KAGAN, A. M., LINNIK, Y. V. and RAO, C. R. (1973). *Characterization Problems in Mathematical Statistics*. Wiley, New York.

KEMPERMAN, J. H. B. (1969). On the optimum rate of transmitting information. *Lecture Notes in Math.* **89** pp. 126–169, Springer-Verlag, Berlin.

KLEINER, B., MARTIN, R. D. and THOMSON, D. J. (1979). Robust estimation of power spectra. *J. Roy. Statist. Soc. Ser. B* **41** 313–351.

KRUSKAL, J. B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'. In *Statistical Computation.* (R. C. Milton and J. A. Nelder, eds.) Academic, New York.

KRUSKAL, J. B. (1972). Linear transformation of multivariate data to reveal clustering. In *Multidimensional Scaling: Theory and Application in the Behavioral Sciences*, I, *Theory.* Seminar Press, New York and London.

LI, G. and CHEN, Z. (1981). Robust projection pursuit estimation for dispersion matrices and principal components. Research Report, Dept. of Statist., Harvard University.

LOGAN, B. F. (1975). The uncertainty principle in reconstructing functions from projections. *Duke Math. J.* **42** 661–706.

LOGAN, B. F. and SHEPP, L. A. (1975). Optimal reconstruction of a function from its projections. *Duke Math. J.* **42** 645–659.

MCDONALD, J. (1982). Unpublished manuscript.

SHEPP, L. A. and KRUSKAL, J. B. (1978). Computerized tomography: the new medical x-ray technology. *Amer. Math. Monthly* **85** 420–439.

STAHEL, W. A. (1981). Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen. Ph.D. Thesis, Swiss Federal Institute of Technology, Zurich.

STONE, C. J. (1977). Nonparametric regression and its applications. *Ann. Statist.* **5** 595–645.

SWITZER, P. (1970). Numerical Classification. In *Geostatistics.* Plenum, New York.

SWITZER, P. and WRIGHT, R. M. (1971). Numerical classification applied to certain Jamaican eocene nummulitids. *Math. Geol.* **3** 297–311.

TUKEY, J. W. (1982). Control and stash philosophy for two-handed, flexible, and immediate control of a graphic display. *Bell Labs. Tech. Memo.*

TUKEY, P. A. and TUKEY, J. W. (1981). Graphical display of data in three and higher dimensions. In *Interpreting Multivariate Data.* (V. Barnett, ed.) Wiley, New York.

VAPNIK, V. N. and ČERVONENKIS, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

VITUSHKIN, A. G. (1978). On representation of functions by means of superpositions and related topics. *Enseign. Math.* Monogr. no. 25.

WEGMAN, E. J. (1972). Nonparametric probability density estimation I: A summary of available methods. *Technometrics* **14** 533–547.

WIGGINS, R. A. (1978). Minimum entropy deconvolution. *Geoexploration* **16** 21–35.

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS 02138

DISCUSSION

JEROME H. FRIEDMAN

Stanford University

I congratulate Professor Huber for an excellent survey of Projection Pursuit methods. Putting together the diverse research in this area into a coherent prospective is a difficult and challenging task. This paper represents an important contribution to this methodology, as well as to statistics in general. Professor Huber was one of the first to recognize the potential of projection pursuit methods and his early support was instrumental in inspiring others to pursue research in this area. This paper integrates his own research with that of others into a unified perspective. After reading this paper, I am of the opinion that projection pursuit may have been a good idea after all. I also congratulate the editors of the *Annals* for soliciting this contribution. Although this paper is a little different than those usually found in *The Annals of Statistics*, I believe it nonetheless brings considerable credit to this journal.

I divide my comments into three parts. The first part provides a brief historical perspective and discusses the trade-off between methodology motivated by purely heuristic (ad hoc) concerns as opposed to that based on sound statistical theory. The second part expands Professor Huber's discussion (Sections 11, 13, and 15) on the relation between the "synthetic" and "analytic" versions of PPDE. Finally, I update his discussion of PPR (Sections 9 and 20) by relating some more recent work in this area.

The original projection pursuit algorithm (Friedman and Tukey, 1974) had a modest goal. It was intended as an adjunct to the PRIM-9 graphical display system (Friedman, Fisherkeller, and Tukey, 1975). PRIM-9 provided the user the ability to perform what one would now term manual projection pursuit. Displayed on the screen of a computer graphics terminal was a two-dimensional projection of a multivariate data set. The parameters (linear combinations) defining the projection could be changed in a continuous manner under user control using graphical input devices (light pen and buttons). As the user interacted with the device by changing the projection parameters, the point cloud would appear to move on the screen. This changing configuration would tend to become either more or less interesting as judged by visual inspection. If it was becoming more interesting, the user would continue to change the parameters in the same manner. If not, he would stop and try to change the projection in a different way. In this way, the user would manually iterate to "interesting" views of his data set.

The "batch" projection pursuit algorithm was intended as sort of an automatic pilot to aid with this strategy. Starting with the current view, the algorithm would try to imitate the strategy that a human would use to try to find a more interesting projection. This projection would then be the starting point for human interaction and so on. The details of the algorithm (projection index, optimization method) were worked out by observing people using PRIM-9. The algorithm tried to imitate the successful approaches that people used to find interesting projections. I believe that this is the reason that it worked so well. As Huber points out, we now know that the projection index we used is related to a form of entropy measure (density squared) and that there may be better ones based on other entropy measures. At the time, however, the only motivation for the procedure was heuristic; it seemed to work well in practice and provide interesting results. There was, at that time, no mathematics to "back it up."

Projection Pursuit Regression (Friedman and Stuetzle, 1981) was also originally motivated by purely heuristic reasoning. The idea was to extend linear regression by viewing it as a numerical optimization procedure and substituting a curve estimate (smooth) for the (univariate) straight line fit in the inner loop of the optimization algorithm. This procedure was iterated by applying it to the residuals of the preceding step. We now know that this algorithm approximately minimized

$$E[Y - \textstyle\sum_{m=1}^{M} f_m(\alpha_m^T x)]^2 \tag{1}$$

with respect the parameters of the projections $\alpha_m^T = (\alpha_{1m} \cdots \alpha_{pm})$ and the projection functions f_m, along with a forward stepwise approach for choosing M. Also, we now have better algorithms for minimizing (1) (Friedman, 1984b,c). At the time, however, PPR was a purely ad hoc procedure with no foundation other than it seemed to work well.

Projection pursuit methods are by no means the only ones to be originally ignored for lack of theoretical justification. Factor analysis, clustering, multidimensional scaling, recursive partitioning, correspondence analysis, soft modeling (partial-least-squares), represent methods that were in common use for many years before their theoretical underpinnings were well understood. Again, the principal justification for their use was that they made sense heuristically and seemed to work well in a wide variety of situations.

I think there are two lessons to be derived from this experience. First, if a method works well in practice, there is probably a strong theoretical reason why, and we ought to try to find it. In the case of projection pursuit, Professor Huber has made substantial contributions to this, as have others (see Huber references). Second, a good way to improve statistical methodology might be to observe how good data analysts use powerful interaction tools. The PRIM-9/projection pursuit connection is twelve years old. With today's powerful technology, along with a new generation of practitioners to use it, many new and powerful methods will likely emerge.

My next set of remarks concerns Huber's discussion of projection pursuit density estimation and the relationship between the "synthetic" and "analytic" approaches. From Huber's discussion, one would get the impression that the analytic approach was preferable to (or at the very least, as good as) the synthetic version. First, it can be shown to exactly replicate product densities and, second, it does not require numerical or Monte Carlo integration. This is the good news. The bad news is that it provides poor estimates, poor enough to generally render it useless in practice.

First, it should be noted that the analytic version can never produce zero density estimates in a region where the initial density $f_0(x)$ has support. The converse is also true; the synthetic version cannot produce estimates $f(x)$ in regions where $f_0(x)$ has no support. However, the symmetry is broken since we have control over $f_0(x)$—we can choose it—and (as Huber points out) it is easy to choose $f_0(x)$ to have wider support than $f(x)$. The converse is, of course, not true. The practical consequence is that the analytic version will produce highly upward biased estimates in regions of low data density.

There is an even more serious problem with the analytic algorithm. It produces estimates with very high variance. This is because it weights the observations to produce marginal density estimates. If $f(x)$ is not very close to $f_0(x)$, then the weights will have high variability, inducing high variance in the marginal estimates; the few observations with high weight will dominate the estimates. Again, the converse is also true. The synthetic algorithm uses weighted Monte Carlo (or numerical quadrature) points to produce marginal estimates. But again, the symmetry is broken by the fact that we can control the variance of the Monte Carlo estimates by controlling the number of Monte Carlo points—the higher the variance of the weights, the more Monte Carlo points we use. This is implicitly done in the accept/reject sampling method employed by Friedman, Stuetzle and Schröeder (1984). Unless one has access to unlimited data, it is not very sensible to vary the data sample size in this manner. One should use all the data (all the time) in a "synthetic" strategy.

I would next like to comment on Huber's discussion concerning the engineering aspects of implementing projection pursuit procedures. As Huber correctly points out, these aspects are seldom emphasized in published accounts and, yet, they are often absolutely crucial to the success of the method. It is very unlikely that an implementation based purely on the published description would work. There are many design decisions that are made at the level of writing the code which are usually based on considerable experimentation. These descriptions are usually the first to be cut out when editors, pressed for space, suggest shortening the manuscript. A more serious consequence is that the apparent success or failure of a new method will be based on how well the complete implementation performs. A good idea poorly implemented will not work well and will likely be judged not good. It is likely that the idea of projection pursuit would have been delayed even further if working implementations of the exploratory (Friedman and Tukey, 1974) and regression (Friedman and Stuetzle, 1981) procedures had not been produced. As data analytic algorithms become more complex, this problem becomes more acute. The best way to guard against this is to become as literate as possible in algorithms, numerical methods and other aspects of software implementation. I suspect that more than a few important ideas have been discarded because a poor implementation performed badly.

Finally, I would like to bring the discussion of projection pursuit regression up to date by relating some recent extensions to the procedure, as well as some improved engineering aspects. I start with the engineering aspects. Huber correctly points out that the critical engineering relates to the method used for curve estimation and the minimization strategy used to find the optimal linear combinations. Both of these have been substantially improved in the most recent implementation (Friedman, 1984c).

The new curve estimation procedure is detailed in Friedman (1984a). I sketch it briefly here. Consider first the abstract version of the PPR algorithm. At any point in the minimization procedure, the best curve, given all the projections and all the other curves, is given by a conditional expectation (Huber, equation 9.7). Thus, it would seem that in the sample version of PPR the proper trade-off between over- and underfitting at any particular stage would be achieved by obtaining the best possible estimates for the corresponding conditional expectations. The quality of the conditional expectation estimates can itself be estimated by cross-validation procedures. The meta parameter of the curve estimation procedure (smoother span) is chosen to optimize the cross-validated predictive-squared-error separately in each projection to which it is applied. In addition, local cross-validation is used to estimate optimal span as a function of abscissa value $z = \alpha^T x$ within each projection. The result is the best possible curve estimate for each projection. This is important. Using a constant global span for all projections discourages the algorithm from iterating towards projections that are good, but for which the particular global span value may not be appropriate for the curve estimate in that projection. Also, an important side effect is that the user is relieved from the burden of having to guess at a good span value. No attempt at robustification is made at the curve estimation stage of the algorithm. The best way to achieve robustification (I believe) is to iteratively invoke the entire PPR fitting procedure with observation weights determined by the residuals from the previous solution in analogy with linear regression robustification.

The new minimization strategy is described in Friedman (1984b,c). It is an alternating optimization procedure based on Gauss-Newton stepping rather than a nested optimization based on Rosenbrock stepping. (The Gauss-Newton procedure was suggested by A. Buja.) The practical consequence is that the algorithm is far more reliable and typically 10–20 times faster than the previous version. This reliability and speed increase are important in extending the size of applications to which PPR can be applied. An even more important consequence is that PPR can now be used as a primitive operation that is itself iterated in more complex algorithms. A straightforward application is to apply the backfitting procedure to the projection parameters as well as the curve estimates, thereby producing estimates that actually minimize (1) with respect to the empirical distribution. These estimates can be considerably different than the stagewise estimates for the projection parameters produced by the previous algorithm, especially in the presence of high association among predictor variables.

The new speed of the PPR algorithm also permits more extensive generalizations. One such generalization is to multiple response variables which, in turn, leads naturally to the classification problem. The multiple response models take the form

$$(2) \qquad \hat{Y}_i = E[Y_i \mid x_1 \cdots x_p] = \bar{Y}_i + \textstyle\sum_{m=1}^{M} \beta_{im} f_m(\alpha_m^T x) \quad (1 \le i \le q)$$

with $\bar{Y}_i = EY_i$, $Ef_m = 0$, $Ef_m^2 = 1$ and $\alpha_m^T a_m = 1$. The coefficients β_{im}, $\alpha_m^T = (\alpha_{1m} \cdots \alpha_{pm})$ and the functions f_m are parameters of the model and are estimated by least squares; the estimates are chosen to be those values that minimize

$$(3) \qquad L_2 = \textstyle\sum_{i=1}^{q} W_i E[Y_i - \hat{Y}_i]^2.$$

The W_i represent a user specified loss metric. Models that take the form (2) are termed SMART for Smooth Multiple Additive Regression Technique. Details of the algorithm for minimizing (3) for SMART models are given in Friedman (1984b,c). From (2), it is seen that each response variable is modeled as a (usually) different linear combination of a set of predictor functions f_m. Each predictor function f_m takes the PPR form: a smooth, but otherwise unrestricted, function of a linear combination of the predictor variables.

SMART models (2) are not the only possible extensions of PPR to multiple responses. A more straightforward approach would be to model each response variable Y_i $(1 \leq i \leq q)$ separately with PPR, obtaining models of the form

$$\hat{Y}_i = \textstyle\sum_{m=1}^{M} f_{im}(\alpha_{im}^T x). \tag{4}$$

Models of this form (4) are contained as special cases of SMART models (2). However, SMART models have the opportunity to take advantage of (possibly high) associations among the response variables to produce much more parsimonious models (fewer predictor functions and linear combinations). This can reduce the variance of estimates, as well as produce more interpretable models.

In the classification problem, a single response variable Y assumes several categorical (unorderable) values $(c_1, c_2, \cdots, c_q)$. This can be converted to a multiple response regression problem by associating dummy (zero/one) variables H_i for each categorical value c_i:

$$H_i = \begin{cases} 1 & \text{if } \; Y = c_i \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Clearly, $\hat{H}_i = E[H_i \mid x_i \cdots x_p] = \Pr\{Y = c_i \mid x_1 \cdots x_p\}$. The $\hat{H}_i$ are estimated via SMART (2, 3). The resulting estimates are used in a minimum (estimated) loss decision rule

$$\hat{Y} = \min\nolimits_{1 \leq j \leq q}^{-1} \{\textstyle\sum_{i=1}^{q} (\pi_i l_{ij}/s_i)\hat{H}_i\}$$

where π_i is the unconditional (prior) probability of $Y = c_i$, l_{ij} is the loss for estimating $\hat{Y} = c_j$ when the truth is $Y = c_i$, and s_i is the total mass of observations for which $Y = c_i$.

SMART modeling has been applied to several problems with success. See Friedman (1984c) for some examples. Its general value will have to be learned from experience.

I once again congratulate Professor Huber for a wonderful paper, and I thank the editors of *The Annals of Statistics* for inviting me to make these comments.

REFERENCES

FRIEDMAN, J. H. (1984a). A variable span smoother. Dept. of Statist., Stanford University, Report LCM005.

FRIEDMAN, J. H. (1984b). Classification and multiple response regression through projection pursuit. Dept. of Statist., Stanford University, Report LCM006.

FRIEDMAN, J. H. (1984c). SMART User's Guide. Dept. of Statist., Stanford University, Report LCM001.

FRIEDMAN, J. H., FISHERKELLER, M. A. and TUKEY, J. W. (1975). PRIM-9: An interactive multidimensional data display and analysis system. Stanford Linear Accelerator Center, Report SLAC-PUB-1408.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statis. Assoc.* **76** 817–823.

FRIEDMAN, J. H. and STUETZLE, W. (1982). Projection pursuit methods for data analysis. In *Modern Data Analysis*. (R. Launer and A. F. Siegel, eds.), Academic, New York.

FRIEDMAN, J. H., STUETZLE, W. and GROSSE, E. (1983). Multidimensional additive spline approximation. *SIAM J. Sci. Statist. Comput.* **4** 291–301.

FRIEDMAN, J. H., STUETZLE, W. and SCHRÖEDER, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79** 599–608.

FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C-23** 881–889.

HASTIE, T. (1983). Nonparametric logistic regression. Dept. of Statist., Stanford University, Report ORION 016.

HENRY, D. H. (1983). Multiplicative models in projection pursuit. Dept. of Statist., Stanford University, Report ORION 025.

MCDONALD, J. A. (1982). Projection pursuit regression on the ORION-I workstation. Film, Bin 88 Productions, Stanford Linear Accelerator Center, Stanford University.

MITTAL, Y. (1983). Two-dimensional projection pursuit tests for goodness-of-fit and equality of distributions. Dept. of Statist., Stanford University.

TIBSHIRANI, R. (1982). Censored data regression with projection pursuit. Dept. of Statist., Stanford University, Report ORION 013.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

FRED L. BOOKSTEIN
The University of Michigan

Professor Huber neatly reminds us how both principal-components analysis and multiple linear regression may be viewed as special cases of PP. In this note I would like to suggest a ramification of PP that incorporates the considerably wider class of *soft models* (iterated regression protocols), variants of canonical correlations analysis, lately developed by Herman Wold.

1. Section 7 of Huber's article considers "questions of k-dimensional projections." Rephrasing this as "one k-dimensional projection," I shall reverse the multiplicities to consider k one-dimensional projections instead. As Huber notes, for a computation of multiple dimensions to be most easily interpretable, each should be characterized uniquely. The simplest identification represents the measurement space $\mathbb{R}^n$ as the direct product $\prod_{i=1}^{k} \mathbb{R}^{n_i}$ of subspaces. In the language of causal modeling, these are *measurement blocks*.

Partition Huber's random vector $\mathbf{X}$ as $(X_1 : X_2 : \cdots : X_k)$, conformally with the projection vector $\mathbf{a} = (a_1 : a_2 : \cdots : a_k)$. Each X_i or a_i is a vector of length n_i, with $\sum_{i=1}^{k} n_i = d$, the original dimension of $\mathbf{X}$. Our "multiple projection" is then the k-vector $(Z_1, \cdots, Z_k) = (a_1^T X_1, \cdots, a_k^T X_k)$. An appropriate normalization sets each vector a_i to length 1, so that the vector $\mathbf{a}$ has Euclidean norm k rather than 1. The space of multiple projections, then, is no longer the unit d-sphere $\mathbb{S}^{d-1}$, but rather the direct product $\prod_{i=1}^{k} \mathbb{S}^{n_i - 1}$ of unit n_i-spheres.

2. PP finds interesting projections by the numerical examination of an objective function Q that measures "interestingness" in some fashion. For the objective functions $Q(Z)$ of a single projection, interestingness seems to have a useful general interpretation: nonnormality. For the multiple PP I am suggesting here, there is a more fundamental aspect of interest: *dependence*. For instance, for $k = 2$ we might use $Q(Z_1, Z_2) = \{\text{correlation}\}$. The projection vector $\mathbf{a} = (a_1, a_2)$ that