Matthews (1985) has carried out many further computations in his Stanford Ph.D. thesis. He works in the context of a random walk on a group and derives the distribution until the walk hits (or is suitably close to) every point. This relates to projection pursuit via Asimov's scheme for the "grand tour." Asimov considers projections that "wiggle around" by small random rotations. His results agree with those reported by Huber in the following sense, it takes a long time to get close to most projections in high dimensions. Therefore, some form of projection pursuit is needed. On the other hand, once an interesting projection has been located, it seems useful to have some kind of grand tour to "wiggle around" in a neighborhood, to try to explore the features of the interesting projection.

REFERENCES

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

DIACONIS, P. (1983). Projection pursuit for discrete data. Stanford University Technical Report #198.

DIACONIS, P. and FREEDMAN, D. (1984). Some asymptotics for graphical projection pursuit. *Ann. Statist.* **12** 793–815.

DIACONIS, P. and GRAHAM, R. L. (1983). The Radon transform on Z_2^k. Stanford University Technical Report #206. To appear in *Pacific J. Math.*

DIACONIS, P. and SHAHSHAHANI, M. (1984). Some theory for projection pursuit regression. *SIAM J. Sci. Statist. Comput.* **5** 175–191.

MATTHEWS, P. (1985). Covering problems for random walks on spheres and finite groups. Doctoral dissertation, Dept. of Statist., Stanford University.

SUDAKOV, V. N. (1980). Typical distributions of linear functionals chosen at random. Colloques Internationaux Du Centre National de la Recherche Scientifique No. 307, pp. 501–511.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

DAVID DONOHO, IAIN JOHNSTONE, PETER ROUSSEEUW, AND WERNER STAHEL

University of California at Berkeley, Stanford University, Technische Hogeschool Delft, and ETH Zürich

In our discussion of this very stimulating paper, we will mostly confine our remarks to some of the general issues Huber raises in the introductory paragraphs.

1. The curse of dimensionality. In paragraph four of the introduction, Huber writes "··· the most exciting feature of PP is that it is one of the very few multivariate methods able to bypass the 'curse of dimensionality' ··· "

Actually, Huber gives no precise definition of the "curse." Perhaps this is best, because there are several curses of dimensionality. Adverse effects of increasing dimension can include: less robustness, greater computational costs, worse mean squared error, and slower convergence to limiting distributions.

In this instance, Huber is concerned with the effects of increasing dimension on the mean-squared-error of smoothers. He points out that kernel and related methods have problems with bias in high dimensions (because they must average over large neighborhoods to keep the variance small). He claims that PP smoothers *do* avoid these problems as long as the "structure" they are to recover is not "highly nonlinear"—without being precise about the meaning of these terms.

Huber's statements can be formulated mathematically. Donoho and Johnstone (1985a) consider what might be called "Projection Pursuit Regression Approximation" (in analogy with what Huber calls PPDA). They approximate a function $f(x_1, x_2)$ of two variables by a sum of n ridge functions, each a polynomial of degree m, and evaluate the quality of the approximation by an L^2-norm with respect to standard bivariate Gaussian measure. They show that for radial functions ($f(x_1, x_2) = g(x_1^2 + x_2^2)$) with p derivatives in L^2, an integrated squared error asymptotic to $N^{-p/1.5}$ is possible, where $N = nm$ is the number of parameters in the approximating sum of ridge functions. This rate is an improvement over the rate $N^{-p/2}$ one would achieve by ordinary methods of approximation (such as bivariate polynomials) for a p-times differentiable function of two variables. However, in the case where f is harmonic ($\Delta f = 0$) but grows so rapidly at ∞ that it has only p derivatives in L^2, the best rate achievable by PPRA is $N^{-p/2}$; this is no better than the rate that bivariate polynomials would achieve.

Using Huber's terminology, radial functions possess the kind of "nearly linear" structure that allows PPRA to work better than usual methods. On the other hand, harmonic functions possess structure nonlinear enough to make PPRA work only at the usual rate.

2. Linear regression. Huber mentions in paragraphs 6 and 7 that a wide variety of multivariate procedures and their robust analogs may be viewed as projection pursuit methods. We would like to point out that ordinary linear regression may also be brought under the PP umbrella.

If $\{(X_i, Y_i), i = 1, \cdots, n\}$ is a standard regression dataset, X_i in $\mathbb{R}^d$, Y_i in $\mathbb{R}$, and Q is a scale-equivariant but *not* location-invariant functional, then the objective

$$Q(\{Y_i - a^t X_i\})$$

defines a measure of interestingness on projections of X—namely, minimizing Q selects a projection $\{a^t X_i\}$ which agrees with the $\{Y_i\}$ closely, so that the residuals $\{r_i = Y_i - a^t X_i\}$ have small scale. Note that if $Q(\{r_i\}) = \sqrt{\sum r_i^2/n}$ then this "most interesting" a is simply the vector of least-squares regression coefficients. For a general Q, the a minimizing Q (when well-defined) is equivariant with respect to the full "regression group": that is, it transforms in the obvious way with respect both to transformations

$$Y_i \rightarrow (Y_i - a^t X_i)/s$$

where $s \neq 0$; *and* with respect to

$$X_i \rightarrow UX_i$$

where U is any nonsingular linear transformation. So each nice Q defines a type of regression estimator; by varying Q one obtains an entire class of regression estimators based on PP.

Work by Rousseeuw (1984) (based on an earlier idea of Hampel (1975)) and by Rousseeuw and Yohai (1985) has shown that it is possible to construct regression estimators far more robust than traditional robust regression methods in the sense of high breakdown point. These methods can withstand contamination of up to 50% of the observations (contamination allowed in *both* the Y- and X-components of an observation) before breaking down.

With hindsight we recognize these estimators as members of the "PP regression" class introduced above. In each case the Q functional is a robust, high-/breakdown measure of scale. Hampel and Rousseeuw use essentially $Q =$ median absolute value; while Rousseeuw and Yohai use a robust M-estimate of scale to get better efficiency in the case of Gaussian errors.

Thus the projection pursuit formalism extends to cover the linear regression problem, and to encompass both the classical least-squares procedure and modern high-breakdown methods. The notion introduced here is best thought of as "PP Linear Regression" to distinguish it from Friedman and Stuetzle's (nonlinear, nonrobust) "PP Regression."

3. Breakdown and projection pursuit. Huber points out in paragraph 8 of his introduction that "... the only known affine equivariant estimators of multivariate location and scatter with high breakdown point ... are based on PP ideas." The only published high-breakdown regression-equivariant estimators are, in the sense of our last remark, also projection pursuit methods (the usual robust regression estimators described in, e.g., Huber (1981) or Krasker-Welsch (1982) do not have a high-breakdown point, nor are they projection pursuit methods).

There is a reason for this apparent relation between high breakdown and projection pursuit. Without getting into details, Donoho, Rousseeuw, and Stahel have found that, for affine-equivariant location estimators and for regression-equivariant regression estimators, breakdown properties are determined by behavior near point configurations of "exact fit": in location, these are situations where most of the data lie in a proper subspace; in regression, where most of the data lie exactly in a regression plane. Note that such configurations are precisely those having a projection in which most of the data collapse to a point.

In other words, high breakdown depends on an estimator's behavior in those situations when certain special kinds of projections occur. Since PP can in principle be used to search for such projections and act appropriately if they occur, the usefulness of PP in synthesizing high-breakdown procedures is not surprising. Notice, however, that PP is not the only way to obtain high-breakdown, equivariant estimators. In multivariate location, Rousseeuw (1983) gives the example of the minimal volume ellipsoid containing at least half the data. Also, not every PP-based estimator which is affine equivariant is going to have high breakdown—this is a corollary of Fill and Johnstone (1984). Breakdown can be characterized by properties of projections; but it is neither necessary nor sufficient to use PP to obtain these properties.

4. On convergence of PPR. In Section 12, Huber establishes a form of weak convergence of the basic PPR algorithm, leaving strong convergence in

$L^2(P)$ an open question. Donoho and Johnstone (1985b) have established such strong convergence in two special cases—where P is either standard Gaussian measure on $\mathbb{R}^d$ or uniform measure on the unit ball in $\mathbb{R}^d$. In these cases $L^2(P)$ may be decomposed as a direct sum of subspaces of polynomials, one for each degree, these subspaces being invariant under projection. The greedy algorithm converges exponentially on any finite sum of these subspaces, and its convergence on all of L^2 (but without any rate) can be derived by approximation. Unfortunately, they see no obvious way to make this argument work for general P.

A statement they can make for general P is: if finite sums of ridge functions are dense in $L^2(P)$, then the greedy algorithm converges *weakly* in $L^2(P)$. Consider the inner product of the mth residual function with the ridge function $g(a^tX)$

$$\begin{aligned} E\{r_m(X)g(a^tX)\}^2 &= E\{E\{r_m(X) \mid a^tX\}g(a^tX)\}^2 \\ &\le \{\sup_a E\{E\{r_m(X) \mid a^tX\}^2\}\}E\{g^2(a^tX)\}. \end{aligned}$$

Using Huber's remark on the maximum marginal norm, one sees that the last term converges to zero as m increases. By linearity, the inner product of r_m with any finite sum of ridge functions tends to zero; if such finite sums are dense in $L^2(P)$, we conclude that $E\{r_m(X)g(X)\}$ tends to zero for every g in $L^2(P)$. In short, when *every* f can be approximated by sums of ridge functions, so that PPR makes sense, the greedy algorithm converges weakly.

For a general strong convergence result, consider the greedy algorithm *with backfitting*. Say that P has the *ridge closure property* if, for every set of n distinct directions, the sums of ridge functions in those directions form a closed subspace of $L^2(P)$. When P has this property, backfitting makes sense, and converges to the projection of f onto that closed subspace. It follows that an approximation to f based on a sum of n ridge functions, f_n, if polished with full backfitting, has a norm no larger than f (since projections have norm 1). On the other hand, the greedy algorithm with backfitting converges weakly if sums of ridge functions are dense. In a Hilbert space, these two facts, $\|f_n\| \le \|f\|$ and $f_n \to f$ weakly, imply strong convergence (just use the Parallelogram Law). Informally, if P is such that approximation by sums of ridge functions makes sense and also backfitting makes sense, then the greedy algorithm with backfitting converges strongly.

It would be interesting to know which P have the required properties. Certainly the Gaussian and the Uniform do. We know of no nice examples (P absolutely continuous) where sums of ridge functions either fail to be dense or fail to be closed.

In the case where P is Gaussian, Donoho and Johnstone have found that the greedy algorithm, although it converges strongly, can do so very slowly, for example on harmonic polynomials. Those examples and the current remarks suggest that backfitting is an important refinement of the basic PPR algorithm.

REFERENCES

DONOHO, D. and JOHNSTONE, I. M. (1985a) When does Projection Pursuit Regression approximate better than Kernel Regression? Preprint.

DONOHO, D. and JOHNSTONE, I. M. (1985b). On convergence of projection pursuit algorithms. Manuscript in preparation.

FILL, J. A. and JOHNSTONE, I. M. (1984). On projection pursuit measures of multivariate location and dispersion. *Ann. Statist.* **12** 127–141.

HAMPEL, F. R. (1975). Beyond location parameters: robust concepts and methods. *Bull. Internat. Statist. Inst.* **46,** 375–382.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

KRASKER, W. and WELSCH, R. E. (1982). Efficient bounded-influence regression estimation. *J. Amer. Statist. Assoc.* **77** 595–604.

ROUSSEEUW, P. (1983). Multivariate estimation with high breakdown point. To appear in *Fourth Pannonian Symp. Math. Statist.*

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

ROUSSEEUW, P. and YOHAI, V. (1985). Robust regression by means of S-estimators. To appear in *Proc. Heidelberg workshop on Robust and Non-linear methods in Time Series Analysis.* Springer, Berlin.

D. DONOHO
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA AT BERKELEY
BERKELEY, CALIFORNIA 94720

I. JOHNSTONE
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

P. ROUSSEEUW
TECHNISCHE HOGESCHOOL DELFT
DEPARTMENT OF MATHEMATICS
AND INFORMATICS
JULIANALAAN 132
2628 BL DELFT
THE NETHERLANDS

W. STAHEL
SEMINAR FÜR STATISTIK
SOL, ETH
8092 ZÜRICH
SWITZERLAND

R. GNANADESIKAN AND J. R. KETTENRING

Bell Communications Research

With the recent flurry of interest in Projection Pursuit (PP), it is both timely and useful to have this paper by Professor Huber which attempts to unify these developments and more classical multivariate dimensionality reduction techniques.

We will limit our comments mostly to PP for finding clusters. The focus will be on the significant gap between the general ideas and existing theory about PP and what is important to and needed by practitioners.

The emphasis in PP on linear combinations and, in this paper, on affine invariance is in conflict with the frequent need to identify subsets of the given variables which contain the cluster structure. PP may help, but it is far from the end of the line as far as reduction of dimensionality and interpretation are concerned. Furthermore, the use of such invariant procedures can be misleading. For instance, the Class III type (squared) distance function, $[a'(x_i - x_j)]^2/a'Sa$, which utilizes an overall covariance matrix S, whether determined robustly or not, may be quite inappropriate for capturing cluster structure because S says nothing about within-cluster variability. Similar comments would seem to apply to Huber's invariant version of the Friedman-Tukey index.