). *Applied Multivariate Statistical Analysis,* 3rd ed. Prentice-Hall, Englewood Cliffs, NJ.

LANG, S. (1972). *Differential Manifolds.* Addison-Wesley, Reading, MA.

MORGAN, F. (1988). *Geometric Measure Theory: A Beginner's Guide.* Academic Press, San Diego.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications,* 2nd ed. Wiley, New York.

WEGMAN, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. *J. Amer. Statist. Assoc.* **85** 664–675.

NEUROBEHAVIORAL UNIT (127a)
V. A. MEDICAL CENTER
385 TREMONT AVENUE
EAST ORANGE, NEW JERSEY 07018