Projection Pursuit Density Estimation

Jerome H. Friedman; Werner Stuetzle; Anne Schroeder

Journal of the American Statistical Association, Vol. 79, No. 387 (Sep., 1984), 599-608.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198409%2979%3A387%3C599%3APPDE%3E2.0.CO%3B2-I

Journal of the American Statistical Association is currently published by American Statistical Association.

Projection Pursuit Density Estimation

JEROME H. FRIEDMAN, WERNER STUETZLE, and ANNE SCHROEDER*

The projection pursuit methodology is applied to the multivariate density estimation problem. The resulting nonparametric procedure is often less biased than the kernel and near-neighbor methods. In addition, graphical information is produced that can be used to help gain geometric insight into the multivariate data distribution.

KEY WORDS: Density estimation; Projection pursuit; Nonparametric methods.

1. INTRODUCTION

The formal goal of nonparametric density estimation is to estimate the probability density of a p-dimensional random vector $\mathbf{X} \in R^p$ on the basis of iid observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ without making the assumption that the density belongs to a particular parametric family. Often in practice, a more important objective is to gain geometric insight into the data distribution in R^p.

Nonparametric estimation of univariate probability density functions has been extensively studied. Examples of successful methods are the related techniques of kernel estimates (Parzen 1962; Rosenblatt 1971), near-neighbor estimates (Loftsgaarden and Quesenberry 1965), and splines (Boneva, Kendall, and Stefanov 1971). A good overview was given by Tapia and Thompson (1978). The direct extension of these methods to multivariate settings, however, has not been as successful in practice. This can be attributed partly to their deteriorating statistical performance, caused by the so-called "curse of dimensionality" (Bellman 1961), which requires very large spans (radii of neighborhoods) to achieve sufficient counts. The resulting estimates are then highly biased. In addition, these methods do not provide any comprehensible information about the structure of the multivariate point cloud.

Our approach to multivariate density estimation is based on the notion of projection pursuit (Friedman and Tukey 1974; Friedman and Stuetzle 1981). It attempts to overcome the curse of dimensionality by extending the classical univariate density estimation methods to higher dimensions in a manner that involves only univariate estimation. As a by-product, graphical information is produced that can be helpful in exploring and understanding the multivariate data distribution.

* Jerome H. Friedman is Professor and Werner Stuetzle is Professor, Statistics Department and Linear Accelerator Center, Stanford University, Stanford, CA 94305. Anne Schroeder is a staff member, Institut National de Recherche en Informatique et Automatique, Le Chesnay, France. The research for this article was supported by Department of Energy Contracts DE-AC-03-76F00515 and DE-AT03-81-ER10843, Office of Naval Research Contract ONR-N-00014-81-K-0340, and Army Research Office Contract DAAG29-82-K-0056.

2. OVERVIEW

The goal of projection pursuit methods is to estimate multivariate functions by combinations of smooth univariate (ridge) functions of carefully selected linear combinations of the variables.

Our projection pursuit density estimation (PPDE) method constructs estimates of the form

$$p_M(\mathbf{x}) = p_0(\mathbf{x}) \prod_{m=1}^{M} f_m(\theta_m \cdot \mathbf{x}), \tag{1}$$

where p_M is the density estimate (or current model) after M iterations of the procedure; p_0 is a given multivariate density function to be used as the initial model; θ_m is a unit vector specifying a direction in R^p, so $\theta_m \cdot \mathbf{x} = \sum_{i=1}^{p} \theta_{mi} x_i$ is a linear combination of the original coordinate measurements; and f_m is a univariate function.

From (1) PPDE is seen to approximate the multivariate density by an initially proposed density p_0, multiplied (augmented) by a product of univariate functions f_m of linear combinations $\theta_m \cdot \mathbf{x}$ of the coordinates. The choice of the initial density is left to the user and should reflect his best a priori knowledge of the data. A Gaussian density with sample mean and sample covariance matrix is often a natural choice. The purpose of PPDE is to choose the directions θ_m and construct the corresponding functions $f_m(\theta_m \cdot \mathbf{x})$. The product of these functions estimates the ratio of the data density to the initial model density.

From (1) we obtain the recursion relation

$$p_M(\mathbf{x}) = p_{M-1}(\mathbf{x}) f_M(\theta_M \cdot \mathbf{x}). \tag{2}$$

Since f_M is used to modify p_{M-1} to obtain p_M, we refer to the f_m as *augmenting functions*.

The recursive definition of model (2) suggests a stepwise approach for its construction. At the Mth iteration, there is a current model $p_{M-1}(\mathbf{x})$ constructed in the previous steps. (For the first step, $M = 1$, the current model is the initial model $p_0(\mathbf{x})$ specified by the user.) Given $p_{M-1}(\mathbf{x})$, we seek a new model $p_M(\mathbf{x})$ to serve as a better approximation to the data density $p(\mathbf{x})$. Thus a direction θ_M and its corresponding augmenting function $f_M(\theta_M \cdot \mathbf{x})$ are chosen to maximize the goodness of fit of $p_M(\mathbf{x})$. We measure relative goodness of fit by the cross-entropy term of the Kullback–Leibler distance

$$W = \int \log p_M(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \tag{3}$$

From (2) we see that W achieves its maximum at the same location as

$$w(\theta_M, f_M) = \int \log f_M(\theta_M \cdot \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (4)$$

Equation (4) is to be maximized under the constraint that $p_M(\mathbf{x})$ be properly normalized, that is, $\int p_M(\mathbf{x})d\mathbf{x} = 1$. For a given direction θ_M and known $p(\mathbf{x})$,

$$f_M(\theta_M \cdot \mathbf{x}) = p^{\theta_M}(\theta_M \cdot \mathbf{x})/p_{M-1}{}^{\theta_M}(\theta_M \cdot \mathbf{x}) \quad (5)$$

is seen to maximize (4). Here p^{θ_M} and $p_{M-1}{}^{\theta_M}$ represent the data and current model marginal densities along the (one-dimensional) subspace spanned by θ_M. Using this f_M for given θ_M, it remains to find the direction θ_M for which (4) achieves the maximum value. The optimal θ_M and its corresponding augmenting function $f_M(\theta_M \cdot \mathbf{x})$ define the new model through (2).

In actual applications the data density $p(\mathbf{x})$ is unknown. We have, instead, a sample of N iid observations $x_1, x_2, \ldots, x_N$ from $p(\mathbf{x})$. The cross-entropy W is estimated by the log-likelihood

$$\hat{W} = \frac{1}{N}\sum_{i=1}^{N} \log p_M(\mathbf{x}_i). \quad (6)$$

Analogously, $w(\theta_M, f_M)$ is estimated by

$$\hat{w}(\theta_M, f_M) = \frac{1}{N}\sum_{i=1}^{N} \log f_M(\theta_M \cdot \mathbf{x}_i), \quad (7)$$

where $f_M(\theta_M \cdot \mathbf{x})$ is an estimate for the ratio of data and model marginals along θ_M. The optimal value θ_M that maximizes $\hat{w}(\theta_M, f_M)$, and thus the loglikelihood $\hat{W}$ of the new model, is determined by numerical optimization.

3. ESTIMATION PROCEDURES

We now discuss the estimation of $f(\theta \cdot \mathbf{x})$, the ratio of data and model marginals along a direction θ. First consider the current model marginal $p_{M-1}{}^{\theta}(\theta \cdot \mathbf{x})$. Without loss of generality, we let θ be the first coordinate axis, that is, $\theta \cdot \mathbf{x} = x_1$. Then

$$p_{M-1}{}^{\theta}(x_1) = \int p_{M-1}(\mathbf{x}) dx_2 dx_3 \cdots dx_n. \quad (8)$$

If $p_{M-1}{}^{\theta}(x_1)$ is continuous, then

$$p_{M-1}{}^{\theta}(x_1) = \lim_{h\to 0} \frac{1}{2h}\int_{x_1-h}^{x_1+h} p_{M-1}{}^{\theta}(z)dz \quad (9)$$

$$= \lim_{h\to 0} \frac{1}{2h}\int_{-\infty}^{\infty} I(x_1 - h \le z \le x_1 + h) p_{M-1}{}^{\theta}(z)dz, \quad (10)$$

where

$$I(s) = 1 \quad \text{if } s \text{ is true}$$
$$= 0 \quad \text{otherwise.} \quad (11)$$

From (8) one has

$$p_{M-1}{}^{\theta}(x_1) = \lim_{h\to 0} \frac{1}{2h}\int I(x_1 - h \le y_1 \le x_1 + h) p_{M-1}(\mathbf{y})d\mathbf{y}$$

$$= \lim_{h\to 0} \frac{1}{2h} E_{p_{M-1}}[I(x_1 - h \le y_1 \le x_1 + h)]. \quad (12)$$

Our estimate of $p_{M-1}{}^{\theta}(x_1)$ is obtained from (12) by using a small finite value for h and employing a Monte Carlo method to estimate the expected value. A Monte Carlo sample $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{N_s}$, of size N_s is generated with density $p_{M-1}(\mathbf{x})$, and

$$\hat{p}_{M-1}{}^{\theta}(x_1) = \frac{1}{2hN_s}\sum_{j=1}^{N_s} I(x_1 - h \le y_{j1} \le x_1 + h) \quad (13)$$

is taken as our estimate of $p_{M-1}{}^{\theta}(x_1)$. Since the choice of x_1 as the direction θ was arbitrary, (13) can be written equally as

$$\hat{p}_{M-1}{}^{\theta}(\theta \cdot \mathbf{x}) = \frac{1}{2hN_s}\sum_{j=1}^{N_s} I(\theta \cdot \mathbf{x} - h \le \theta \cdot \mathbf{y}_j \le \theta \cdot \mathbf{x} + h) \quad (14)$$

for any θ. Note that the same Monte Carlo sample can be used for all θ and $\mathbf{x}$. In Appendix B, we discuss in detail procedures for generating a Monte Carlo sample from the density $p_{M-1}(\mathbf{x})$.

By assumption, the data represent a sample from $p(x)$ that can be used, in analogy with (14), to estimate the data marginal $p^{\theta}(\theta \cdot \mathbf{x})$ by

$$\hat{p}^{\theta}(\theta \cdot \mathbf{x}) = \frac{1}{2hN}\sum_{i=1}^{N} I(\theta \cdot \mathbf{x} - h \le \theta \cdot \mathbf{x}_i \le \theta \cdot \mathbf{x} + h). \quad (15)$$

From (5) our estimate of the augmenting function becomes

$$f_{\theta}(\theta \cdot \mathbf{x}) = \frac{N_s \sum_{i=1}^{N} I(\theta \cdot \mathbf{x} - h \le \theta \cdot \mathbf{x}_i \le \theta \cdot \mathbf{x} + h)}{N \sum_{j=1}^{N_s} I(\theta \cdot \mathbf{x} - h \le \theta \cdot \mathbf{y}_j \le \theta\, \mathbf{x} + h)}. \quad (16)$$

This is just the ratio of the fraction of observation counts to the fraction of Monte Carlo counts in an interval of width $2h$ centered at $\theta \cdot \mathbf{x}$. To help stabilize the denominator, we choose h to always include exactly αN_s Monte Carlo observations. In this case (16) becomes

$$f_{\theta}(\theta \cdot \mathbf{x}) = \frac{1}{\alpha N}\sum_{i=1}^{N} I(\theta \cdot \mathbf{x} - h \le \theta \cdot \mathbf{x}_i \le \theta \cdot \mathbf{x} + h). \quad (17)$$

The fraction α is called the span; it is a parameter of the procedure. In actual applications the span can be adjusted

based on visual inspection of the augmenting functions f_m and the histogram estimates of the marginals of data and model along the directions θ_m. The binwidths of the histograms should be chosen small so that they estimate the marginals without much bias. The goal is to pick as large a span as possible, which will make the augmenting functions f_m smooth, subject to the constraint that the histograms of data and model along each of the directions should not differ systematically.

4. REDUNDANT-VARIABLE ELIMINATION

For purposes of interpretation, it is desirable that models be parsimonious. That is, models should involve only as many variables as are required for an adequate description of the data. For models constructed by projection pursuit, this means that each solution linear combination θ_M should involve only those predictor variables that are necessary. Because of sampling fluctuations, it often happens that several variables enter into a solution linear combination (usually with small coefficients), but their removal will not substantially affect the quality of the solution. This is especially true if some of the variables are highly correlated.

Redundant variables entering into a solution linear combination θ_M can be eliminated by the following (reverse) stepwise procedure. Each nonzero coefficient is in turn set to zero, keeping all other coefficients at their solution values; the corresponding augmenting function is computed and the log-likelihood is obtained. That variable x_U for which this (deleted) log-likelihood W_U is largest becomes a candidate for elimination. Let x_L be the variable for which the deleted log-likelihood W_L is smallest, and let W_C be the log-likelihood for the complete solution (no variables deleted). If

$$W_C - W_U > \beta(W_C - W_L), \tag{18}$$

then the elimination procedure stops and the complete solution is accepted. Otherwise x_U is deleted (coefficient set to zero) and the above procedure is repeated (next iterative pass) for all variables with nonzero coefficients. This iterative procedure terminates when the candidate variable for an iteration x_U cannot be delected (i.e., when (18) is true). The quantity β is a user-specified parameter.

5. TERMINATION CRITERIA

As with any stepwise procedure, one needs a criterion for stopping the iteration at some (Mth) step. Stopping too soon can increase the bias of the estimator, and not stopping soon enough can unduly increase its variance. Optimal termination of stepwise procedures has been studied (see Stone 1974 and references therein); these methods can be applied here. In practice, stepwise procedures are often terminated subjectively, based on an inspection of successive values of the goodness-of-fit criterion.

PPDE can provide several additional aids in judging whether a new step enhances the model enough to be included. One can compare $p_{M-1}{}^{\theta_M}(\theta_M \cdot \mathbf{x})$ (the current model marginal along θ_M) with $p^{\theta_M}(\theta_M \cdot \mathbf{x})$ (the actual data marginal along θ_M). The ratio of these two densities would be the Mth augmenting function. Since θ_M is chosen to maximize (in the likelihood sense) the difference between data and model marginals, their comparison in this projection represents a genuine comparison of the full multivariate densities for quality. Our experience indicates that graphical comparisons are most effective.

Graphical inspection of $f_M(\theta_M \cdot \mathbf{x})$ can also be used to judge whether it should be included in the model. If the graph of $f_M(\theta_M \cdot \mathbf{x})$ versus $\theta_M \cdot \mathbf{x}$ displays a noisy pattern with no systematic tendency, then its inclusion will likely only increase the variance of the density estimate. On the other hand, a definite tendency indicates that $f_M(\theta_M \cdot \mathbf{x})$ is dealing with a genuine inadequacy of the present model.

6. EXPRESSION OF THE RESULTS

From the formal point of view, the result of applying PPDE is an estimate of the data density specified by the initial model, a series of directions (unit vectors) $\theta_m \in R^p$, and a corresponding set of augmenting functions $f_m(\theta_m \cdot \mathbf{x})$. The augmenting functions can be stored as specific values associated with each observation. Because of their inherent smoothness, however, this representation is highly redundant and a bit cumbersome. For this reason, we approximate each augmenting function f_m by a cubic spline function

$$s_m(z) = \sum_{l=1}^{L} \beta_{lm} B_l(z). \tag{19}$$

The $B_l(z)$ are basic cubic B-splines (de Boor 1978), and the β_{lm} are determined by a least squares fit of s_m to the observations $(\theta_m \cdot \mathbf{x}_i, f_m(\theta_m \cdot \mathbf{x}_i))$, $i = 1, \ldots, N$. The number of knots L is chosen to be inversely proportional to the span α (17). The internal knots are placed such that equal numbers of Monte Carlo observations fall between each pair.

7. EXAMPLES

We illustrate the use of PPDE by applying it to three examples. In all examples except the last, the initial model $p_0(\mathbf{x})$ was taken to be Gaussian with sample mean and sample covariance matrix, the logarithm of the likelihood of the data sample under the initial model was arbitrary set to zero, and the size of the Monte Carlo sample was taken to be twice the data sample size.

The first example is especially simple and was chosen primarily to illustrate the functioning of the algorithm. In it 225 observations were generated in two dimensions from a uniform mixture of three Gaussian distributions, with unit covariance matrices and centers at the vertices of an equilaterial triangle of sidelength six. Figure 1 shows the true density function. Because the data for this example are only two-dimensional, it is possible to monitor

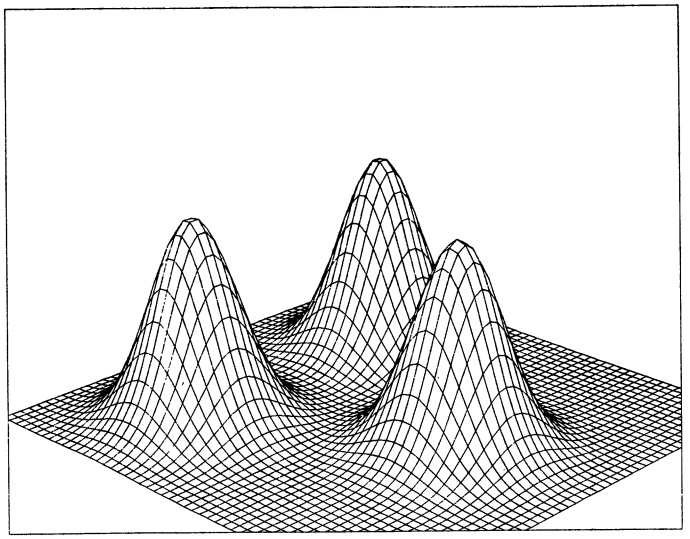

Figure 1. True Density Function—Gaussian Mixture.

the progress of the PPDE procedure as it attempts to iteratively construct the two-dimensional density from one-dimensional projections.

Figure 2 shows the initial model approximation $p_0(\mathbf{x})$, a Gaussian with sample mean and sample covariance matrix. Figure 3a shows superimposed histograms of the data (···) and the Monte Carlo sample drawn from $p_0(\mathbf{x})$ (X) as projected onto the first solution linear combination $\theta_1 = (.0, 1.0)$, and Figure 3b shows the first augmenting function $f_1(\theta_1 \cdot \mathbf{x})$. (Note that augmenting functions are not well defined and thus can behave strangely in regions where both the data density and current model estimate are both very small. This has no effect on the quality of the resulting density estimate.) The estimate after the first iteration (Figure 3c) achieves an increase in log-likelihood of 62.8. Taken by itself this number does not mean much, but it is useful to compare the relative effects of adding additional terms to the model.

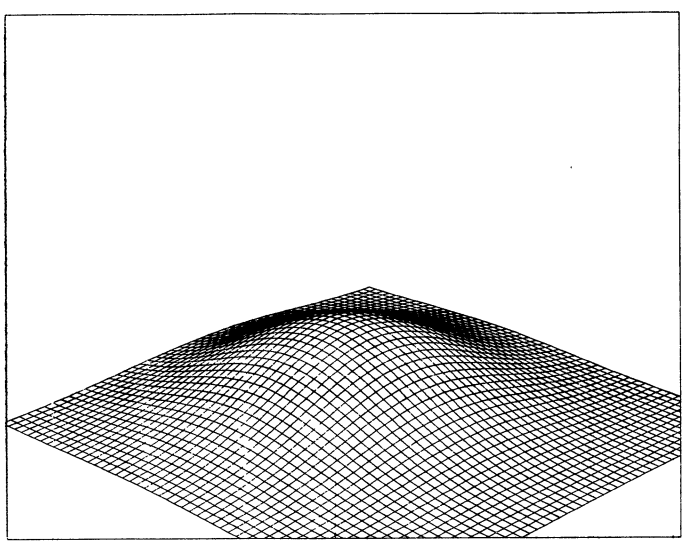

Figure 2. Initial Model $p_0(x)$—Gaussian With Sample Mean and Sample Covariance.

Figure 4a shows the data and a Monte Carlo sample drawn from $p_1(\mathbf{x})$ as projected on the second solution linear combination $\theta_2 = (.87, .49)$. Figure 4b shows the second augmenting function $f_2(\theta_2 \cdot \mathbf{x})$ plotted versus $\theta_2 \mathbf{x}$. The increase in log-likelihood of 75.7 and Figure 4 a b indicate that $p_2(\mathbf{x})$ is a substantial improvement over $p_1(\mathbf{x})$. This is confirmed by Figure 4c, which illustrates $p_2(\mathbf{x})$. The three-peak structure of $p(\mathbf{x})$—the true data density—is now reproduced in the estimate $p_2(\mathbf{x})$.

Figure 5a shows the superimposed histograms of data and Monte Carlo as projected onto the third solution direction $\theta_3 = (.89, -.45)$. The model marginal is systematically too low in the left peak and too high in the valley. Figure 5b shows the augmenting function for the third term. Figure 5 a and b and the increase in log-likelihood of 43.3 indicate that inclusion of a third term should improve the estimate. Indeed it does improve, as demonstrated in Figure 5c.

The second example is designed to lend some credibility to the claim that PPDE will outperform standard, nonadaptive density estimation methods in certain situations. We compare PPDE to the k nearest neighbor density estimator (KNNE). Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be an iid sample from some unknown density $p(\mathbf{x})$. KNNE estimates the density at some point $\mathbf{x}_0$ by

$$\hat{p}(\mathbf{x}_0) = k/n\text{vol}_k(\mathbf{x}_0), \tag{20}$$

where $\text{vol}_k(\mathbf{x}_0)$ is the volume of the smallest sphere centered at $\mathbf{x}_0$ and containing k observations. The number k of near neighbors is a parameter of the estimator to be chosen by the user; it controls the trade-off between bias and variance of the estimate. In our comparisons we always picked the optimal value of k, the value that makes KNNE work best. Note that this, of course, is impossible to do in practice when the true underlying density is unknown.

We compare PPDE and KNNE for three different scenarios. The first one is the same as in Example 1: the true density is taken to be a uniform mixture of three standard Gaussians, centered at the vertices of an equilateral triangle of sidelength six. In the second scenario, the data are five-dimensional. The distribution in the first two dimensions is exactly the same as before; the remaining three variables have independent Gaussian distributions with standard deviation 3. The structure of the density thus lies in the first two variables; the three remaining variables only add noise. The noise standard deviation was chosen to give all five variables roughly the same variance. The third scenario is essentially the same as the second, but there are seven noise dimensions instead of three and the data, therefore, are 10-dimensional. In all three scenarios the sample size is $N = 225$.

An important question concerns what measure of distance between true and estimated density to use when comparing the estimators. We want the measure to reflect how well the estimators can reproduce the main feature of the underlying density, that is, the presence of three clusters. The Kullback–Leibler distance would be an ob-

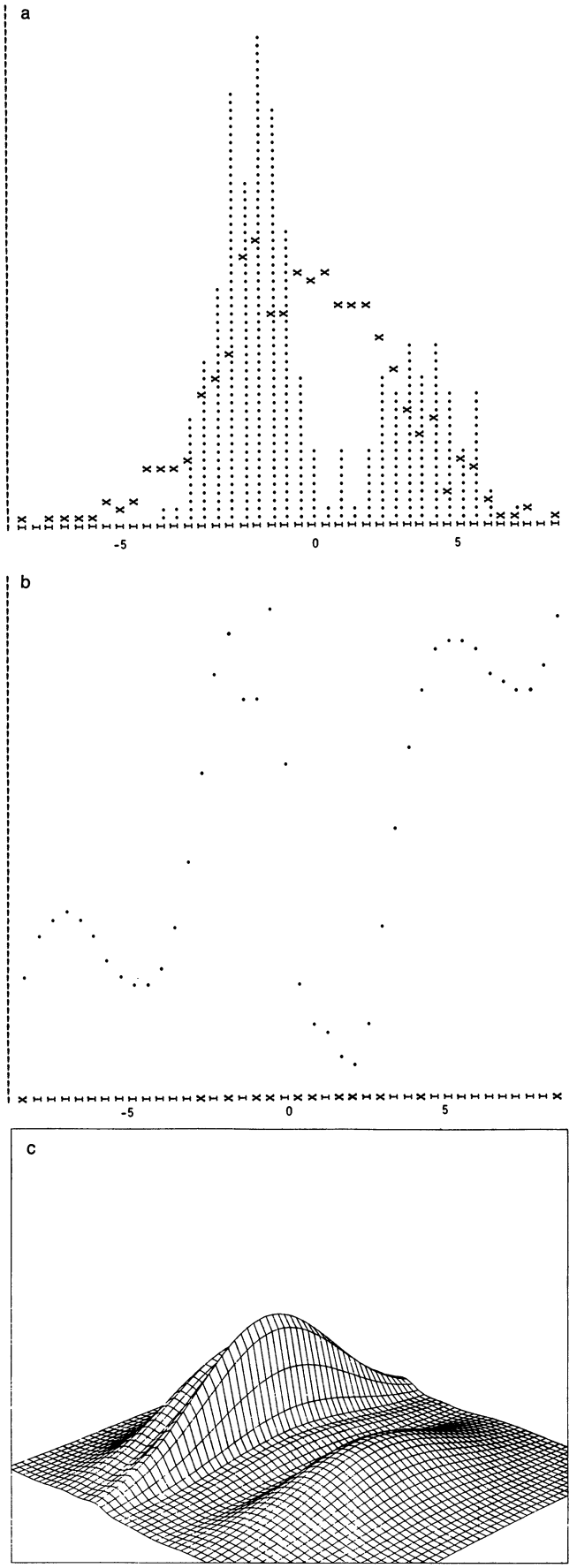

Figure 3. PPDE Estimate—First Iteration. (a) Data (···) and model (X) marginals along θ_1 = (0, 1); (b) first augmenting function $f_1(\theta_1 \cdot x)$; (c) model after the first iteration.

vious choice, but it overemphasizes differences in the tails, where the KNNE cannot be expected to do very well. We chose to base our comparison on the expected squared error, integrated over a region E in space containing most of the mass of the true underlying density:

$$\text{EISE} = \int_E E(\hat{p}(\mathbf{x}) - p(\mathbf{x}))^2 \, d\mathbf{x}. \tag{21}$$

To ease interpretation, we do not report EISE but, rather, the expected percentage of variance explained.

$$\text{PVE} = 100\,(1 - \text{EISE/var}), \tag{22}$$

where

$$\text{var} = \int_E (p(\mathbf{x}) - \bar{p})^2 \, d\mathbf{x} \tag{23}$$

and

$$\bar{p} = \frac{1}{\text{vol}(E)} \int_E p(\mathbf{x}) \, d\mathbf{x}. \tag{24}$$

We chose the region E to be elliptical with principal axis along the coordinate directions. The lengths of the principal axis are 12 for coordinates 1 and 2, and 15 for the noise coordinates.

One small problem that we have ignored so far is normalization of the KNNE estimate. We are obviously not interested in measuring error that is simply due to the KNNE estimate's not being normalized properly. For each of the three scenarios described previously, we thus multiply the KNNE estimates with a factor chosen to minimize the expected integrated squared error. This factor of course can be found only in a simulation situation in which the true underlying density is known. Table 1 shows the results of our simulation. The pattern is quite clear. The two estimators have about the same performance for scenario 1. The performance of KNNE deteriorates completely, however, as noise variables are added; it explains only 9% of the variance in 10 dimensions. The performance of PPDE is influenced to a much lesser extent; it still explains 63% of the variance in 10 dimensions.

The difference in performance between KNNE and PPDE for the ten-dimensional problem is further illustrated by Figure 6 a–h. Each figure shows the cross-section of a density in the plane spanned by the first two coordinate axes, that is, $p(x, y, 0, \ldots, 0)$ plotted as a function of x and y. Figure 6 a–d shows the cross-sections of the PPDE estimates for the first four of our Monte Carlo trials; Figure 6 e–h shows the corresponding cross-sections of the KNNE estimates. We note that in each case the PPDE estimate reproduces the salient structure, whereas the KNNE estimate has little to do with reality.

The next example illustrates the use of PPDE in a purely data-analytic setting. For this example, we used data from the diebetes study of Reaven and Miller (1979). For each of 145 subjects in the study, five variables were measured: relative weight, a measure of glucose tolerance, a second measure of glucose tolerance (glucose area), a measure of insulin secretion (insulin area), and a measure of how glucose and insulin interact (SSPG).

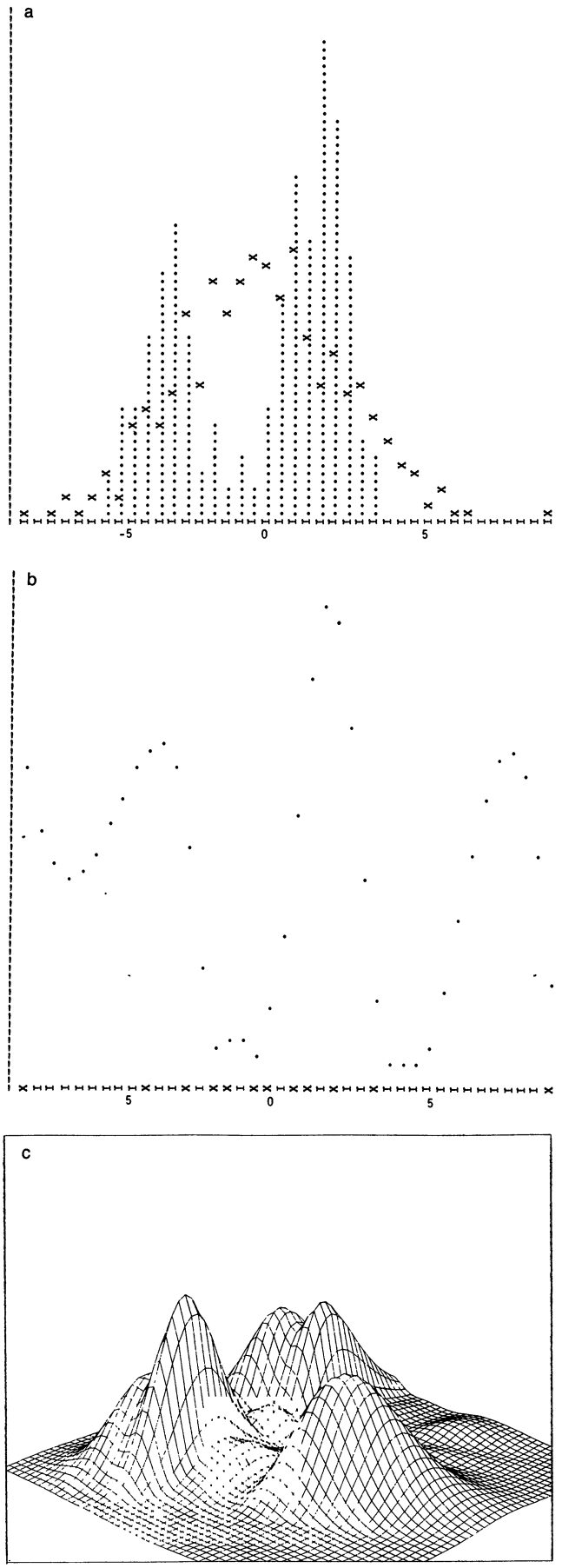

Figure 4. PPDE Estimate—Second Iteration. (a) Data and model marginals along θ_2 = (.87, .49); (b) second augmenting function $f_2(\theta_2 \cdot x)$; (c) model after the second iteration.

The two measures of glucose tolerance exhibit a high degree of linear association (r = .96), so the first one is not considered.

Our purpose with this example is to see how well the four-dimensional data density $p(\mathbf{x})$ can be represented as a product of two two-dimensional marginal densities $p_{ab}(x_a, x_b)\, p_{cd}(x_c, x_d)$. If the data density could be factored into such a product for a specific pairing of variables, (ab) (cd), then all of the information about the data structure in the full four-dimensional space would be contained in the two scatterplots—variable a versus variable b and variable c versus variable d.

Unlike in the previous examples, the initial model is not explicitly defined; it is given by the factored approximation

$$p_0(\mathbf{x}) \equiv p_{ab}(x_a, x_b)p_{cd}(x_c, x_d), \tag{25}$$

with a specific choice for the variables a, b, c, and d. The two-dimensional marginal densities in (25) are taken to be the actual data as projected onto the subspaces spanned by (x_a, x_b) and (x_c, x_d), respectively. Since it is not our purpose to provide explicit density estimates, it is not necessary to have an explicit (computable) representation for $p_0(\mathbf{x})$. All that is necessary is that we be able to draw a sample from it. Such a sample of size N (here N = 145) is generated by randomly permuting the observation labels of the $(x_a x_b)$ pairs with respect to the $(x_c x_d)$ pairs. Let $(r_1, r_2, \ldots, r_N)$ be a random permutation of the integers $(1, 2, \ldots, N)$. The Monte Carlo sample from the initial model is taken to be the four-tuples

$$x_{1a}\, x_{1b} x_{r1c}\, x_{r1d}$$

$$x_{2a}\, x_{2b}\, x_{r2c}\, x_{r2d}$$

$$\vdots$$

$$x_{Na}\, x_{Nb}\, x_{rNc}\, x_{rNd}.$$

As many Monte Carlo observations as needed can be obtained by repeating this procedure with different random permutations.

Table 2 shows the increase in log-likelihood achieved by PPDE, after two and four iterations, starting with the three different initials models (25) specified by the three distinct groupings (a, b), (c, d). It is clear that the least improvement was associated with

$$p_0(\mathbf{x}) = p_{13}(x_1, x_3)p_{24}(x_2, x_4), \tag{26}$$

indicating that this factorization gives the best representation of the actual data density. Figure 7a shows a scatterplot of x_1 versus x_3, and Figure 7b shows x_2 versus x_4 for the data sample. The results in Table 2 indicate that this is the best pair of plots in which to view the four-dimensional data structure. Starting with an initial density equal to the product of the two (two-dimensional) densities shown in Figure 7 a and b, however, PPDE was able to construct a model with substantially greater likelihood (see Table 2), indicating that Figure 7 a and b does not reveal all of the data structure in the full four-dimensional space.

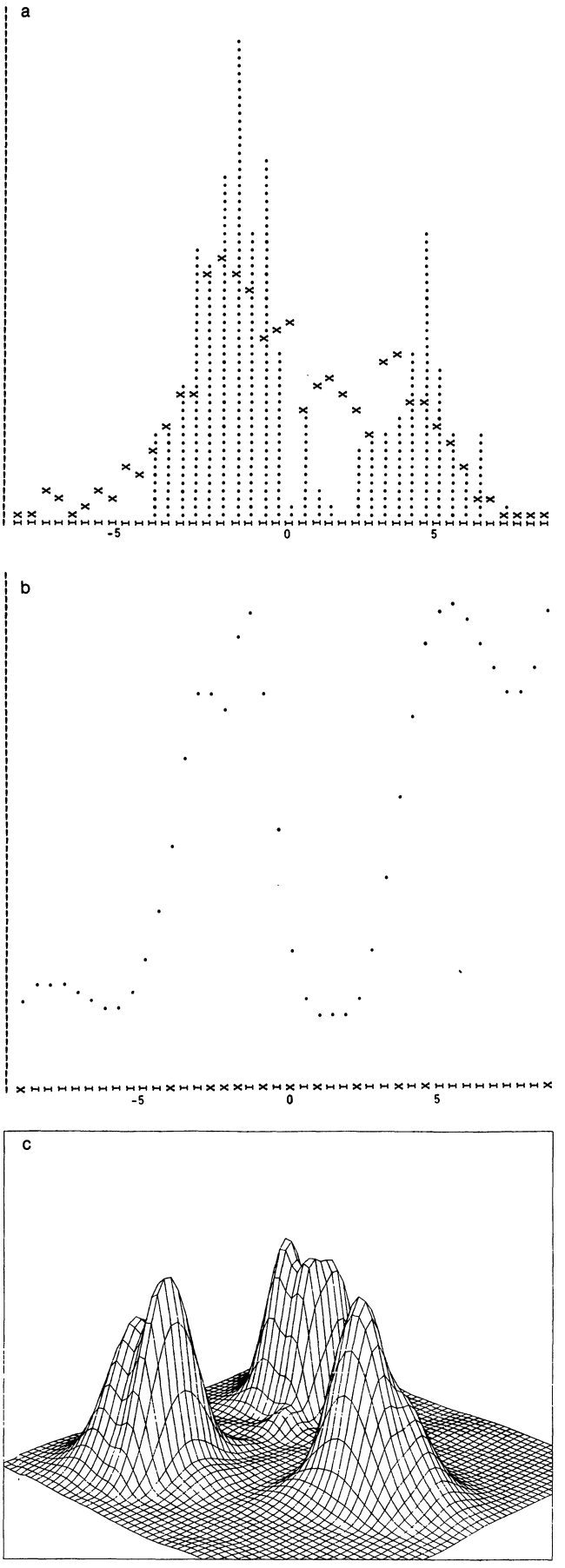

Figure 5. PPDE Estimate—Third Iteration. (a) Data and model marginals along θ_3 = (.89, −.45); (b) third augmenting function $f_3(\theta_3 \cdot x)$; (c) model after the third iteration.

Table 1. Percentage of Variance Explained by PPDE and KNNE Estimators

	PPDE		KNNE	
Dimensions, p	Monte Carlo Estimate (%)	Standard Deviation	Monte Carlo Estimate (%)	Standard Deviation
2	79	3	80	1
5	69	7	42	2
10	63	5	9	2

This is verified in Figure 7c, where the 145 data points (·) and 145 Monte Carlo points (+'s) drawn from $p_0(\mathbf{x})$ (26) are shown projected on the plane spanned by the first two solution directions, $\theta_1 = (-.29, -.37, .38, .80)$ and $\theta_2 = (-.08, .92, -.37, -.11)$. The horizontal axis is $\theta_1 \cdot \mathbf{x}$ and the vertical axis is $(\theta_2 - (\theta_2 \cdot \theta_1)\theta_1) \cdot \mathbf{x} / \| \theta_2 - (\theta_2 \cdot \theta_1)\theta_1 \|$. The data have roughly the same shape as the factored approximation (26) but are more tightly concentrated, especially in the circular ball centered at (0, 0).

8. DISCUSSION

As a formal estimator of a multivariate density function, PPDE shares advantages common to projection pursuit procedures. Since all estimation is carried out in a univariate setting, the high bias inherent in other multivariate nonparametric density estimators can often be avoided. The PPDE estimate is given in a concise functional form, (1) and (19); and it can be graphically represented. The graphical representation can be used to adjust the main parameters of the procedure, the span α and the number M of terms in the model.

Bias is encountered with PPDE when many terms are required to provide a good representation of the true data density, but only a few can be estimated because of insufficient sample size. In these cases it is important that the first few terms be able to approximate a wide variety of functions so that the most salient features of the data density can be modeled. In the limit $M \to \infty$, any density function can be represented by (1) (for any strictly positive p_0), but even for moderate M, functions of this form constitute a rich class. In addition, the choice of initial model p_0 permits the user to introduce any knowledge he may have concerning the density function, thereby allowing a further reduction in bias.

The success of PPDE will, of course, depend on the particular nature of the actual data density. Examples of

Table 2. Increase in (Log) Likelihood of PPDE Solutions From Factored Initial Model $p_o(\mathbf{x}) = p_{ab}(x_a, x_b)p_{cd}(x_c, x_d)$ (third example)

Combination				Number of Iterations	
a	b	c	d	2	4
1	2	3	4	85.7	124.1
1	3	2	4	47.1	76.3
1	4	2	3	85.8	122.1

density functions requiring large M in (1) are those with highly concave isopleths or spherically nested isopleths of the same density value (unless, of course, this structure is anticipated and incorporated in the initial model p_0).

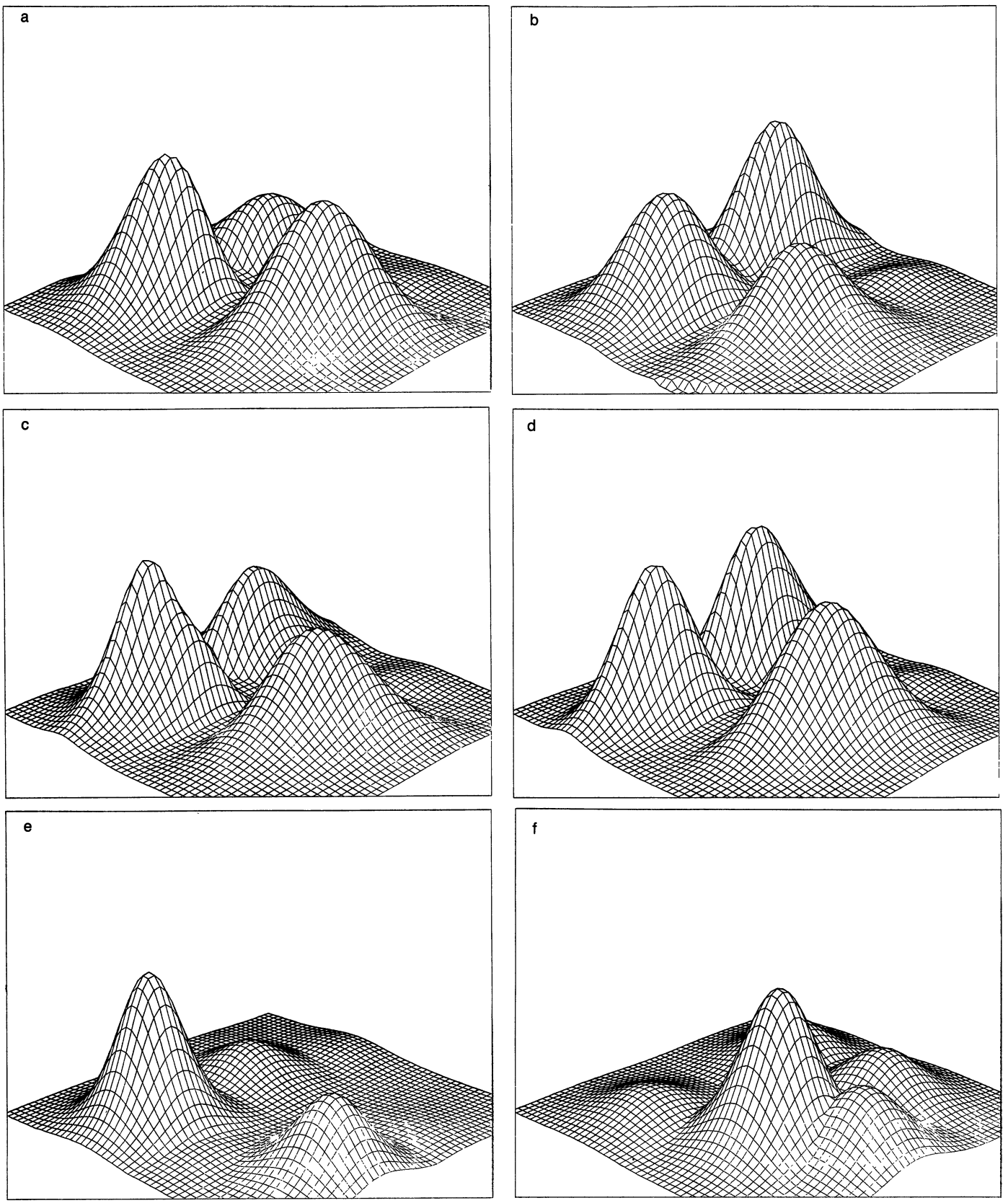

Figure 6 (continued on p. 607).

APPENDIX A: BACK ADJUSTMENT OF THE AUGMENTING FUNCTIONS

The basic iterative procedure described in Section 2 is (using the language of linear regression) stagewise in that each θ_M and its corresponding f_M are chosen as the solution to an optimization problem, holding all previous θ_m and $f_m (m < M)$ fixed. It is sometimes possible to improve the goodness of fit of the model by refitting all $f_m (m \leq M)$ after a new term is included in the model. This is done in an iterative manner similar to the basic (outer) iteration procedure except that the directions $\theta_m (1 \leq m \leq M)$ are held fixed to avoid the (costly) numerical optimization. At each stage m of this inner iterative procedure, $f_m(\theta_m \cdot \mathbf{x})$ is readjusted to maximize the log-likelihood (3) given all $f_j(\theta_j \cdot \mathbf{x}), j \neq m$. One complete pass through this inner iteration produces a new set of augmenting functions, which comprise a model with (possibly) higher likelihood. Since this pass has changed each f_m, it is possible that yet another pass can increase the likelihood still further. Thus the passes themselves are iterated until no increase in likelihood is observed.

We now discuss the calculation of a new $f_m(\theta_m \cdot \mathbf{x})$, given the $f_j(\theta_j \cdot \mathbf{x})$ for $j \neq m$. Let

$$p_{(m)}(\mathbf{x}) = p_0(\mathbf{x}) \prod_{j \neq m} f_j(\theta_j \cdot \mathbf{x}) = p_M(\mathbf{x}) / f_m(\theta_m \cdot \mathbf{x}). \quad \text{(A.1)}$$

We seek a new function $f_m'(\theta_m \cdot \mathbf{x})$ that maximizes

$$\hat{w}(\theta_m, f_m') = \int \log f_m'(\theta_m \cdot \mathbf{x}) p(\mathbf{x})\, d\mathbf{x} \quad \text{(A.2)}$$

subject to the constraint

$$\int p_{(m)}(\mathbf{x}) f_m'(\theta_m \cdot \mathbf{x})\, d\mathbf{x} = 1. \quad \text{(A.3)}$$

The solution is

$$f_m'(\theta_m \cdot \mathbf{x}) = p^{\theta_m}(\theta_m \cdot \mathbf{x}) / p_{(m)}{}^{\theta_m}(\theta_m \cdot \mathbf{x}), \quad \text{(A.4)}$$

where the numerator and denominator represent the corresponding marginal densities. These marginal densities are estimated as described in Section 3. The resulting estimate for f_m' then replaces f_m in the new model $p_M(\mathbf{x})$.

APPENDIX B: MONTE CARLO SAMPLING

To apply PPDE, it is necessary to draw a Monte Carlo sample from both the initial model $p_0(\mathbf{x})$ and the current model $p_{M-1}(\mathbf{x})$. For many choices of $p_0(\mathbf{x})$, there exist special algorithms that allow efficient direct sampling (see, e.g., Rubinstein 1981). Densities for which this is not the case can be sampled using the accept/reject method (Kronmal and Peterson 1981).

Suppose a Monte Carlo sample drawn from a density $q(\mathbf{x})$ is available and one wishes a sample drawn from $p(\mathbf{x})$. Let $r(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$ and

$$\gamma = \max_{\mathbf{x}} r(\mathbf{x}). \quad \text{(B.1)}$$

Consider each Monte Carlo observation $\mathbf{x}_i$ in turn. Draw a uniform random number u_i in the interval [0, 1]. If $u_i \gamma \leq r(\mathbf{x}_i)$, accept the ith observation; otherwise reject it. The accepted Monte Carlo observations will be a random sample drawn from $p(\mathbf{x})$.

This accept/reject method can be used to draw a random sample from any density $p(\mathbf{x})$. The efficiency of the procedure (number accepted, divided by number accepted plus number rejected) will be greater the more closely $q(\mathbf{x})$ resembles $p(\mathbf{x})$ in the sense of low variability of $p(\mathbf{x})/q(\mathbf{x})$.

The form of $p_{M-1}(\mathbf{x})$ ((1) and (2)) permits reasonably efficient sampling by the accept/reject method. First, the random sample drawn from $p_{M-2}(\mathbf{x})$, available from the previous iteration, is considered. From (1) and (A.4), we have

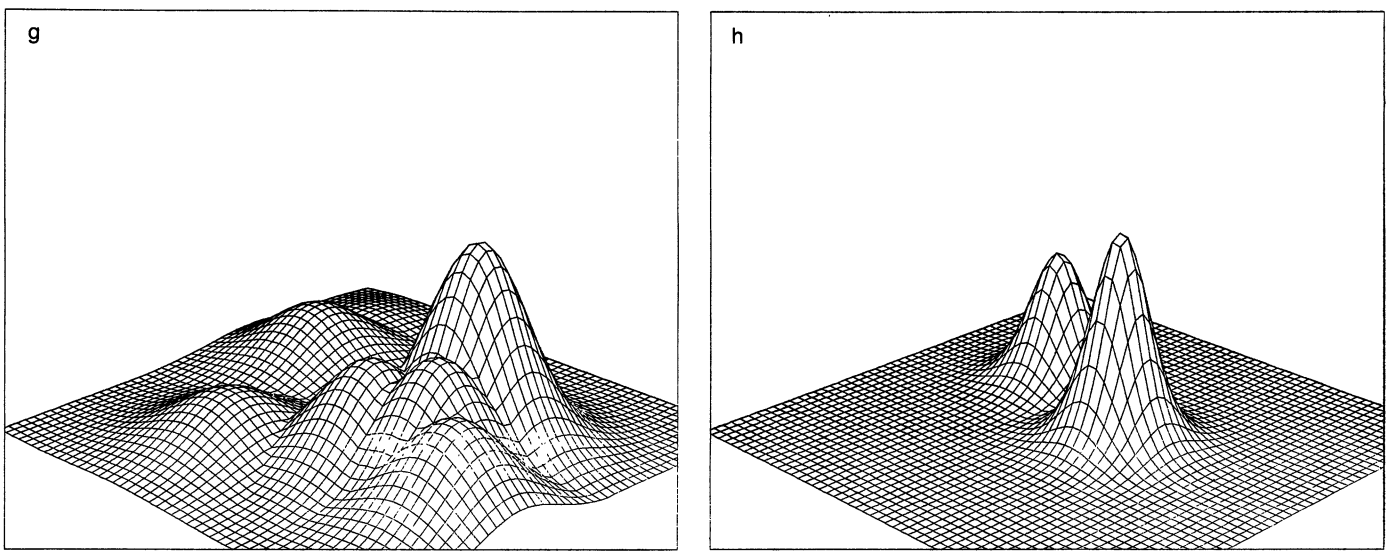

Figure 6. Comparison of Projection Pursuit and k Nearest Neighbor Density Estimates on a 10-Dimensional Problem. (a–d) Two-dimensional cross-sections of PPDE estimates for first four Monte Carlo trials; (e–h) two-dimensional cross-sections of KNNE estimates for first four Monte Carlo trials.

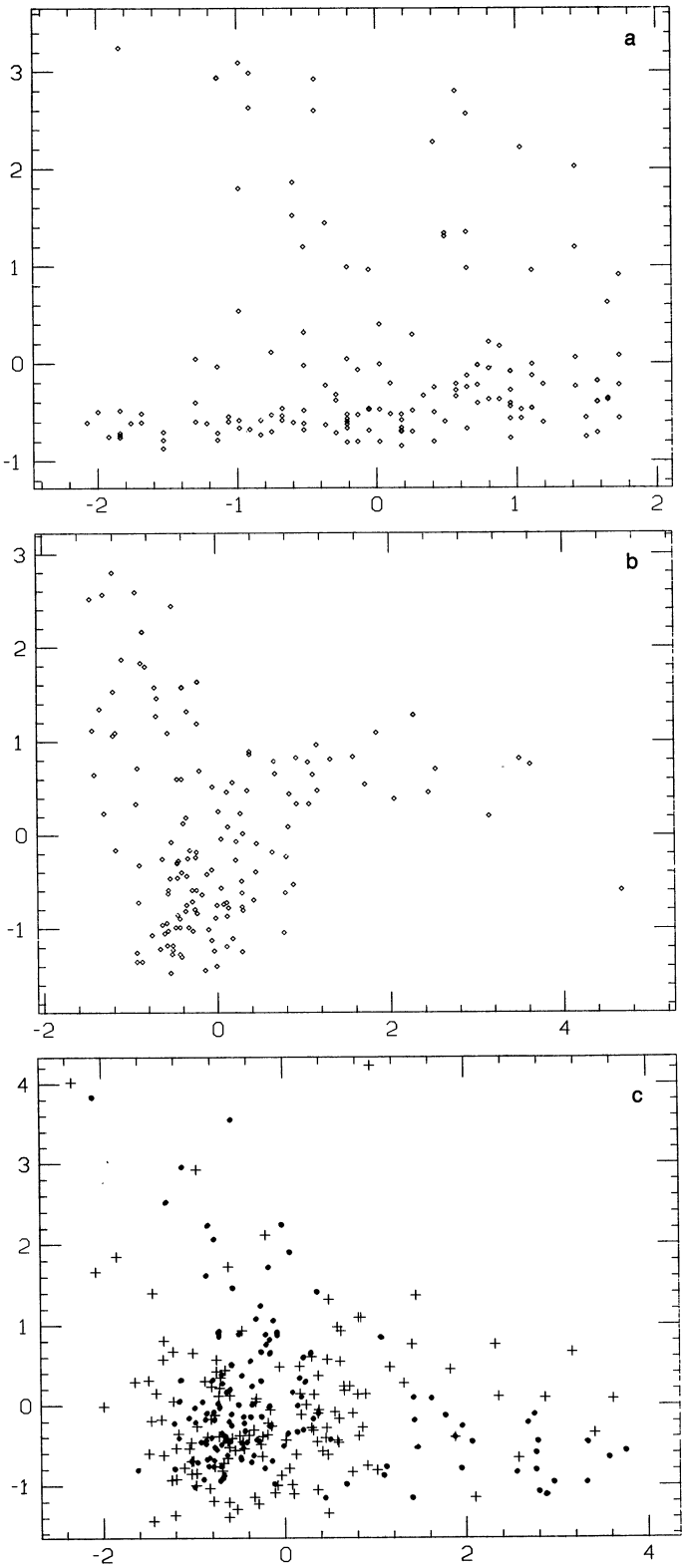

Figure 7. Exploratory Analysis—Density Factorization. (a) Diabetes data, x_1 versus x_3; (b) diabetes data, x_2 versus x_4; (c) diabetes data (·) and Monte Carlo from factored model $p_0(\mathbf{x})(+)$ projection onto plane spanned by first two solution linear combinations.

$$\frac{p_{M-1}(\mathbf{x})}{p_{M-2}(\mathbf{x})} = \prod_{m=1}^{M-2} \frac{f_m'(\theta_m \cdot \mathbf{x})}{f_m(\theta_m \cdot \mathbf{x})} f_{m-1}'(\theta_{M-1} \cdot \mathbf{x}). \quad \text{(B.2)}$$

We estimate the maximum value of (B.2) by its largest value over the data sample. Applying the accept/reject procedure to the Monte Carlo sample drawn from p_{M-2} yields a (smaller) sample drawn from $p_{M-1}(\mathbf{x})$. The remaining Monte Carlo observations are drawn from $p_{M-1}(\mathbf{x})$ by applying the accept/reject procedure to a sample from $p_0(\mathbf{x})$, using

$$\frac{p_{M-1}(\mathbf{x})}{p_0(\mathbf{x})} = \prod_{m=1}^{M-1} f_m'(\theta_m \cdot \mathbf{x}). \quad \text{(B.3)}$$

Again, the maximum value of (B.3) is estimated by its largest value over the data sample,

$$\hat{\gamma} = \max_{1 \le i \le N} \prod_{m=1}^{M-1} f_m'(\theta_m \cdot \mathbf{x}_0). \quad \text{(B.4)}$$

A small problem can arise because the maxima of (B.2) and (B.3) might not be assumed at one of the original data points. We could encounter a Monte Carlo observation $\mathbf{y}$ with $r(\mathbf{y}) > \gamma$. In this case, $\mathbf{y}$ is included in the sample L times, where L is the integer part of $r(\mathbf{y})/\hat{\gamma}$. The quantity $r(\mathbf{y}) - L$ is then used with the standard accept/reject procedure to determine whether $\mathbf{y}$ is accepted yet another time.

[*Received January 1982. Revised March 1984.*]

REFERENCES

BELLMAN, R.E. (1961), *Adaptive Control Processes*, Princeton, N.J.: Princeton University Press.

BONEVA, L.I., KENDALL, D.G., and STEFANOV, I. (1971), "Spline Transformations," *Journal of the Royal Statistical Society*, Ser. B, 33, 1–70.

DE BOOR, D. (1978), *A Practical Guide to Splines*, New York: Springer-Verlag.

FRIEDMAN, J.H., and TUKEY, J.W. (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers*, C23, 881–890.

FRIEDMAN, J.H., and STUETZLE, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, 76, 817–823.

KRONMAL, R.A., and PETERSON, A.V. (1981), "A Variant of the Acceptance-Rejection Method for Computer Generation of Random Variables," *Journal of the American Statistical Association*, 76, 446–451.

LOFTSGAARDEN, D.O., and QUESENBERRY, C.P. (1965), "A Nonparametric Estimate of a Multivariate Density Function," *Annals of Mathematical Statistics*, 36, 1049–1051.

PARZEN, E. (1962), "On the Estimation of a Probability Density Function and the Mode," *Annals of Mathematical Statistics*, 33, 832–837.

REAVEN, G.M., and MILLER, R.G. (1979), "An Attempt to Define the Nature of Chemical Diabetes Using Multidimensional Analyses," *Diabetologia*, 16, 17–24.

ROSENBLATT, M. (1971), "Curve Estimates," *Annals of Mathematical Statistics*, 42, 1815–1842.

RUBINSTEIN, R.Y. (1981), *Simulation and the Monte Carlo Method*, New York: John Wiley.

STONE, M.H. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

TAPIA, R.A., and THOMPSON, J.R. (1978), *Nonparametric Probability Density Estimation*, Baltimore: Johns Hopkins University Press.