REFERENCES

ANDREWS, D. F (1972). Plots of high dimensional data. *Biometrics* **28** 125–36.

ANDREWS, D. F., GNANADESIKAN, R. and WARNER, J. L. (1971). Transformations of multivariate data. *Biometrics* **27** 825–40.

ANDREWS, D. F., GNANADESIKAN, R. and WARNER, J. L. (1973). Methods for assessing multivariate normality. In *Multivariate Analysis III.* (P. R. Krishnaiah, ed.) 95–116. Academic, New York.

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

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

GNANADESIKAN, R., KETTENRING, J. R. and LANDWEHR, J. M. (1982). Projection plots for displaying clusters. In *Statistics and Probability: Essays in Honor of C. R. Rao.* (G. Kallianpur et al., eds.) 269–80, North-Holland, Amsterdam.

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.

ROY, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.* **24** 220–38.

ROY, S. N. (1957). *Some Aspects of Multivariate Analysis.* Wiley, New York.

BELL COMMUNICATIONS RESEARCH
MORRISTOWN, NEW JERSEY 07960

TREVOR HASTIE AND ROBERT TIBSHIRANI[1]

Stanford University and Stanford Linear Accelerator Center

We would like to thank Professor Huber for this far-reaching yet penetrating discussion of projection pursuit methods.

Our comments will touch upon three areas: inference, the relation of PPDE to the Iterative Proportional Scaling algorithm, and the extension of PPR models to other settings.

1. Inference. Professor Huber discusses only briefly (Section 21) the problem of inference for PP models. But if PP is to be used for data analysis, we feel that this is an important question. We will concentrate on the PPR model, although qualitatively our findings should apply to PPDE and perhaps to other PP procedures as well. Suppose that we have fit a one-term PPR model of the form $\hat{y} = g(\hat{\mathbf{a}}'\mathbf{x})$ to a set of data with p predictors and n observations. An important question is: Is the direction $\hat{\mathbf{a}}$ really "significant," or just an artifact of our search over all possible directions? We can answer this by comparing the observed decrease in the corrected sum of squares $D = \sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2$

[1] This work was supported by the Department of Energy under contracts DE-AC03-76SF and DE-AT03-81-ER10843, and by the Office of Naval Research under contract ONR N00014-81-K-0340, and by the U.S. Army Research Office under contract DAAG29-82-K-0056.

to its null distribution. To simplify the discussion, let's assume that D has been scaled by the true error variance. In the special case in which $g(\cdot)$ is forced to be linear, standard regression theory gives us the exact null distribution. Then $g(\hat{\mathbf{a}}'\mathbf{x})$ is just the least squares solution, and assuming that the y's are normally distributed, D is distributed χ^2_p. For example, if $p = 5$, this means that D must be at least 11.07 to be significant at the 5% level (one-sided). From this we can see the (inherent) adjustment for searching over all possible linear combinations $z = \mathbf{a}'\mathbf{x}$: a single predictor that produces a decrease of 11.07 is significant at $<.1\%$! And the larger p is, the more significant z has to be.

In the projection pursuit model, we would expect this effect to be at least as large. Note that in the case when the $g(\cdot)$ was linear, the degrees of freedom of $D(p)$ is just the number of independent parameters in the direction $(p - 1)$ plus the number in the linear fit (2) minus 1 for the constant model. Thus for a PPR model, we might guess that the degrees of freedom is $p - 1$ plus the number of degrees of freedom on the smooth minus 1. Cleveland (1979) and Tibshirani and Hastie (1984) provide a method of computing the degrees of freedom of a smooth: for a running lines smoother, expressible in the form $\hat{\mathbf{y}} = S\mathbf{y}$, the degrees of freedom is exactly $\text{tr}(S)$. The smoother matrix S depends on the abscissa values and the span of the smoother (but *not* on $\mathbf{y}$). (We use the term "degrees of freedom" to denote the mean of the distribution of D. This distribution is actually a little more spread out than the corresponding χ^2 distribution.)

We found the null distribution of D by simulation to find out if the degrees of freedom do indeed add. To fix attention on a set of covariates, we used the ozone concentration data of Breiman and Friedman (1984), considering only the five variables arrived at in their paper. We generated data from the null model $y_i = \varepsilon_i \sim \mathcal{N}(0, 1)$, found the first direction and computed D. This was done 500 times, and the average value of D was computed. Figure 1 (solid curve) displays the results, for span .1 and larger spans. Now $\text{tr}(S)$ was 11.54 (this number is actually based on the direction $\mathbf{a}' = (.80, -.38, .37, -.24, -.14)$, but depends very little on the direction used), hence if the degrees of freedom add, the expected value of D should be $(5 - 1 + 11.54 - 1) = 14.54$. This number and the numbers for larger spans are shown in Figure 1 (broken curve). For smaller spans, the actual degrees of freedom is almost twice the expected number! As the span increases, this effect starts to disappear. There is some kind of interaction between the directional search and span size, an effect that would be important to understand.

Determining the variability of a PPR fit is another (closely related) problem: the bootstrap (Efron, 1979) can help us here. Figure 2 shows the estimated function from a single PPR term for the ozone concentration data, the span of the smoother fixed at .1. This is the smooth for the direction $\hat{\mathbf{a}}' = (.80, -.38, .37, -.24, -.14)$, and explained 72% of the residual variation. Figure 3 shows smoothed bootstrap histograms of 200 values of $\mathbf{a}^*$. Each was obtained by sampling with replacement from the *tuples* $(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_{330}, \mathbf{y}_{330})$ and fitting a one-term PPR model. Also shown (dotted histograms) are the results for linear $g(\cdot)$. (Note that $-\hat{\mathbf{a}}^*$ gives the same fit as $\hat{\mathbf{a}}^*$; we chose the one having the smallest angle

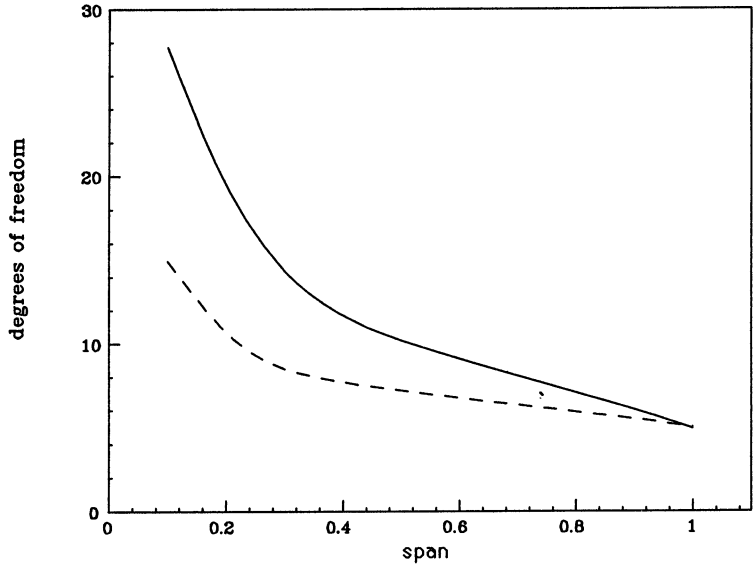

FIG. 1. *The solid curve shows the average decrease in residual sum of squares when we fit a one-term PPR model to null data, as a function of the span of the smoother. The broken curve is what we obtain by adding the degrees of freedom of the direction and the smooth.*

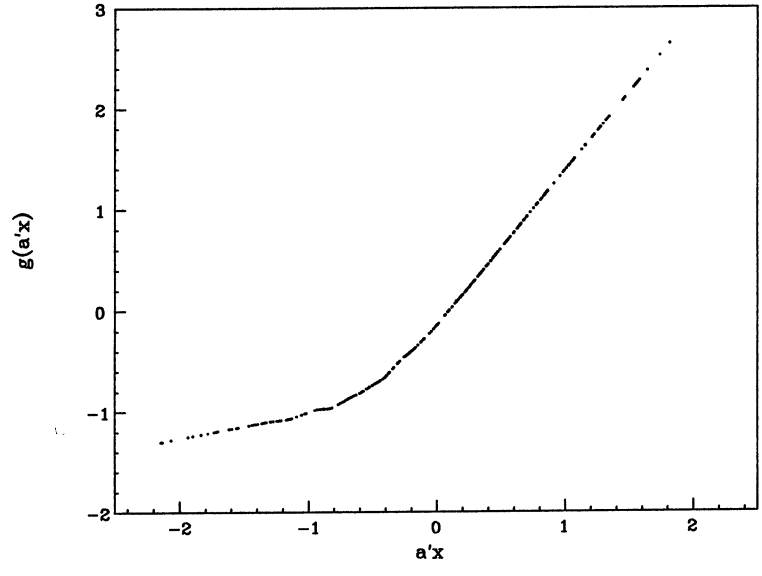

FIG. 2. *The estimated smooth function for the first term in the PPR model.*

with $\hat{\mathbf{a}}$.) The PPR model shows only slightly more variability than the linear regression model; interestingly, some of the distributions are offset from the original $\hat{\mathbf{a}}_i$ (vertical spikes). Figure 4 shows the results for the second term of a two-term PPR model, which explained another 6% of the residual variation (the

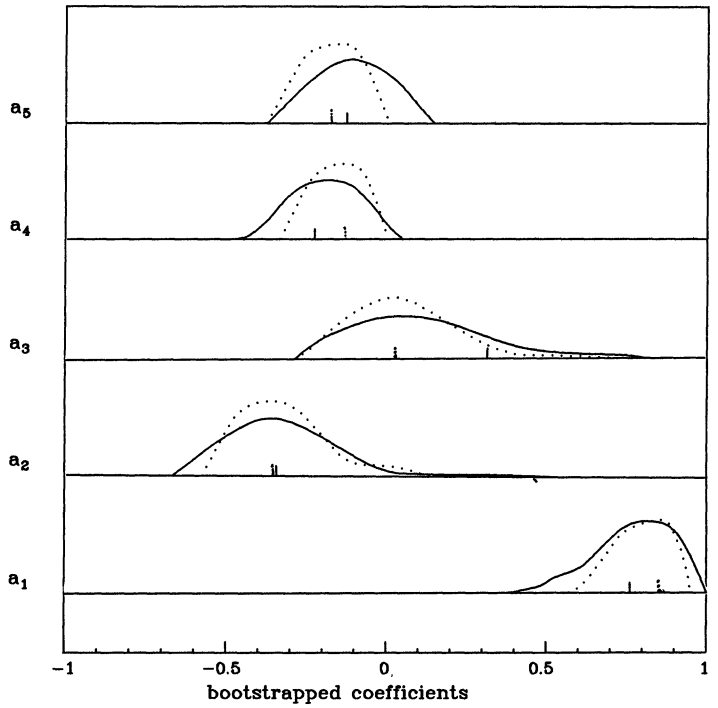

FIG. 3. *Smoothed histograms of the bootstrapped coefficients for the one-term PPR model. Solid histograms are for span .1; the dotted histograms are for linear* $g(\cdot)$.

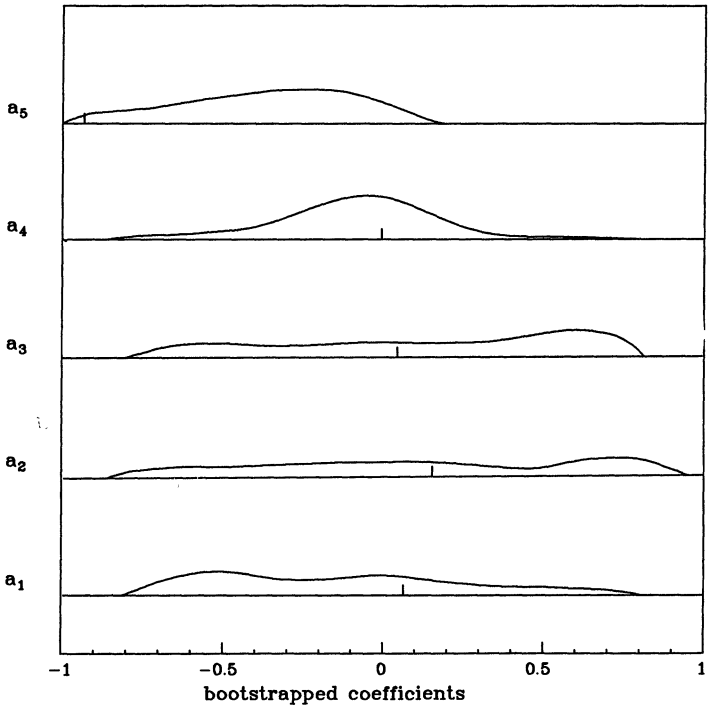

FIG. 4. *The histograms of the bootstrapped coefficients for the second term in the PPR model, span* = 1.

first term, somewhat surprisingly, looked almost identical to Figure 3). The histograms show a great deal of variability. A. C. Ehrenberg notes that many standard regression textbooks tell us on one page not to interpret multiple regression coefficients, then on the next page, they give a data example and interpret each of the coefficients! These bootstrap results make us even more cautious about interpreting the directions of a PPR model.

2. PPDE and log-linear models for contingency tables. There exist strong analogies between PPDE and PPDA and the Maximum Likelihood Estimation (MLE) of log-linear models for contingency tables (see Fienberg, 1977, for a full practical treatment of the latter, and Bishop et al., 1974, for a more theoretical treatment).

A k-way contingency table can be represented by a k-dimensional array X, where $x(i_1, i_2, \cdots, i_k)$ is the proportion of observations (out of N) occurring simultaneously in category i_1 of variable 1, i_2 of variable 2, etc. There is a corresponding *true* array of probabilities P and the data are assumed to be a multinomial sample from this model. The class of *log-linear* models $\mathcal{G}$ specify models for P of the form $g(i_1, \cdots, i_k) = g_0 \Pi g_{l_1,\cdots,l_q}(i_{l_1}, \cdots, i_{l_q})$ where $l_1, \cdots, l_q$ is any subset of the variables $1, \cdots k$. The following are examples of log-linear models for the array P:

1) $g(i_1, \cdots, i_k) = g_0 g_1(i_1)$: one nonuniform marginal
2) $g(i_1, \cdots, i_k) = g_0 g_1(i_1) \cdots g_k(i_k)$: complete independence.
3) $g(i_1, \cdots, i_k) = g_0 g_{12}(i_1, i_2) \cdots g_{1k}(i_1, i_k) \cdots g_{(k-1)k}(i_{k-1}, i_k)$: no third-order interaction.

There exists in fact a *hierarchy* of models ranging in complexity from the uniform model g_0 to the saturated model P itself. The functions g_a are discrete, e.g. $g_1(i_1)$ has I_1 values: $g_1(1), g_1(2), \cdots, g_1(I_1)$, where I_1 is the number of categories in variable 1. These I_1 values are regarded as parameters.

Typically the functions are estimated by maximizing the multinomial likelihood, which corresponds exactly to minimizing an estimate of the Kullback-Leibler distance, or Relative Entropy:

$$\hat{E}(P, G) = \sum_{i_1,\cdots,i_k} x(i_1, \cdots, i_k) \log x(i_1,\cdots,i_k)/g(i_1,\cdots,i_k)$$

as in Section 13–15 in Huber. The *Iterative Proportional Scaling Algorithm* (IPS, Deming and Stephan, 1940) solves the likelihood equations. For just one term, as in model 1 above, the algorithm finds g_1 so that model and data marginal agree in coordinate 1. This is exactly the PPDE estimation procedure of marginal adjustment as outlined in Dr. Huber's paper, and as described in Friedman, Stuetzle and Schroeder (1984). For models such as 3 above with many terms, the likelihood equations require that the particular model and data marginals agree.

For example, since the term g_{12} is included, one set of MLE equations is: $g(i_1, i_2, +, +, \cdots, +) = x(i_1, i_2, +, +, \cdots, +)$. A similar set occurs for each term in the model, and is a direct consequence of MLE in the exponential family. The iterative proportional scaling algorithm cycles through adjusting each model marginal in turn, until at convergence all the constraints are satisfied. This iterative proportional scaling corresponds exactly to the back adjustment procedure of Friedman et al. for continuous data. A typical procedure in log-linear modeling is to select the appropriate model terms in a manual forward or backward stepwise fashion. An automatic search would correspond to discrete projection pursuit, where the "directions" are restricted to this special set of marginal "projections."

Diaconis (1983) defines the notion of a general projection for discrete data. He uses the Radon transform on fairly arbitrary partitions of the original table. He proposed selecting a projection which minimizes the marginal relative entropy, or the projection "furthest from uniform." This corresponds exactly to selecting the term in a one-term log-linear model by MLE, where the set of possible terms includes these general projections (the relative entropy factors into *constant + marginal entropy*). One could then build up a discrete PP multiplicative model by successively adding these general projection terms, with fitting performed by the IPS algorithm.

3. Generalized projection pursuit regression. Generalized Linear Models (Nelder and Wedderburn, 1972) are a class of likelihood based linear regression models defined in particular for the exponential family. The mean μ is linked to the vector of covariates $\mathbf{x}$ via the *link* function g: $g(\mu) = \beta'\mathbf{x}$. An example is logistic regression in which the mean is p, the probability of success, and $g(p) = \text{logit}(p) = \log(p/(1-p)) = \beta'\mathbf{x}$. The models are fit by maximum likelihood. Hastie and Tibshirani (1984) generalize this class to include additive models of the form $g(\mu) = \sum s_j(x_j)$ where $s_j(\cdot)$ are general smooth functions. The functions are estimated using the *Local Scoring* procedure, which stated simply amounts to using the Fisher Scoring procedure to minimize the expected log-likelihood. They also discuss the simple modifications needed to fit *generalized projection pursuit models* of the form

$$g(\mu) = \sum_{k=1}^{m} s_k(\mathbf{a}_k'\mathbf{x})$$

using local scoring. This allows us to do PPR in a large class of problems that include logistic regression, the Cox regression model for censored data, constant CV models and the whole class of Quasi-Likelihood models as discussed in McCullagh and Nelder (1983).

Acknowledgements. We thank Leo Breiman for allowing us to use his air pollution data, and Jerry Friedman for the use of his PPR program SMART and helpful discussion.

REFERENCES

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practise.* MIT Press, Cambridge, MA.

BREIMAN, L. and FRIEDMAN, J. H. (1982). Estimating optimal transformations for multiple regression and correlation. Dept. of Statist. technical report, Orion 16, Stanford University.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

DEMING, W. E. and STEPHAN, F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11** 427–444.

DIACONIS, P. (1983). Projection pursuit for discrete data. Stanford University technical report no. 198.

EFRON, B. (1979). Bootstrap methods: another look at the jacknife. *Ann. Statist.* **7** 1–26.

FIENBERG, S. E. (1977). *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge, MA.

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

HASTIE, T. and TIBSHIRANI, R. (1984). Generalized additive models. Stanford University, Lab. for Comput. Statist. report no. 2.

MCCULLAGH, P. and NELDER, J. (1983). *Generalized Linear Models.* Chapman Hall, London.

NELDER, J. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

TIBSHIRANI, R. and HASTIE, T. (1984). Local likelihood estimation. Stanford technical report 97 and unpublished Ph.D. thesis (1st author), Dept. of Statist. Stanford University.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

M. C. JONES

University of Birmingham

Professor Huber presents a most interesting paper reviewing the broad area within multivariate data analysis now encompassed by the term "projection pursuit." My own comments relate to recent research in this field undertaken at the University of Bath, UK, by myself and Professor Robin Sibson. Our work focussed on the basic projection pursuit algorithm thought of as an exploratory tool applied to point clouds—as a method for finding "interesting" low-dimensional "views" of a multivariate data set—in the spirit of Friedman and Tukey (1974); as such, these comments are most relevant to Section II of the current paper.

Initially, we had access only to Friedman and Tukey's pioneering paper and during much of the course of our work remained unaware of the more recent work by Professors Huber, Friedman and others. Bearing this in mind, the close agreement between many of Professor Huber's ideas and our own, which are outlined below, seems quite remarkable.

The particular implementation of the projection pursuit method described by Friedman and Tukey allowed considerable scope for improvement on both theoretical and practical grounds. Consequently our aim was to provide a new