THE SINGULARITIES OF FITTING PLANES TO DATA

By Steven P. Ellis

University of Rochester

This paper is dedicated to the memory of the late Charlie Odoroff
—colleague and ponderer of point clouds.

Plane-fitting, for example, linear regression, principal components or projection pursuit, is treated from a general perspective. It is shown that any method of plane-fitting satisfying very mild hypotheses must have singularities, that is, data sets near which the procedure is unstable. The well-known collinearity phenomenon in least squares regression is a special case. Severity of singularities is also discussed.

The results, which are applications of algebraic topology, may be viewed as putting limits on how much can be done through robustification to stabilize plane-fitting.

1. Introduction. Let a topological space $\mathbf{X}$ be the sample space. Let δ be a data analytic procedure on $\mathbf{X}$ taking values in a topological space $\mathbf{A}$. Now, δ may not be defined everywhere in $\mathbf{X}$, so let $\mathbf{X}' \subset \mathbf{X}$ be the set of points at which δ is defined. If $x_0 \in \mathbf{X}$ is a limit point of $\mathbf{X}'$, we call x_0 a *singularity* of δ if $\lim_{x \to x_0} \delta(x)$ does not exist. (Multi)collinear data sets are singularities of least squares regression [Belsley, Kuh and Welsch (1980), Section 3.1]. The breakdown point [Donoho and Huber (1983)] of an estimator of location is apparently related to its singularities at infinity.

Singularities of data analytic procedures are important because it is desirable that a data summary or description not vary wildly when the data are perturbed only a little. So if a procedure has singularities, it seems desirable to remove them. For example, if $n > 2$ is the sample size, say that a map λ: $\mathbb{R}^n \to \mathbb{R}$ ($\mathbb{R}$ = reals) is a measure of location if it satisfies the constraint: $\lambda(x, \ldots, x) = x$ for every $x \in \mathbb{R}$. Suppose λ is the "shorth" [Andrews, Bickel, Hampel, Huber, Rogers and Turkey (1972)], that is, it assigns to a data set the mean of the shortest half of the data. One easily sees that λ has singularities.

However, by robustifying λ, the singularities can be removed. Let $S \subset \mathbb{R}^n$ be the set of singularities of λ and let $U \subset \mathbb{R}^n$ be a tight neighborhood about S. If $\mathbf{x} \in S$, let $\mu = \text{median}(\mathbf{x})$; if $\mathbf{x} \in \mathbb{R}^n \setminus U$ ($\setminus$ indicates set-theoretic subtraction), let $\mu(\mathbf{x}) = \lambda(\mathbf{x})$; and if $\mathbf{x} \in U \setminus S$, let $\mu(\mathbf{x})$ be an appropriate convex combination of the median and shorth of $\mathbf{x}$. Then at most data sets $\mu = \lambda$, but the robustified shorth μ has no singularities.

In this paper, the singularities of plane-fitting techniques like linear regression, principal components or projection pursuit [Huber (1985)], are investigated from a general perspective. A plane-fitter is defined generically to be a map satisfying a very mild constraint. We show that, for topological reasons, plane-fitters, unlike measures of location, have singularities which cannot be robustified away.

We also discuss the severity of singularities of plane-fitters. If x_0 is a singularity of δ, then by definition the images under δ of neighborhoods of x_0 cannot be arbitrarily small. The larger these images are, the more unstable δ is near x_0 and, we say, the more severe is the singularity at x_0. We show that, apparently, a severe singularity of a plane-fitter cannot be replaced by a mild one without in general creating new singularities.

The proofs, which rely on algebraic topology, are sketched in Section 3.

2. Results. Let $n > p > k > 0$ be fixed integers. A data set will be 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.) By a plane we will mean a linear manifold, that is, an affine subspace of $\mathbb{R}^p$.

By fitting a plane we will mean assigning to $Y \in \mathscr{Y}$ a k-dimensional plane which reflects the linear structure of Y. Choose a particular plane-fitting technique δ. For simplicity, we confine attention to the map Φ defined as follows. If $Y \in \mathscr{Y}$ and the k-plane $\delta(Y)$ is defined, let $\Phi(Y)$ be the k-dimensional subspace (i.e., plane through the origin) of $\mathbb{R}^p$ parallel to $\delta(Y)$. From now on, assume that Φ is defined on a dense subset $\mathscr{Y}'$ of $\mathscr{Y}$. The range of Φ is the Grassman manifold Γ consisting of all k-planes in $\mathbb{R}^p$ passing through the origin [Boothby (1975), page 63].

A map $\Phi: \mathscr{Y}' \to \Gamma$ will be a plane-fitter if it satisfies a certain constraint which we now develop. If $Y \in \mathscr{Y}$, it rows, $y_1, \ldots, y_n$, determine the plane,

$$\left\{ \sum_1^n a_i y_i : \sum_1^n a_i = 1 \right\}.$$

Let $\Delta(Y) \in \Gamma$ denote the subspace of $\mathbb{R}^p$ parallel to this plane. Let $\mathscr{P}_k = \{Y \in \mathscr{Y} : \dim \Delta(Y) = k\}$. It is natural to require a plane-fitter Φ to assign to each data set lying on a unique k-plane, the plane through the origin parallel to the one on which the data lie. That is:

(2.1)   If $Y \in \mathscr{P}_k$, then $\Phi(Y)$ is defined and $\Phi(Y) = \Delta(Y)$.

For example, $\Phi = $ *least squares regression* can be extended in a natural way to satisfy (2.1).

Now, (2.1) turns out not to be a good choice for the defining constraint of plane-fitting. Say that $Y \in \mathscr{Y}$ is degenerate and write $Y \in \mathscr{D}$ if $\dim \Delta(Y) < k$. Then $\mathscr{D} \subset \overline{\mathscr{P}_k}$. (Here the overbar indicates the closure of a set.) A little reflection shows that if Φ satisfies (2.1), every degenerate data set is a singularity of Φ. It is tempting to robustify Φ in the vicinity of $\mathscr{D}$ in order to

free it of degenerate singularities. However, the modified Φ would no longer satisfy (2.1).

To make room for such robustified plane-fitters, we must weaken (2.1) by permitting $\mathscr{P}_k$ to be replaced there by smaller classes of data sets. If $\mathscr{P} \subset \mathscr{P}_k$ is to be used in place of $\mathscr{P}_k$, it must be rich enough to lead to an interesting definition of plane-fitting. To insure this, the main requirement we make is that it be possible to choose, in a continuous manner, a set of representatives in $\mathscr{P}$ for the planes in Γ.

Say that a map $R: \Gamma \to \mathscr{P}_k$ is a RID (right inverse of Δ) if it is continuous and for every $\xi \in \Gamma$ the k-plane determined by the rows of $R(\xi)$ is parallel to ξ, formally $\Delta[R(\xi)] = \xi$. An example of a RID is the map R_0 defined as follows. Let $e_1, \ldots, e_p$ be a fixed basis of $\mathbb{R}^p$. If $\xi \in \Gamma$, let $R_0(\xi) \in \mathscr{P}_k$ be the matrix whose ith row is the projection of e_i onto ξ, $1 \le i \le p$, and whose remaining rows are zero. (Recall $n > p$.)

We weaken (2.1) as follows.

There exists $\mathscr{P} \subset \mathscr{Y}$ satisfying:

(2.2a) $\mathscr{P}$ is an open subset of $\mathscr{P}_k$,

(2.2b)   there exists a RID, R, such that $R(\Gamma) \subset \mathscr{P}$,

(2.2c)   if $Y \in \mathscr{P}$, then $\Phi(Y)$ is defined and $\Phi(Y) = \Delta(Y)$.

It is easy to see that a Φ with no degenerate singularities can satisfy (2.2) providing $\mathscr{D} \cap \overline{\mathscr{P}} = \varnothing$. Here is an example of such a $\mathscr{P}$.

2.1 EXAMPLE [Example 1.2, page 11, Ellis (1989)]. Let $\alpha, \varepsilon \in (0, 1)$ and suppose $k > 1$. Define $\mathscr{P}$ to be the collection of $Y \in \mathscr{P}_k$ having the following property. Delete any collection of up to $n\varepsilon$ rows of Y. If $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_k > 0$ are the first k eigenvalues of the sample covariance matrix of the remaining rows, then $\lambda_k/\lambda_1 > \alpha$ and $\lambda_1 > \alpha/(1 - \alpha)$. Thus, $\mathscr{D} \cap \overline{\mathscr{P}} = \varnothing$ in a robust way. Providing $\alpha < 1 - p\varepsilon(1 - \varepsilon)^{-1}$, $\mathscr{P}$ satisfies (2.2a, b), at least for large n.

With this preparation we state our first result.

2.2 THEOREM.   *Suppose Φ satisfies (2.2). Then Φ has at least one singularity. If $n > 2k$, Φ has a nondegenerate singularity.*

Multicollinearity in least squares regression is the best known example of the phenomenon described in the theorem.

The result shows that robustifying Φ around $\mathscr{D}$ is not enough to rid it of singularities. If all the degenerate singularities are smoothed out, Φ will then have nondegenerate singularities, which will exist in any case if n is large enough. Since it seems unreasonable to regard Φ as a plane-fitter if it does not satisfy the mild constraint (2.2), we conclude that plane-fitters can never be completely stabilized.

We next turn to the issue of the severity of singularities.

2.3 EXAMPLE. *Simple L^1 regression.* Let r be a positive integer and take $n = 4r + 4$, $p = 2$ and $k = 1$. Providing it exists uniquely, let $\hat{\beta}(Y)$ be the slope of the L^1-best fitting line for regressing the second column of $Y \in \mathscr{Y}$ on the first. Let $\Phi(Y)$ be the line $\{(x, \hat{\beta}(Y)x), x \in \mathbb{R}\} \in \Gamma$.

Let $0 = x_0 < \cdots < x_r$ and let $Y_0 \in \mathscr{Y}$ be the matrix whose rows are $(-x_r, 0), \ldots, (-x_0, 0)$, $(x_0, 0), \ldots, (x_r, 0)$, $(-x_r, -x_r), \ldots,$ $(-x_0, -x_0), (x_0, x_0), \ldots, (x_r, x_r)$. The rows of Y lie on two lines which cross at the origin. Now let $\varepsilon > 0$, let $\beta_1 \in [0, 1]$ and consider the data set Y_1 which is like Y_0 except the two rows $(-x_0, -x_0) = (x_0, x_0) = (0, 0)$ are replaced by $(-\varepsilon, -\beta_1\varepsilon)$, $(\varepsilon, \beta_1\varepsilon)$. Then $\hat{\beta}(Y_1) = \beta_1$. So if Y lies in a small neighborhood $\mathscr{U}$ about Y_0, by perturbing Y a small amount, one can obtain a data set whose fitted slope can be any value in $[0, 1]$. Thus, Y_0 is a singularity of Φ. However, while $\hat{\beta}(\mathscr{U} \cap \mathscr{Y}')$ contains $[0, 1]$, it is bounded (providing $\mathscr{U}$ is small enough), so the instability of $\hat{\beta}$ near Y_0 has bounded amplitude. Hence, $\Phi(\mathscr{U} \cap \mathscr{Y}')$ is a proper subset of Γ.

This has an inferential interpretation. Based on Y_0, a likelihood ratio confidence interval for the slope in a regression model with i.i.d. double exponential errors will contain $[0, 1]$ for any positive confidence coefficient. However, the confidence interval will be bounded and hence will provide some information. At any data set near Y_0, one would observe similar behavior.

2.4 EXAMPLE. *Principal components.* If $Y \in \mathscr{Y}$, let $\Phi(Y) \in \Gamma$ be the plane spanned by the first k eigenvectors of the sample covariance matrix of Y, providing it is well-defined. Let $Y_0 \in \mathscr{Y}$ have mean-centered, orthonormal columns. Then arbitrarily close to Y_0, one can find $Y \in \mathscr{Y}$ s.t. $\Phi(Y)$ is any k-plane through the origin. That is, Φ maps every neighborhood of the singularity Y_0 *onto* Γ.

As in the last example, this has inferential implications. Consider the model in which the data come from a multivariate normal population. Assume the plane ξ, spanned by the first k eigenvectors of the population covariance matrix is well-defined, though unknown. Then for any positive confidence coefficient, the likelihood ratio confidence set for ξ based on this model and computed at Y_0 will equal *all of* Γ. A confidence set for ξ at a data set near Y_0 would also be huge.

Call a singularity Y_0 mild if, as in example 2.3, it has a neighborhood $\mathscr{U}$ s.t. $\Phi(\mathscr{U} \cap \mathscr{Y}') \neq \Gamma$. Call Y_0 severe if, as in Example 2.4, for every neighborhood $\mathscr{U}$ of Y_0, $\Phi(\mathscr{U} \cap \mathscr{Y}') = \Gamma$. In a qualitative sense, the size of the oscillations (and hence instability) of Φ is greater near a severe singularity than near a mild one.

Having seen that robustification cannot purge Φ of singularities, there is at least the hope that Φ can be stabilized somewhat by replacing severe singularities by mild ones. However, the next theorem and example suggest that, in general, this can be done only at the cost of creating new singularities. If R is

a RID and $Y_0 \in \mathcal{Y}$, define $\mathscr{C}_R(Y_0)$ to be the cone

$$\mathscr{C}_R(Y_0) = \{(1 - t)Y_0 + tY_1 : 0 \le t \le 1, Y_1 \in R(\Gamma)\}.$$

2.5 THEOREM. *Suppose Φ satisfies (2.2) and $Y_0 \in \mathcal{Y} \setminus \mathscr{P}$ is not a severe singularity of Φ. Then for any RID $R: \Gamma \to \mathscr{P}$, Φ has a singularity in $\mathscr{C}_R(Y_0) \setminus \{Y_0\}$.*

2.6 EXAMPLE (Example 2.4, continued.). Let Y_0 be as before and take $\mathscr{P} = \mathscr{P}_k$. Suppose $n > 3p$ and the first $2p$ rows of Y_0 are 0. Let R_0 be the RID described just before (2.2). By changing rows $p + 1$ through $2p$ in R_0, one can obtain a RID R whose last $n - 2p$ rows are 0 and s.t. the columns of $R(\xi)$ are mean-centered for all ξ. It is easy to see that Φ is defined and continuous everywhere on $\mathscr{C}_R(Y_0) \setminus \{Y_0\}$.

3. Sketches of proofs. We prove Theorem 2.5 first. A good reference on algebraic topology is Munkres (1984).

PROOF OF THEOREM 2.5. Suppose there is a neighborhood $\mathcal{U}$ of Y_0 and a RID $R: \Gamma \to \mathscr{P}$ s.t. $\Phi(\mathcal{U} \cap \mathcal{Y}') \ne \Gamma$ and Φ has no singularities in $\mathscr{C}_R(Y_0) \setminus \{Y_0\}$. Let the cone on Γ, $C(\Gamma)$, be obtained from $\Gamma \times [0, 1]$ by identifying $\Gamma \times \{0\}$ to a point. Let $\pi: \Gamma \times [0, 1] \to C(\Gamma)$ be the natural projection. Define $Z: C(\Gamma) \to \mathscr{C}_R(Y_0)$ by $Z \circ \pi(\xi, s) = (1 - s)Y_0 + sR(\xi)$, $(\xi, s) \in \Gamma \times [0, 1]$. Z is well-defined and continuous.

For $t \in (0, 1]$, let $A_t = \pi(\Gamma \times (0, t])$. Pick $\varepsilon \in (0, 1)$ s.t. $Z(A_\varepsilon) \subset \mathcal{U}$. Let $\mathscr{S} \subset \mathcal{Y}$ be the set of singularities of Φ, let $\psi(Y) = \lim_{Y' \to Y} \Phi(Y')$ ($Y \in \mathcal{Y} \setminus \mathscr{S}$) and if $a \in A_1$, let $\varphi(a) = \psi \circ Z(a) \in \Gamma$. Then φ is continuous on A_1, $\varphi(A_\varepsilon) \ne \Gamma$ and, by (2.2), $\varphi \circ \pi(\xi, 1) = \xi$ for $\xi \in \Gamma$. We will show that this is impossible.

Consider the following commutative diagram.

$$
\begin{array}{ccccc}
A_\varepsilon & \xrightarrow{\;j\;} & A_1 & \xleftarrow{\;i\;} & \pi(\Gamma \times \{1\}) \\
\varphi' \downarrow & \searrow^{\hat{\varphi}} & \downarrow \varphi & & \\
\Gamma' & \xrightarrow[\;\eta\;]{} & \Gamma & &
\end{array}
$$

Here i, j and η are inclusions (Γ' is defined later). Since $\pi(\Gamma \times \{1\})$ is a deformation retract of A_1 and $\varphi \circ i: \pi(\Gamma \times \{1\}) \to \Gamma$ is a homeomorphism, φ_* is an isomorphism. Similarly, A_ε is a deformation retract of A_1 so $\hat{\varphi}_* = \varphi_* \circ j_*$ is an isomorphism.

By hypothesis, there exists $\xi_0 \in \Gamma \setminus \varphi(A_\varepsilon)$. Let $\Gamma' = \Gamma \setminus \{\xi_0\}$ and let $\varphi': A_\varepsilon \to \Gamma'$ denote $\hat{\varphi}$ regarded as a map into Γ'. Since $\hat{\varphi}_*$ is an isomorphism, $\varphi_{*}' \circ (\hat{\varphi}_*)^{-1}$ is injective. But Γ is a compact connected manifold without boundary. Hence, by duality, $H_d(\Gamma; \mathbb{Z}/2)$ is nontrivial ($d = \dim \Gamma$, $\mathbb{Z} =$ integers) while $H_d(\Gamma'; \mathbb{Z}/2)$ is trivial, which is a contradiction.

PROOF OF THEOREM 2.2. The result is immediate from Theorem 2.5 and the following with $m = k$. $\square$

3.1 Lemma [Lemma 4.1, page 34, Ellis (1989)]. *There exists an infinitely differentiable RID R: $\Gamma \to \mathscr{P}$. For any such R, if $m \leq p$ and $np > (n + p - m)(m - 1) + k(p - k) + p + 1$, the set*

$$\{Y_0 \in \mathscr{Y}: \text{there exists } Y \in \mathscr{C}_R(Y_0) \setminus R(\Gamma) \text{ such that } \dim Y < m\}$$

has Lebesgue measure 0.

REFERENCES

ANDREWS D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. and TUKEY, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton Univ. Press.

BELSLEY, D. A., KUH, E. and WELSCH, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Wiley, New York.

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

DONOHO, D. L. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday* (P. J. Bickel, K. A. Doksum and J. L. Hodges, Jr., eds.) 157–184. Wadsworth, Belmont, Calif.

ELLIS, S. P. (1989). On the singularities of fitting planes to data. Technical Report 89/01, Depts. Statistics and Biostatistics, Univ. Rochester.

HUBER P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435–525.

MUNKRES, J. R. (1984). *Elements of Algebraic Topology.* Benjamin/Cummings, Menlo Park, Calif.

DEPARTMENT OF STATISTICS
UNIVERSITY OF ROCHESTER
ROCHESTER, NEW YORK 14627