cycle around a loop of three steps: (1a, 1b) separately regress the dependent block score (by PPR) upon each of the two predictor blocks; (2) regress the same dependent block score (by PPR) upon a small hybrid block of dimension 2 consisting of the pair of partial predictors from step 1; (3) regress (by PPR) this bivariate two-block predictor upon the variates of the dependent block. The prediction function becomes a revised dependent variable for step 1, and so forth until convergence, one hopes.

The extension of PP to two-block and multiple-block designs involves two themes: the search for k projections rather than one, and the iterative refinement of projections by alternating regression. Such an incorporation into PP of the two main themes of soft modeling should considerably enhance its power for the point clouds of complicated dimensional structure that arise in biometrics, interdisciplinary developmental studies, and all the other arenas for which "theoretical knowledge," in Wold's phrase, "is scarce." I thank Professor Huber and the editor of these *Annals* for the opportunity to see and explain this connection.

REFERENCES

BOOKSTEIN, F. L. (1980). Data analysis by partial least squares. In *Evaluation of Econometric Models.* (Kmenta and Ramsey, eds.) 75–88, Academic, New York.

BOOKSTEIN, F. L. (1982). The geometric meaning of soft modeling, with some generalizations. In *Systems under Indirect Observation: Causality, Structure, Prediction* (Jöreskog and Wold, eds.) II:55–74. North-Holland, Amsterdam.

BOOKSTEIN, F. L. (1984). Tensor biometrics for changes in cranial shape. *Ann. Hum. Biol.* **11** 413–437.

JÖRESKOG, K. and WOLD, H., eds. (1982). *Systems under Indirect Observation: Causality, Structure, Prediction.* 2 vols. Contributions to Economic Analysis, No. 139. North-Holland, Amsterdam.

KMENTA, J. and RAMSEY, J. B., eds. (1980). *Evaluation of Econometric Models.* Academic, New York.

LYTTKENS, E. (1972). Regression aspects of canonical correlations. *J. Multivariate Anal.* **2** 418–439.

WOLD, H. (1975). Path models latent variables: the NIPALS approach. In *Quantitative Sociology: International Perspective on Mathematical and Statistical Modeling.* (H. M. Blalock et al., eds.) 307–357, Academic, New York.

WOLD, H. (1980). Model construction when theoretical knowledge is scarce. In *Evaluation of Econometric Models* (Kmenta and Ramsey, eds.) 47–74, Academic, New York.

WOLD, H. (1982). Soft modeling: the basic design and some extensions. In *Systems under Indirect Observation: Causality, Structure, Prediction.* (Jöreskog and Wold, eds.) II:1–54, North-Holland, Amsterdam.

CENTER FOR HUMAN GROWTH AND DEVELOPMENT
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109

ANDREAS BUJA AND WERNER STUETZLE

University of Washington

Peter Huber's paper is interesting and important. In our opinion its main contributions are:

- The formulation of abstract versions of PPDE and PPR operating on distributions instead of samples. This complements the more intuitive understanding by the inventors of these procedures (Friedman and Stuetzle, 1981a,b; Friedman, Stuetzle and Schroeder, 1984) and makes PP methods amenable to theoretical analysis.

- The classification and analysis of projection indices. An especially nice contribution in this direction is the interpretation of the original Friedman-Tukey projection index as a measure of non-Gaussianity.
- The specification of several unsolved theoretical problems concerning consistency and convergence rates of PP procedures. This will hopefully stimulate additional study and lead to a more thorough understanding of the new methods.

To use the author's own terminology from his paper "Applications versus abstraction: the selling out of mathematical statistics?" (Huber, 1975): Peter Huber has helped take PP from the groping phase into the squeezing phase and opened up a new area of research in mathematical statistics. Congratulations!

We will now comment in more detail on some sections of the paper.

Section 2. In this section the author gives some estimates for the time needed to visually inspect higher dimensional spaces and estimates that it would take 3 to 4 hours to exhaustively inspect a four-dimensional space. Although certainly true under the stated conditions, this does seem to be a far too pessimistic estimate in practice. We have two reasons for this belief:

- If we watch a Grand Tour (Asimov, 1985), we get information not only from the position of the observations at a given time, but also from the direction and speed of their motion on the screen. Looking at 1,000 projections in a Grand Tour conveys more information than looking at 1,000 projections onto randomly chosen planes. How much more information we get, i.e. how well the human visual system can, for example, detect clusters in the speed and direction of the movement of dots, is an interesting open question. In our experience with Grand Tour implementations we found that clusters could be detected through motion even when the static pictures would not show any structure at all. Of course, the information conveyed by motion tends to show up in the static pictures as well, at least at some point, but we argue that the wealth of information in dynamic graphics transcends the mere purpose of finding static informative pictures. In the situation mentioned above, the Grand Tour tells us that in the two dimensions spanned by the screen coordinates there is no significant structure, whereas the two dimensions coded by the speed reveal clusters. This is a statement about the shape of the point cloud in a four-dimensional subspace.
- There are not very many examples of structure detectable by projection, but only if the projection plane is correct to within, say, five degrees. The only example that came to mind is a cylindrical hole all the way through the data. Note that in the case of the parallel planes produced by RANDU, which supposedly are difficult to detect, the squint angle is indeed five degrees, but this of course does not mean that the projection plane has to be within ±2.5 degrees of one particular plane in order for the viewer to detect the structure. Any plane which is orthogonal or within ±2.5 degrees of orthogonality to

> the RANDU planes will show them, and a rough calculation indicates that this condition will be satisfied for about one out of every 24 randomly chosen projection planes.

The RANDU planes do suggest several questions about PP. First, it seems doubtful that any sample version of PP would pick up the planes, because of the smoothing involved in computing projection indices. Second, even if the sample estimate of the projection index had minima at projections which show the RANDU planes, the valleys might be much too narrow to be found by conventional optimizers, which depend on the presence of a nontrivial slope in the index. The fact that the planes are visible only for projections correct to within ±2.5 degrees translates most likely into a PP-index which consists of flat mesas with insufficient slope to hint at the locations of the narrow canyons which contain the minima. In spite of being promoters of PP methods ourselves, we are not quite convinced that this example makes a strong case for PP. On the opposite, it might highlight some unresolved problems.

It is difficult to say how successful manual search and the Grand Tour are in practice, as one never knows what one is missing. Still we may say that the structures we have found in actual data sets rarely ever appear only in a very narrow neighborhood of planes. (Of course that might be the reason why they were found in the first place.) A case in point are the four-dimensional particle physics data shown in the PRIM-9 movie (Fisherkeller, Friedman and Tukey, 1974). If this data set is viewed using a Grand Tour, the four connected rods near which the data are concentrated are apparent almost immediately.

Another relevant point is illustrated by this data set: The two outermost rods are formed by a small number of observations only, and yet they can be seen without much trouble. The reason why we are able to perceive them is that the five or so points (out of a data set of 500) in each of the sparse rods show a linear arrangement of striking purity. In the human visual system, the clarity of a structure can counterbalance sparsity of the data, whereas automated statistical procedures tend to discard minorities, no matter how striking the features they show.

The particle physics data are particularly suitable for visual inspection with three-dimensional rotations and Grand Tour methods because their structure (the four rods) is one-dimensional. The situation is less gratifying when the structure has intrinsic dimensionality higher than three or four. Then we run into the problem that "most projections look normal" according to Diaconis and Freedman (1984). It is typical that the illustration for this effect presented by Huber is a hypercube in seven dimensions, i.e. a configuration which spans a solid seven-dimensional body. The need for optimization over projections becomes more urgent for structure with low codimension, but it does not seem likely to replace dynamic graphics. The ideal environment for graphical exploration would probably combine PP with the Grand Tour and three-dimensional rotations.

Section 8. In the beginning of this section, the author sketches several ways to proceed, after an interesting projection has been found. He states that his option 2 corresponds to PP classification. This is not the case. PP classification, as suggested by Friedman and Stuetzle, is very similar to PP regression. In the case of binary response (2 classes) the idea is to construct a PP model for the conditional probability $P(Y = 1 \mid \mathbf{x})$ of getting a class 1 observation, given predictor vector $\mathbf{x}$. Because probabilities should be positive, it is attractive to build a multiplicative model

$$P(Y = 1 \mid \mathbf{x}) \sim \textstyle\prod_{j=1}^{m} g_j(\mathbf{a}_j^T \mathbf{x}).$$

One possibility for a projection index is the residual sum of squares. This approach was implemented and tested by Henry (1983).

Section 9. At the end of his discussion of PPR the author states that "in general, it is possible to improve the fit by various versions of backfitting: omit one of the earlier summands g_j, determine the best possible replacement and then iterate. Usually, the directions $\mathbf{a}_j$ are kept constant in this process." This idea, suggested by Friedman and Stuetzle (1981), sounds somewhat ad hoc, but can be easily justified. Suppose we have fixed directions $\mathbf{a}_1 \cdots \mathbf{a}_m$, and we wish to find functions $g_1 \cdots g_m$ that minimize the expected residual sum of squares

$$\mathbf{E}(r_m^2) = \mathbf{E}(Y - \textstyle\sum_{j=1}^{m} g_j(\mathbf{a}_j^T \mathbf{x}))^2.$$

It is easy to see that this quantity is minimized if the model agrees with the response in the marginals along $\mathbf{a}_1 \cdots \mathbf{a}_m$:

$$\mathbf{E}(Y \mid \mathbf{a}_k^T \mathbf{x} = z) = \mathbf{E}(\textstyle\sum_{j=1}^{m} g_j(\mathbf{a}_j^T \mathbf{x}) \mid \mathbf{a}_k^T \mathbf{x} = z),$$

or

$$\mathbf{E}(Y - \textstyle\sum_{j \neq k} g_j(\mathbf{a}_j^T \mathbf{x}) \mid \mathbf{a}_k^T \mathbf{x} = z) = g_k(z), \quad k = 1 \cdots m.$$

This is a linear system of equations for the g_j. If we knew all but one of the functions, we could immediately get the remaining one using the above equation. The obvious idea now is to start with trial guesses for the g's and then in cyclical order replace each one by an improved version according to the above equation, until convergence. This way of solving systems of linear equations is well known in numerical analysis; it is called the Gauss-Seidel method. A proof of convergence for the backfitting process is also contained in a forthcoming paper by Breiman and Friedman (1985).

Section 15. This section, where the author discusses several variants of projection pursuit density estimation, merits close scrutiny. Let $f(\mathbf{x})$ denote the true unknown data density, from which we have an i.i.d. sample $\mathbf{x}_1 \cdots \mathbf{x}_n$, and let $g_0(\mathbf{x})$ denote an initial guess for $f(\mathbf{x})$. We wish to find directions $\mathbf{a}_j$ and functions h_j such that $g_0(\mathbf{x}) \prod h_j(\mathbf{a}_j^T \mathbf{x}) = f(\mathbf{x})$. (This is what the author calls the "synthetic view" of PPDE resp. PPDA, in contrast to the "analytic view" to be discussed below.) Suppose we have already estimated k terms of the model, and denote $g^k(\mathbf{x}) = g^0(\mathbf{x}) \prod_{j=1}^{k} h_j(\mathbf{a}_j^T \mathbf{x})$. We wish to determine the next direction $\mathbf{a}_{k+1}$ such that it maximizes the marginal relative entropy $\mathbf{E}(f_{\mathbf{a}}, g_{\mathbf{a}}^k)$, and then put $h_{k+1} = f_{\mathbf{a}_{k+1}}/g_{\mathbf{a}_{k+1}}^k$.

We will now concentrate on the estimation of the model marginal, because this is a crucial point where the various algorithmic approaches differ. Without restriction of generality, we pick $\mathbf{a}_{k+1}$ to be the first coordinate direction, and for notational simplicity we denote the marginal of any multivariate density $p(\mathbf{x})$ along this direction by $p_1(z)$.

The first approach mentioned in the paper is to generate a sample $\mathbf{y}_1 \cdots \mathbf{y}_N$ from g^k and then estimate g_1^k by a kernel estimate, such as

$$\hat{g}_1^k(z) = (1/2N\Delta)\ \textstyle\sum_{i=1}^N \mathbf{I}(z - \Delta \le y_{i1} \le z + \Delta)$$

This is the approach used in the original implementation of PPDE. The sample was generated using the accept/reject method. The process is described in detail in Friedman, Stuetzle and Schroeder (1984).

The second approach mentioned in the paper is to generate a sample from g^0 instead of g^k, and estimate the marginal by

$$\tilde{g}_1^k(z) = (1/2N\Delta)\ \textstyle\sum_{i=1}^N \mathbf{I}(z - \Delta \le y_{i1} \le z + \Delta)(g^k(\mathbf{y}_i)/g^0(\mathbf{y}_i)).$$

It has the advantage that g^0 can be chosen to make generation of the sample easy; we can for example pick it to be multivariate Gaussian with the same mean and covariance matrix as the sample. This approach was also considered at the time when PPDE was invented, and was rejected. The reason was that, vaguely speaking, sampling from g^0 instead of g^k leads to a loss of efficiency in estimating the model marginal, which would have to be compensated for by an increase in the Monte Carlo sample size N. As the program seems to spend most of its time estimating marginals, for which the work grows at least linearly in N, and only a small part generating the Monte Carlo observations, it seemed reasonable to generate Monte Carlo observations according to g^k and keep N smaller.

To see precisely in what sense sampling from a distribution other than g^k will in general result in a loss of efficiency of the marginal estimate, we will now compute the expected value and variance of $\tilde{g}_1^k$ and find out how they depend on the choice of g^0. Note that $\hat{g}_1^k$ is equal to $\tilde{g}_1^k$ if we choose $g^0 = g^k$.

For the expected value we obtain

$$\begin{aligned}\mathbf{E}(\tilde{g}_1^k(z)) &= (1/2\Delta)\mathbf{E}_{g^0}(\mathbf{I}(z - \Delta \le Y_1 \le z + \Delta)(g^k(\mathbf{Y})/(g^0(\mathbf{Y})))\\ &= \mathbf{E}_{g^k}\mathbf{I}(z - \Delta \le Y_1 \le z + \Delta).\end{aligned}$$

This means that the expected value does not depend on g^0; the bias of the marginal estimate is controlled solely by the window width Δ.

Let us now compute the variance of $\tilde{g}_1^k$ and see how it depends on the choice of g^0. We use the ANOVA-like decomposition

$$\operatorname{var} u(\mathbf{Y}) = \mathbf{E}(\operatorname{var}(u(\mathbf{Y}) \mid Y_1)) + \operatorname{var}(\mathbf{E}(u(\mathbf{Y}) \mid Y_1))$$

with

$$u(\mathbf{Y}) = \mathbf{I}(z - \Delta \le Y_1 \le z + \Delta)(g^k(\mathbf{Y})/g^0(\mathbf{Y})).$$

This gives

$$\begin{aligned}\operatorname{var} \tilde{g}_1^k(z) = {} & \mathbf{E}(\mathbf{I}(z - \Delta \le Y_1 \le z + \Delta)\operatorname{var}(g^k(\mathbf{Y})/g^0(\mathbf{Y}) \mid Y_1))\\ & + \operatorname{var}(\mathbf{I}(z - \Delta \le Y_1 \le z + \Delta)(g_1^k(Y_1)/g_1^0(Y_1))).\end{aligned}$$

The first term vanishes if

$$g^k(y_1 \cdots y_k) = c(y_1)g^0(y_1 \cdots y_k),$$

i.e. if the two densities are proportional for any fixed value of the first coordinate. If we make the first term vanish, the variance still depends on the choice of function $c(y_1)$. By choosing this function, we allocate the precision of our estimate—we influence, for which values of the first coordinate we estimate the marginal more precisely and for which we estimate it less precisely. Optimal allocation of the mass of g^0 in the marginal depends on the measure we choose to define the distance between true and estimated marginal. The point is that, whatever allocation of precision we choose, g^0 should be proportional to g^k for any fixed value of the first coordinate. This condition is satisfied in particular if $g^0 = g^k$.

The third approach was assigned to a student for investigation shortly after the first successful implementation of PPDE, and after some trials was discovered not to work at all. The reason is quite fundamental. The approach is based on the analytic view of PPDA as sketched by the author in Section 11.3: iteratively construct a function, say $u(\mathbf{x})$, such that the true density $f(\mathbf{x})$, multiplied by this function, is equal to the initial guess $g_0(\mathbf{x})$:

$$f(\mathbf{x})u(\mathbf{x}) = g_0(\mathbf{x}).$$

In the paper, $u(\mathbf{x})$ is the reciprocal of a product of ridge functions, but this is incidental. The essential point is that such a function $u(\mathbf{x})$ exists only if the initial guess g_0 is absolutely continuous with respect to $f(\mathbf{x})$. We have no control over $f(\mathbf{x})$, and so we cannot make sure that $u(\mathbf{x})$ exists. On the other hand, we can pick a strictly positive initial guess $g_0(\mathbf{x})$ and thus make sure that an augmenting function $u(\mathbf{x})$ with

$$g_0(\mathbf{x})u(\mathbf{x}) = f(\mathbf{x})$$

does exist.

It is instructive to have a brief look at a situation where an implementation of the analytic view will run into problems. Suppose the sample consists of two isolated clusters, and our initial guess $g_0(\mathbf{x})$ is a multivariate Gaussian density with the same mean and covariance as the sample. Let us say we have found a direction $\mathbf{a}_1$ such that in the marginal along $\mathbf{a}_1$ the clusters are nicely separated. The marginal of the Gaussian of course will again be Gaussian, and its center will be somewhere between the modes of the data marginal. We now have to find a function $v(z)$ such that the marginals of $f(\mathbf{x})v(\mathbf{a}_1^T\mathbf{x})$ and $g_0(\mathbf{x})$ along $\mathbf{a}_1$ agree. In sample terms this means that we have to determine weights for the observations such that a kernel density estimate applied to the projections of the weighted observations is equal to the marginal of $g_0(\mathbf{x})$. This will be impossible because no observations project between the projections of the two clusters, and therefore no conceivable weighting will give a nonzero estimate (unless we choose the width of the kernel very large, in which case we end up with a highly biased estimate).

REFERENCES

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

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). To appear in *J. Amer. Statist. Assoc.* Also published as Technical Report Orion-010 (1982), Dept. of Statist., Stanford University.

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

FISHERKELLER, M. A., FRIEDMAN, J. H. and TUKEY, J. W. (1974). Prim-9: an interactive multidimensional data display and analysis system. SLAC-Pub-1408, Stanford Linear Accelerator Center, Stanford, California.

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

FRIEDMAN, J. H. and STUETZLE, W. (1981b). Projection pursuit methods for data analysis. In *Modern Data Analysis.* (A. F. Siegel and R. Launer, eds.) Academic, New York.

FRIEDMAN, J. H., STUETZLE, W. and SCHROEDER, A. (1984). Projection pursuit density estimation. *J. Amer. Statist. Assoc.* **79** 599–608. Also published as Stanford Technical Report Orion-002 (1981), Dept. of Statist., Stanford University.

HENRY, D. H. (1983). Multiplicative models in projection pursuit. SLAC-Report 261, Stanford Linear Accelerator Center, Stanford, California.

HUBER, P. J. (1975). Applications vs. abstraction: the selling out of mathematical statistics? *Suppl. Adv. Appl. Probab.* **7** 84–89.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195

PING CHENG AND C. F. J. WU

Academia Sinica, Beijing and University of Wisconsin, Madison

Professor Huber's stimulating paper has greatly advanced our knowledge of the projection pursuit methodology. Our discussion will be confined to the convergence of the projection pursuit density approximation method (PPDA). In Proposition 14.3 he proved the uniform and L_1-convergence of the PPDA by assuming that the density f can be deconvoluted with a Gaussian component. This is a very strong smoothness condition on f. Our original attempt was to prove his conjecture that the convergence still holds under more general smoothness condition on f. Failing this, we have instead found a smoothed version of the PPDA that converges uniformly and in L_1 to f with no smoothness condition required on f. Our modification is described as follows.

Let $\{g^{(k)}\}$ be the sequence of approximating densities defined in Proposition 14.3. Define the *smoothed* approximating density $\bar{g}^{(k)}$ by convoluting $g^{(k)}$ with a normal density

$$\bar{g}^{(k)} = g^{(k)} * N(0, \sigma_k^2 I_d), \tag{1}$$

where σ_k satisfies

$$\tau_k \sigma_k^{-2d} \to 0 \quad \text{and} \quad \sigma_k \to 0 \quad \text{as} \quad k \to \infty \tag{2}$$