AD-A107 816 HARVARD UNIV CAMBRIDGE MASS DEPT OF STATISTICS F/6 12/1
DENSITY ESTIMATION AND PROJECTION PURSUIT METHODS.(U)
SEP 81 P J HUBER N00014-79-C-0512
UNCLASSIFIED PJH-7 NL

RESEARCH REPORT

P.J.H. - 7

DENSITY ESTIMATION AND PROJECTION PURSUIT METHODS

Peter J. Huber
September 1981

This work was facilitated in part by National Science Foundation Grant MCS-79-08685 and Office of Naval Research Contract N00014-79-C-0512.

Department of Statistics
Harvard University
Cambridge

Density estimation and projection pursuit methods

Peter J. Huber

Harvard University, Cambridge MA

ABSTRACT

Recently, Friedman and Stuetzle (1981) have proposed to use projection pursuit methods for multidimensional density estimation. We discuss conceptual and technical issues in density estimation and describe several variant approaches to projection pursuit density approximation and estimation. A new measure of fit between multivariate densities is introduced (maximum marginal relative entropy), and some of its properties are derived. It seems to be particularly well adapted to projection pursuit density approximation.

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

1. Why density estimates?

Density estimates are used in two quite distinct contexts

(1) to aid with the visual analysis of data;

(2) to serve as substitutes for the unknown true density.

One occasionally hears the claim that density estimates are not really needed, and that statisticians would do better by using empirical measures in an intelligent fashion than by devising arbitrary density estimates. How far is this claim justified?

With regard to (1), the criticism is clearly unfounded: it is easier to perceive multimodality and to approximately locate the modes in a histogram with well-chosen bin width than in the empirical cumulative. In a two-dimensional scatterplot, differences in the underlying density of about 30% tend to go unnoticed, but not so in a density estimate. Compare in particular Figs. 1 to 5 (pp. 45-49) of Fryer's discussion of the paper by Boneva, Kendall and Stefanov (1971).

With regard to (2), the criticism may contain some truth. For instance, in adaptive estimation of location it is certainly most transparent to describe the empirical version of the score function $-f'/f$ as being the logarithmic derivative of a smooth estimate $\hat{f}$ of the density f, but the smoothing constant (kernel width) and other peculiarities of $\hat{f}$ must be determined not by any *ad hoc* density estimation criterion, but by standards derived from the ultimate adaptation requirement. Compare Stone (1975).

The archetypal situation where the empirical measure (in the raw or in a smoothed form) substitutes for the unknown true distribution is the so-called *bootstrap* (cf. Efron 1977): we are given a single random sample $(x_1, \ldots, x_n)$ of size n from some unknown true distribution F, and we should like to learn about the sampling properties of a certain statistic $T_n = T_n(x_1, \ldots, x_n)$. We do this by drawing N random samples of size m (with replacement) from the empirical measure F_n (or from a smoothed version thereof), and we base our conclusions on averages over the population consisting of the N values of T_m calculated from these random samples. Since we shall want to keep the order of the Monte Carlo error $O(N^{-1/2})$

somewhat below the intrinsic error $O(n^{-1/2})$ caused by the finiteness of the original sample, we usually choose N larger than n (but not exaggeratedly so). Usually, we take $m=n$, since mostly we are interested in the properties of statistics calculated from all the available data, but also the limiting case $m \to \infty$ is of particular interest.

It happens that most statistics possess some built-in smoothing properties, evidenced by the fact that they tend to work reasonably well also with rounded and grouped data. As a consequence, it is usually adequate to use the raw, unsmoothed empirical measure F_n, at least as long as m is small (say $m \leq n$). If m is substantially larger than n, we can get into serious trouble (an obvious example is furnished by the bootstrap estimate of the variance of the sample median). More generally, the bootstrap performs poorly with the raw F_n whenever the statistic T_n is concerned with local features of the distribution (e.g. for many statistics based on nearest neighbor techniques). Then it becomes essential to bootstrap from a wisely chosen density estimate. Compare in particular Silverman's recent (1981) paper on using bootstrap methods for estimating and testing the number of modes.

All approaches to density estimation involve at least one arbitrary smoothness parameter (the bin width in histograms, the kernel width in kernel smoothers, a Lagrange multiplier in the penalized likelihood approach, etc.). The optimal choice of this parameter depends on the unknown smoothness of the true underlying density, but also, and perhaps even more so, on the objectives of the investigator. If he is using density estimates as an aid to visual data scrutiny, it is important to try several choices of these parameters. Cross validation in the sense of Wahba and Wold (1975) is a great tool, but -- like the automatic exposure control in modern cameras -- it should be used with circumspection, and must be overridden on occasion: if we are looking for rare rough spots in an otherwise very smooth curve, automatic cross-validation will tend to oversmooth and to smooth away the spots of interest.

2. One dimensional density estimates

In comparison to higher dimensional problems, one dimensional density estimation is easy and straightforward. There are several good and reasonably well understood approaches (plain

histograms, kernel estimates, penalized likelihood estimates, plus some others, each with several variants (e.g. whether one chooses bins with constant width or with constant occupancy numbers in a histogram, and similarly for kernel estimates). See *Tapia and Thompson (1978)*.

The choice between these approaches often is more a matter of taste than a question of actual statistical performance. For our purposes, we shall need computationally fast estimates. The histogram density is fast, but it may not be smooth enough. Probably the fastest smooth estimate is obtained by fast guessing a histogram with too narrow bin width (e.g. equal to the resolution interval of the ultimate density estimate), and then smoothing this histogram by a simple kernel smoother. Compare Freedman and Diaconis (1980) for bin width recommendations.

3. Two and three-dimensional density estimates and their graphical representation

In two dimensions, kernel estimates still work well, and they lend to neat (and computationally efficient) graphical representations, namely to the ones familiar from cartography: smooth isodensity contours with coloring.

Two-dimensional histograms, or more precisely, stereographic or perspective projections of brick-histograms, are widely used but have serious deficiencies: it is often difficult to make out which brick sits on a particular point (x,y) of the base plane, and small bricks can disappear behind a wall of high ones.

Penalized likelihood estimates and related approaches based on spline smoothing may have even more serious drawbacks for dimensions two and higher. In particular, they exhibit awkward cusp and jump behavior (cusps at the support points). Moreover, they are expensive to compute.

Dimension 3 is the highest that allows direct visual representation of densities. Crystallographers for example routinely make 3-dimensional electron density maps by drawing contours -- 2-dimensional section by 2-dimensional section -- on stacks of parallel perspex sheets; more recently and more conveniently, the same kind of pictures are now generated on high power graphic display screens. Another possibility -- also dependent on kinematic display facilities -- is to use coloured scatter plots: show large random samples from a density that has been modified to emphasize high density regions and to downplay low density regions. This colouring might for instance be achieved by replacing the original density f by a rescaled power f^a. This generalizes the "sharpening by erosion" of Tukey and Tukey (1980).

4. Multidimensional density estimates

In dimensions higher than 3, direct visualisation of density estimates is impossible (or at least highly impractical, our visual cortex is too much geared to 3 dimensions). In addition, we run into the so-called "curse of dimensionality": high dimensional space is mostly empty. For instance, if points are uniformly distributed in the 10-dimensional unit ball, then a ball with radius 1/2 only contains a fraction $2^{-10} = 1/1024$ of the points.

As a consequence, we must think hard about what we want to look at, and what our estimates are supposed to achieve when they are used as substitutes for the unknown true density.

Ordinarily, we would prefer projections to sections (but projections of "thick" sections might sometimes be interesting). It becomes debatable whether we should first identify the interesting projections (either by a *priori* considerations, or by projection pursuit methods applied to the point cloud), and then use a low dimensional density estimate on the projected data, or whether we should first find a full p-dimensional density estimate and then determine the interesting low dimensional marginals. With density estimates based on projection pursuit ideas, the two approaches are in fact very similar.

Density estimates may serve as powerful visual clustering methods. What kinds of features can we hope to detect in high dimensional data, and by what methods?

First, we must be aware that because of the "curse of dimensionality" even large regions of zero density are difficult or impossible to spot.

Small high density clumps on the other hand can be spotted by appropriate nearest neighbor techniques (but not by standard kernel estimates with constant kernel width). More gen-

orally, nearest neighbor techniques will be able to spot substructures consisting of points concentrated near low dimensional (not necessarily linear) submanifolds.

Projection pursuit methods (to be discussed below, see also Huber 1981b) appear to be the first and so far only ones able to recognize substructures consisting of points concentrated near linear manifolds of low codimension (e.g. hyperplanes).

Linear substructures of intermediate dimensions (dimension near $p/2$ in a p-dimensional space, where p is large) may be even more difficult to spot. An (admixed) possibility consists in forming a minimal spanning tree, breaking it into connected components by deleting all links exceeding a certain length, and then doing robust principal component analyses on the components (cf. Huber 1981a, Ch. 8, and Chen and Li, 1981).

If we prepare density estimates for visual inspection, it is usually better to use slight undersmoothing. But if we plan to use them in a bootstrap, then it is usually better to stay on the "blind", slightly oversmoothed side, in particular if we want to check whether a peculiar local feature might have been a mere random effect. Compare Silverman (1980).

If a density estimate is to be used solely as a visual aid, it does not do much harm if it occasionally has negative values. But if it is to be inserted into a bootstrap or a similar procedure, we better make sure that it is a genuine probability density. Moreover, the representation of the density estimate then should be chosen such that it is easy to draw random samples from it.

I hope this short discussion has made it clear that we cannot expect any single approach to density estimation to do well under all circumstances.

5. Projection pursuit density approximation

Projection pursuit density estimates were recently proposed by Friedman and Stuetzle (1981) as an extension of the projection pursuit regression idea. The basic idea is to search for a one dimensional projection for which the projection of the pointcloud and the marginal density of the desired p dimensional density estimate disagree most, and then to improve the estimate by modifying its marginal density in that particular direction. We shall now show that the Friedman-Stuetzle approach solves a minimum problem posed in terms of relative entropy.

We first discuss a non-sampling version, namely the problem of approximating a density f in R^p by a function g (if possible a probability density) having a special functional form.

For example, we might approximate f by an additive decomposition of the form

$$f(x) \approx g^{(k)}(x) = \sum_{j=1}^{k} h_j(a_j^T x), \tag{5.1}$$

where $x \in R^p$, $a_j \in R^p$, and the h_j are functions of one real variable.

It is clear that under mild regularity conditions we can obtain arbitrarily accurate approximations of this form. Assume for instance that the characteristic function ψ of f is absolutely integrable, then f has the harmonic representation

$$f(x) = (2\pi)^{-p} \int \psi(s) e^{-is^T x} ds \tag{5.2}$$

Any finite Riemann sum approximating the integral on the right hand side then yields an additive decomposition of the type (5.1).

Though the representation (5.1) has awkward features: while $g^{(k)}$ converges to f uniformly on compacts when the mesh of the Riemann sum is extended and refined, the approximating sums clearly are not Lebesgue integrable, and they certainly do not exhibit the global features of a "typical" density (i.e. having a hump in the middle, and decaying to 0 as $|x| \to \infty$). We can insure integrability by multiplying the right hand side of (5.1) by a fudge factor like $\exp(-|x|^2)$, or equivalently, by picking a fixed, strictly positive, smooth probability density in R^p (e.g. the standard Gaussian), and taking all densities relative to that instead of Lebesgue measure. Perhaps even worse, it is difficult to make g everywhere positive and to keep it positive and integrating to 1 while more terms are added to the sum.

Nevertheless, additive decompositions have technical advantages, the main one being that it is readily easy to calculate marginal densities.

The following is a somewhat more specific proposal for such a decomposition. First, we transform the distribution by an affine transformation such that it is centered at 0 and has unit

covariance matrix (or alternatively, use a robust procedure, such that the central part of the distribution is considered at 0 and has unit covariance). Then approximate f by a decomposition of the form

$$f(x) \approx g^{(k)}(x) = (2\pi)^{-p/2}\sigma^{-1}e^{-\frac{1}{2}x'x}\sum_{j=0}^{k} h_j(a_j'x), \tag{5.3}$$

with $h_0 = 1$. Note the formal similarity to an Edgeworth expansion.

Correspondingly, a multiplicative decomposition

$$f(x) \approx g^{(k)}(x) = \prod_{j=1}^{k} h_j(a_j'x) \tag{5.4}$$

looks more attractive. Note that if $k=p$, and if the a_j are linearly independent vectors, then $g=g^{(k)}$ is a product density and can be written as

$$g(x) = \prod_{1}^{p} g_j(x_j) \tag{5.5}$$

in a suitable coordinate system, while the g_j are one-dimensional probability densities. Clearly, such a decomposition is well suited to represent (approximate) product densities, but it may yield awkward representations (i.e. involving too many terms) that are difficult to interpret, if the density f happens to be a mixture instead of an approximate product.

Note that we are iteratively decomposing $\log f$, and thus closely emulate projection pursuit of y. Another possibility to create positivity is to additively decompose $\sqrt{f}$. Incidentally, since the square root transform stabilizes the variance of the histogram, the latter approach also has some advantages from a sampling point of view.

Any scheme for iteratively determining directions a_j and successive approximations $g^{(k)}$ shall be called a *projection pursuit approximation*.

The quality of the approximation of g to f can be measured in many ways, e.g. by

(1) relative entropy

$$E(f,g) = \int \log(f/g)\, f\, dx,$$

or by

(2) Hellinger distance

$$H(f,g) = \int(\sqrt{f}-\sqrt{g})^2 dx,$$

or by any other measure of distance between distributions (e.g. Prohorov distance, bounded Lipschitz metric, etc.). Since we are working with densities, we are naturally more attracted to (1) or (2) than to the other distances mentioned. Among them, (1) is of course particularly well suited to an *additive decomposition* of $\log f$, (2) to an additive decomposition of $\sqrt{f}$.

Note that both Hellinger distance and relative entropy are invariant under *affine* transformations. Hellinger distance is a metric, relative entropy is not (it is not symmetric in its two arguments), but it follows from Jensen's inequality that $E(f,g) \geq 0$, and that $E(f,g)=0$ implies that $f=g$ a.s.

We may allow $k<p$ in the decompositions. In the absence of fudge factors g then fails to be Lebesgue integrable, but $E(f,g)$ still may be finite. However, in such cases it is usually preferable to pad g by $p-k$ "bland" (e.g. Gaussian) components, in order to turn it into a genuine probability density.

6. Minimization of relative entropy

In this section we are concerned with optimal choices for the directions a_j and the factors h_j in the decomposition (5.4). Assume first that $k=p$ and that the a_j are fixed, linearly independent vectors in R^p. We may choose a_j to be the jth coordinate direction without loss of generality. We first consider the problem of finding the best approximation in the relative entropy sense of the given density f by a product density

$$g(x) = \prod_{1}^{p} g_j(x_j),$$

where the g_j are one-dimensional probability densities.

Relative entropy

$$E(f,g) = \int[\log f - \sum \log g_j(x_j)]\, f\, dx_1 \cdots dx_p$$

is minimized by minimizing

$$-\sum_j \int \log g_j(x_j) f(x_1, \ldots, x_p)\, dx_1 \cdots dx_p,$$

which in turn is minimized by minimizing

$$-\int \log g_j(x_j)\, f_j(x_j)\, dx_j$$

for each j separately, where

$$f_j(x_j) = \int f(x_1, \ldots, x_p)\, dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_p$$

is the jth marginal density.

Since $E(f_j, g_j) \geq 0$ for $g_j \neq f_j$, the minimum density is achieved for the unique choice

$$g_j = f_j.$$

We note in passing that this calculation at the same time proves that Shannon entropy

$$E_{Sh}(f) = -\int \log f(x)\, f\, dx$$

satisfies

$$E_{Sh}(f) \leq \sum_j E_{Sh}(f_j),$$

with equality if $f = \prod_j f_j$.

By letting the p directions a_j vary simultaneously and freely, we may more generally approximate f by the best possible product density (we do not worry about existence of the maximum for the moment):

$$g(x) = \prod_{j=1}^{p} g_j(a_j^T x)$$

It is not clear whether a stepwise projection pursuit approach will achieve the best possible approximation by a product density: first solve a maximum problem to find a_1, then find a_2 and so on.

But if f is an exact product density in a suitable coordinate system, then a properly designed stepwise approach will pick up the unique factors one at a time. This is a non-trivial result involving some subtle properties of entropy, it shall be sketched briefly.

It is convenient to use random variable terminology and to write $E_{Sh}(X) = E_{Sh}(f)$ if the one-dimensional random variable X has density f. Then the functional

$$Q(X) = E_{Sh}(\lambda \cdot \sigma(\lambda)) = \log \sigma(X) - E_{Sh}(X),$$

where $\sigma(X)$ is the standard deviation of X, is clearly invariant:

$$Q(sX + t) = Q(X), \quad s \neq 0,$$

and has the property that for independent random variables X, Y:

$$Q(X + Y) \leq \max(Q(X), Q(Y)),$$

and that the inequality is strict unless both X and Y are normal, see Huber (1981b). Note that $Q(X) = E(f, g)$, where g is a normal density with the same mean and variance as f.

Now assume that f is a product density, without loss of generality we may assume that

$$f(x) = \prod_j f_j(x_j),$$

and that the relative entropies are in decreasing order:

$$E(f_1, g_1) \geq E(f_2, g_2) \geq \cdots \geq E(f_p, g_p).$$

The above-mentioned inequality for Q implies that the maximum of Q is reached at a factor of f, namely at the factor f_1 with the largest relative entropy $E(f_1, g_1)$. We now divide out this factor and replace f by

$$f^*(x) = f(x) g_1(x_1) / f_1(x_1)$$

Then if f^* is subjected to the same process as f before, the second factor f_2 is picked out, and so on.

Playing around with a few symbolic examples has led to the following observations. Minimization of Shannon entropy $E_{Sh}(f, \cdot)$ and of related possibilities seems to be a numerically unstable process if the covariance matrix of f is poorly conditioned (i.e. if the ratio between the largest and the smallest eigenvalue is large). This is made worse by the fact that there will be many local minima in general, and that the minima sometimes are deep but narrow even if the covariance matrix is well conditioned. So the minima may be difficult to find.

If f has finite second moments, then the best Gaussian approximation to f in the relative entropy sense has the same mean vector and covariance matrix as f. This is easily shown by a variational argument. In view of the preceding remarks it would therefore seem advisable to apply an affine transformation such that in the new coordinate system f has mean 0 and unit covariance matrix, and to start the approximation process with $g^{(0)} = \varphi$ being the standard p-

dimensional normal distribution. This choice also automatically provides the "ideal" padding mentioned earlier.

Let g be an approximation to the p-dimensional density f. We shall now attempt to improve the approximation by replacing $g(x)$ by $\tilde{g}(x)=g(x)h(x_1)$, where h depends on the first coordinate only. Note that g and $\tilde{g}$ determine the same conditional density given x_1. An intuitively attractive choice then is to determine h such that the marginal distribution of $\tilde{g}$ in direction x_1 agrees with that of f. This indeed minimizes relative entropy.

Lemma 6.1. *Relative entropy $E(f,\tilde{g})$ is minimized by the choice*

$$h(x_1) = f_1(x_1)/g_1(x_1),$$

where f_1 and g_1 are the marginal densities of f and g in direction x_1.

Proof. We note that the conditional density, given x_1, is the same for g and for $\tilde{g}$, namely

$$g(\cdot|x_1) = g(x_2,\ldots,x_p|x_1) = g(x)/g_1(x_1),$$

and that $g_1(x_1)h(x_1)$ is the marginal density of $\tilde{g}$. Thus

$$\begin{aligned} E(f,\tilde{g}) &= \int(\log f - \log \tilde{g})\, f\, dx \\ &= \int[\log f(\cdot|x_1) - \log g(\cdot|x_1) + \log f_1(x_1) - \log(g_1(x_1)h(x_1))]\, f\, dx, \end{aligned}$$

which is minimized by minimizing

$$\int[\log f_1 - \log(g_1 h)]\, f\, dx = E(f_1, g_1 h),$$

and this clearly is achieved for the unique choice $g_1 h = f_1$. ∎

For the minimizing choice of $\tilde{g}$ we have

$$E(f,g) - E(f,\tilde{g}) = E(f_1,g_1).$$

If we can choose the projection direction u, then the best possible improvement of relative entropy that can be achieved through replacing $g(x)$ by $\tilde{g}(x) = g(x)h(u^T x)$ clearly is obtained with

$$h = f_u / g_u,$$

where f_u and g_u are the marginal densities of f and g respectively in direction u, and where u is chosen such that it maximizes

$$E(f,g) - E(f,\tilde{g}) = E(f_u,g_u) \tag{6.1}$$

(at the moment, we are not concerned about the existence of such an u; maximization within a prescribed relative error tolerance is in fact good enough for all practical purposes).

The procedure just described shall be referred to as the *projection pursuit approximation method* in the following.

7. Convergence in maximum marginal relative entropy

We can use *maximum marginal relative entropy*

$$E^*(f,g) = \sup_u E(f_u,g_u)$$

as a measure of discrepancy between two p-dimensional densities f and g. Clearly, because of (6.1), $E^*(f,g) \le E(f,g)$. Since any distribution is uniquely characterized by the set of its marginals in all possible directions (Cramér-Wold theorem), we have

$$E^*(f,g_n) \to 0 \;\Rightarrow\; f=g \;\Rightarrow\; E(f,g_n) \to 0.$$

Since each step of projection pursuit decreases $E(f,g)$ by $E^*(f,g)$, the latter quantity clearly converges to 0, and, for any given $\varepsilon > 0$, it takes at most $E(f,g)/\varepsilon$ steps to reach a density $\tilde{g}$ for which $E^*(f,\tilde{g}) \le \varepsilon$.

Maximum marginal relative entropy $E^*(f,g)$ seems to be a concept particularly well suited to projection pursuit approximation, and it deserves closer study. Let f be a fixed probability density, while g_n is an arbitrary sequence of probability densities.

Clearly, $E(f,g_n) \to 0$ implies $E^*(f,g_n) \to 0$. I conjecture that the reverse implication is false, but I do not have a counterexample.

We now shall derive some consequences of E^*-convergence.

Proposition 7.1. *$E^*(f,g_n) \to 0$ implies that $g_n \to f$ in the sense of weak convergence of the underlying measures.*

Basically, this proposition is just another version of the Cramér-Wold theorem, compare Billingsley (1968, p.48). We prove it with the aid of two auxiliary lemmas of independent interest.

Lemma 7.2. Let $z > -1$, then

$$z - \log(1+z) \le \frac{1}{4}\min(1, z^2).$$

Proof. For $z \ge 1$ the assertion is true, since it holds for $z=1$ and the left hand side is monotone increasing. We note that

$$k(z) = z - \log(1+z) - z^2/4$$

vanishes for $z=0$, and its derivative

$$k'(z) = \frac{z(1-z)}{2(1+z)}$$

satisfies $k'(z) < 0$ for $-1 < z < 0$, $k'(z) > 0$ for $0 < z < 1$. Hence the assertion also holds for $|z| < 1$.

∎

Lemma 7.3. For any two densities f and g with respect to Lebesgue measure in R^k we have

$$\int |f-g|\,dx \le 4\,|E(f,g)|^{1/2}.$$

Proof. Put $h = g/f - 1$, then

$$\begin{aligned} E(f,g) &= -\int \log(g/f)\, f\,dx \\ &= \int [\log(g/f) + g/f - 1]\, f\,dx + \int (1 - g/f)\, f\,dx \\ &= \int [(h - \log(1+h)]\, f\,dx + \int_{f>0} (f-g)\,dx \\ &\ge \frac{1}{4}\int \min(1, h^2)\, f\,dx \end{aligned}$$

By the Chebyshev inequality (or more precisely, by repeating its proof) we obtain for $\delta \le 1$

$$P_f\{|h| \ge \delta\} \le \frac{4E(f,g)}{\delta^2}.$$

Assume now that $E(f,g) \le 1/4$, and put $\delta = |4E(f,g)|^{1/2}$, then

$$P_f(|h| \ge \delta) \le \delta$$

Let C be the set $\{x \mid |h(x)| \ge \delta\}$, then $|g-f| \le \delta f$ on C^c. We have

$$\int_{C^c} f\,dx \ge 1 - \delta,$$

and

$$\int_C g\,dx \le \int_C (f + \delta f)\,dx \le (1+\delta)\delta,$$

hence

$$\int_{C^c} g\,dx \ge 1 - 2\delta.$$

Thus

$$\begin{aligned} \int |f-g|\,dx &= \int_C |f-g|\,dx + \int_{C^c} |f-g|\,dx \\ &\ge \delta \int_C f\,dx + \int_{C^c} f\,dx + \int_{C^c} g\,dx \ge 1 - 4\delta. \end{aligned}$$

This proves the lemma, since for $E(f,g) \ge 1/4$ the assertion is trivially true.

∎

Corollary 7.4. Convergence in relative entropy $E(f, g_n) \to 0$ implies L_1 convergence (i.e. in total variation) and convergence in Hellinger distance.

Proof. L_1-convergence is an immediate consequence of Lemma 7.3, and convergence in Hellinger distance follows from the remark that $(\sqrt{f} - \sqrt{g})^2 \le |f-g|$.

∎

Proof of Proposition 7.1. In view of Lemma 7.3, $E^*(f, g_n) \to 0$ implies that the marginal densities show uniform L_1-convergence:

$$\sup_a \int |f_a - g_{n,a}|\,dx \to 0.$$

Hence, the characteristic functions ψ_a of the marginals converge uniformly, and since the characteristic function ψ of any density f is related to the characteristic functions ψ_a of its marginals f_a by

$$\psi(ta) = E\left\{e^{itZ_a}\right\} = \psi_a(t),$$

it follows that the characteristic functions ψ_n of g_n converge uniformly to ψ:

$$|\psi(ta) - \psi_n(ta)| \le \sup_a \int |f_a - g_{n,a}|\,dx$$

Hence, g_n converges weakly.

∎

If f is sufficiently smooth, so that its characteristic function is absolutely integrable, and if

the sequence g_k is generated by the projection pursuit method described at the end of the preceding section, then I conjecture that $g_k \to f$ uniformly and in the L_1 sense. Actually, I can prove only a very special case (which however covers the principal intended application).

Proposition 7.3. Assume that f can be deconvolved with a Gaussian component:

$$f = \bar{f} * \varphi$$

where $\bar{f}$ is some density, and φ is normal $N(0,\sigma^2 I_p)$. Then the sequence $g^{(k)}, k=0,1,\cdots$ of approximating densities, constructed by projection pursuit as in Section 6, converges uniformly and in L_1 to f.

Proof. Each of the $g^{(k)}$ allows a decomposition

$$g^{(k)} = g^{(k)} * \varphi.$$

This is proved by induction. It clearly is true for $k=0$, since $g^{(0)}$ itself is Gaussian.

Now assume that $g=g^{(k)}$ can be deconvolved. For the following argument it is essential that φ is a product density in all orthogonal coordinate systems, so that convolutions can be calculated coordinatewise. Choose the coordinate system such that a_1 is the first coordinate direction. Then the representation

$$g = g(\,|x_1)g_1(x_1)$$

induces the representation

$$g = g(\,|x_1)g_1(x_1)$$

with $g(\,|x_1)$ and $g_1(x_1)$ being obtained by convolving their bared counterparts with $(p-1)$- and 1-dimensional normal distributions respectively. If in the last displayed expression we replace $g_1(x_1)$ by $f_1(x_1) = f_1(x_1)*N(0,\sigma^2)$, we obtain the deconvolution of the next term $g^{(k+1)}$.

The characteristic functions of $g^{(k)}$ and $\bar{g}^{(k)}$ are related by $\psi^{(k)}(s) = \bar{\psi}^{(k)}(s)\exp(-\sigma^2|s|^2/2)$, and similarly for ψ. Hence ψ and $\psi^{(k)}$ are majorized by $\exp(-\sigma^2|s|^2/2)$.

Thus the $\psi^{(k)}$ are absolutely integrable, and $g^{(k)}$ can therefore be represented as

$$g^{(k)} = (2\pi)^{-p}\int \psi^{(k)}(s)\,e^{-is^Tx}\,ds.$$

Since the sequence $\psi^{(k)}$ converges uniformly, it now follows from the majorization of ψ and $\psi^{(k)}$ that the sequence

$$|f(x)-g^{(k)}(x)| \le (2\pi)^{-p}\int|\psi(s)-\psi^{(k)}(s)|\,ds$$

converges uniformly to 0. L_1 convergence follows trivially.

8. Projection pursuit density estimation

It is straightforward to change the projection pursuit approximation procedure of Section 6 into a density estimator. The first step is to standardize the predictand in R^p by an affine transformation so that it is centered at 0 and that its covariance matrix is the unit matrix. Note that there are some very delicate robustness questions our density estimates should not be sensitive to occasional outliers, but they should be able to pick up long tails in the underlying distribution. These two requirements are in some sense contradictory, and we lack a rational theory for dealing with them. But, from a pragmatical point of view, we note that all density estimators have trouble coping with isolated points in the tails (they tend to produce equally isolated bumps in the estimate), so it is probably wise to set aside these points, to disregard them in the estimation procedure, but to show them as remarkable points in the pictures we produce.

Since it seems to be better if the initial estimate has too heavy tails than if it has too light tails, we would seem to have the choice of either using the classical, non-robust mean and covariance matrix together with a Gaussian $g^{(0)}$, or else robust location and covariance estimates together with a density $g^{(0)}$ that is heavier tailed than the Gaussian. I believe the simpler first version to be good enough in most cases.

The zeroth order density estimate $g^{(0)}$ thus ordinarily is the p-dimensional standard normal. The approximation steps now can be described as follows.

Let

$$g(x) = g^{(k)}(x) = g^{(0)}(x)h_1(a_1^Tx)\cdots h_k(a_k^Tx)$$

be the current density estimate.

According to Section 6, we now should determine a direction $a = a_{k+1}$ such that it maximizes $E(f_a, g_a)$, and then put $h_{k+1} = f_a / g_a$.

For a given a, f_a is straightforward to estimate: project the given sample in direction a, yielding $z_i = a^t x_i$, $i = 1, \ldots, n$, and then calculate a one dimensional density estimate $\hat{f}_a$ based on $(z_1, \ldots, z_n)$.

The projection g_a of the current density estimate g is a well defined, known quantity, and from the point of view of theory does not present any problem. However, we *may run into* trouble with its actual calculation, in particular since it has to be calculated inside a minimization loop.

Direct numerical integration almost certainly is too slow. There are several speeding Monte Carlo approaches. The first one is to replace g by a sample $y_1, \ldots, y_M$ from g, and then to estimate g_a in the same way as $\hat{f}_a$. This variant may not be easy to implement (how does one sample efficiently from g ?). A second one is to take a random sample $y_1, \ldots, y_M$ from $g^{(k)}$ (i.e. from the p-dimensional standard normal distribution), to put $v_i = h_1(a^t y_i), \ldots, h_k(a^t y_i)$, and to create a histogram with bin width Δ and value

$$t_j = \sum \{ h_1 \cdots h_k \mid v_i \approx a^t x_i \approx v_i + \Delta \}$$

for the jth bin. Then apply a kernel smoother to this histogram to obtain the estimate $\hat{g}_a$.

This approach will be predictably poor if $g = g^{(k)}$ has a *narrow, but high density* peak somewhere, because then the random sample from $g^{(k)}$ puts so few observations there that the value $\hat{g}_a$ is inaccurately estimated. If this happens, it may be best to abandon temporarily the algorithm, to locate the offending peak, *to replace* $g^{(k)}$ *by a mixture of* two differently scaled normals (one of them at the peak in question), and then to restart the projection pursuit algorithm with the new $g^{(k)}$.

We proposed here to use marginal relative entropy as the criterion to be maximized when determining a new projection direction. This certainly is the conceptually purest approach. But conceivably, other measures of discrepancy might offer advantages from a sampling or computational point of view, the matter needs further investigation.

2. *Maximization of Hellinger distance*

Hellinger distance is more complicated to maximize than relative entropy. Assume that g is the current approximation to f, and that we want to *find a function* $h(x_1)$ of the first coordinate alone such that the Hellinger distance between the probability densities f and

$$\tilde{g}(x) = g(x) h(x_1)$$

is maximized. Clearly, g and $\tilde{g}$ determine the same conditional densities given x_1, say $q = q(x_2, \ldots, x_p | x_1)$, so $h = \tilde{g}_1 / g_1$ is the quotient of the marginal densities in direction x_1. We can maximize $H(f, \tilde{g})$ by maximizing

$$\int \sqrt{f \tilde{g}}\, dx = \int \left[\int \sqrt{f q}\, dx_2 \cdots dx_p \right] \sqrt{\tilde{g}_1}\, dx_1$$

under the side condition

$$\int \tilde{g}_1\, dx_1 = 1.$$

We write for short

$$w(x_1) = \int \sqrt{f q}\, dx_2 \cdots dx_p.$$

Then the Schwarz inequality implies that

$$\int \sqrt{f \tilde{g}}\, dx = \int w \sqrt{\tilde{g}_1}\, dx_1 \le \left\{ \int w^2\, dx_1 \right\}^{1/2},$$

with equality iff $\sqrt{\tilde{g}_1}$ is proportional to w. Hence we should take

$$\sqrt{\tilde{g}_1} = c\, w,$$

with the constant c such that the side condition is satisfied, and the optimal h then is $h = \tilde{g}_1 / g_1$.

We noted earlier that for Hellinger distance an additive decomposition of $\sqrt{f}$ might be more germane. Assume

$$\sqrt{f} \approx \sqrt{g} = \sum_1^k u_j(a_j^t x)$$

In this case it is convenient to ignore the requirement that g should integrate to 1.

If $\sqrt{g}$ is the current approximation, then the best approximation of the form

$$\sqrt{g} + u(x_1)$$

is obtained by minimizing

$$\int (\sqrt{f} - \sqrt{g} - u)^2 \, dx,$$

which clearly is minimized by the choice

$$u(x_1) = \int (\sqrt{f} - \sqrt{g}) dx_2 \cdots dx_p.$$

In neither of the two problems the search for the optimal direction a of projection appears to allow a simple and intuitive formalization (as in the relative entropy case, where the relative entropy of the marginal densities had to be maximized).

In the sampling case, even graver inconveniences seem to arise, since it not at all clear whether the expressions for u and for u can be estimated in a computationally efficient fashion.

10. Consistency of projection pursuit density estimates

By re-interpreting some of the results obtained in Section 7, consistency of projection pursuit density estimates is almost trivial to prove. But the proof at the same time shows why consistency here is an almost useless concept.

Assume that we are given a sample of size n in R^p; let

$$\mu = \frac{1}{n} \sum_{i}^{n} \delta_{x_i}$$

be the empirical measure. We now apply a spherically symmetric normal kernel smoother to obtain the density estimate

$$f = \mu * N(0, \sigma^2 I_p).$$

Note that the marginal density f_a of f in direction a is obtained by applying the one-dimensional kernel $N(0, \sigma^2)$ to the projection

$$\mu_a = \frac{1}{n} \sum_{i} \delta_{a^t x_i}$$

of the original data.

It follows that if we iterate the projection pursuit estimate minimizing relative entropy, using a kernel smoother with Gaussian kernel in the projections, it converges in the L^1-sense to the p-dimensional kernel estimate f.

Since f can be deconvoluted with a normal component, it follows from Proposition 7.5 that the projection pursuit estimate converges uniformly and in L_1.

Since the p-dimensional kernel estimate f is consistent under very weak assumptions on the true underlying density f if σ tends to 0 slowly while the sample size n goes to ∞, the projection pursuit estimate is consistent too, provided we iterate it enough so that it approximates f sufficiently closely.

This result is not very helpful, however. After all, the main reason for using the projection pursuit estimate is that the sample size is too small for a p-dimensional kernel estimate to make sense, and we certainly would not want to iterate the projection pursuit estimate so far *that it approximates the latter*. The following example may illustrate these issues.

Example. Take a sample of size n from the standard normal $N(0, I_p)$ in p dimensions. Note that an equal mixture of two normal densities $N(\pm a, \sigma^2)$ is unimodal for $a \leq \sigma$, bimodal otherwise. Thus we may say that two sample points are merged in the p-dimensional kernel estimate $f = \mu * N(0, \sigma^2 I_p)$, if their Euclidean distance is $\leq 2\sigma$, and that they are separated otherwise. Then the expected number of merged pairs can be calculated as

$$m = \tfrac{1}{2} n(n-1) q,$$

where

$$q = P(|x_i - x_j| \leq 2\sigma) = \chi^2_p(2\sigma^2).$$

Numerically, we obtain with $p = 10$, $n = 10^6$, $\sigma = 0.1$, that $m = 0.4$. In other words, in the p-dimensional kernel estimate all 10^6 points, except maybe 1 or 2 merged pairs, are still separate. The one-dimensional marginal estimates on the other hand will be quite smooth. Note that for one-dimensional estimates a kernel width $kn^{-1/5}$ minimizes the asymptotic mean square error, and that the constant k is such that for $n = 10^6$, the choice $\sigma = 0.1$ is very nearly optimal (cf.

Wegman 1972, p. 548).

A somewhat more meaningful consistency result might be the following conjectural one.

Let f be the true underlying density and assume that it has been standardized such that its central part is approximated by the standard normal $g^{(0)}$. Let

$$g^{(k)}(x) = g^{(0)}(x)h_1(a_1^t x) \cdots h_k(a_k^t x)$$

be the k-th projection pursuit approximation to the true density f, obtained by minimizing maximum marginal relative entropy. Assume that the sequence $a_1, a_2, \cdots$ is uniquely determined, and that $E^*(f, g^{(k)}) < \varepsilon_k$ for some sequence ε_k. Let $\hat{a}_{k,n}$, $\hat{g}^{(k,n)}$ be the sequences obtained by projection pursuit density estimation from samples from f of size n, using a Gaussian $N(0, \sigma_n^2)$ kernel smoother on the one-dimensional projections, with $\sigma_n \to 0$ as the sample size $n \to \infty$ (the precise order of convergence needs to be specified). Then, for each fixed k, we would expect that $\hat{a}_{k,n} \to a_k$ and $\lim E^*(f, \hat{g}^{(k,n)}) < \varepsilon_k$, and that the convergence is much faster than the convergence of the p-dimensional kernel estimate.

Since we do not know the true f nor ε_k, except perhaps in synthetic sampling situations, we would in practice need an analogue of Mallows's C_p-statistic, telling us when to stop the projection pursuit approximation process.

The kernel width to be used for smoothing the one-dimensional marginals needs investigation. Presumably, one should use a width $hn^{-1/6}$ approximately optimal with regard to expected mean square error for a three times differentiable density (even if the underlying density is expected to be somewhat rougher), and one should in general tend to oversmooth rather than to undersmooth.

11. *Sampling questions*

Projection pursuit density estimation and relative entropy used as a distance measure raise some peculiar sampling questions. This section discusses them in a preliminary fashion.

We would like to approximate the p-dimensional density f by a function g of a certain form. Since f is unknown, we must approximate an estimate $\hat{f}$ instead. Here, in general, $\hat{f}$ is not a genuine p-dimensional density estimate, but stands short for the collection of *estimates* $\hat{f}_\alpha$ of the one-dimensional projections f_α of f. By $\hat{f}$ we denote the collection of expectations $\bar{f}_\alpha = E(\hat{f}_\alpha)$.

It is hardly realistic to aim for an "optimal" approximation. For reasons of computational expediency, the density estimates $\hat{f}_\alpha$ will have to be histogram estimates with constant bin width, smoothed with a simple smoother (possibly an inhomogeneous one - more smoothing in the tails). But anyway, we should like to keep the total approximation error $g - \hat{f}$ small, and then, common sense tells us that we should try to adjust the estimation and approximation procedures such that the bias $\bar{f} - f$, the random error $\hat{f} - \bar{f}$ and the approximation error $g - \hat{f}$ are all of about the same order of magnitude.

Let $E(f_\alpha, g_\alpha)$ be relative entropy (or some other measure of discrepancy between one-dimensional densities) and put

$$E^*(f, g) = \sup_\alpha E(f_\alpha, g_\alpha),$$
$$\bar{E}(f, g) = \operatorname{ave}_\alpha E(f_\alpha, g_\alpha).$$

Somewhat arbitrarily, we aim for the following approximate relations between bias, random error and approximation error. The bias should be smaller than the average random error, and the approximation error should be larger than the average random error, but it may fall below the maximum random error. In formulas:

$$E^*(f, \bar{f}) \approx \bar{E}(\bar{f}, f) < E^*(g, \hat{f}) \approx E^*(\bar{f}, \hat{f}) \tag{11.1}$$

Before we discuss these errors further, we insert some auxiliary remarks of a more general nature. We note that if the difference $\Delta f = g - f$ is infinitesimally small, then a Taylor expansion gives for relative entropy

$$e(f, g) = e(g, f) = \frac{1}{2}\int \left|\frac{\Delta f}{f}\right|^2 f \, dx, \tag{11.2}$$

neglecting higher order terms. Thus, relative entropy is, essentially, a weighted average of the squared relative error. Compare also Lemma 7.2, and the proof of Lemma 7.3.

The above formula (11.2) indicates that special care is needed in the tails of f, where f is small and the relative error may be unacceptably large. The problem can be partly alleviated by smoothing more heavily *in the tails, but since this increases the relative bias*, this does not help too much. Thus, we should truncate the tails, but note that the relative sampling error

$$E(\bar{f},\hat{f}) \approx \frac{1}{2}\int\left|\frac{\bar{f}-\hat{f}}{\hat{f}}\right|^2 \bar{f}\,dx$$

has the expectation

$$\frac{1}{2}\int\frac{\mathrm{var}(\hat{f})}{\hat{f}}\,dx, \tag{11.3}$$

and if we use a kernel estimate with kernel $K_\sigma(z) = \sigma^{-1}K(z/\delta)$, that is

$$\hat{f}_n(x) = \frac{1}{n}\sum K_\sigma(x-x_i),$$

we obtain that asymptotically (for large n and small σ) the integrand of (11.3) is independent of x:

$$\frac{\mathrm{var}(\hat{f}_n)}{\hat{f}_n} \approx \frac{1}{n\sigma}\int K^2 dx$$

Hence, truncating the range of integration is tricky: the computed value may be unstable and depend heavily on the choice of the truncation points.

It is an open question by how much *loc* trades with the tails, plus perhaps a modification of $E(f,g)$ (see below) are able to mitigate these awkward features of relative entropy.

An estimate alternative to a kernel estimate with constant width is a histogram estimate with constant occupancy number n_j and variable bin width. Technically, it is somewhat easier to deal with histograms where the bin width has been adjusted such that the *expected* occupancy number $E(n_j) = np_j$ is the same for all k bins, with $p_j = 1/k$; asymptotically, both versions are equivalent. Then, the ratio between the estimated and the expected average density in bin j is $f/\bar{f} = n_j/(np_j)$, this bad quotient also makes sense in the end bins, where $\hat{f}/\bar{f} = 0,\infty$ is undesirable. For this estimate, the relative bias in the end bins is intolerably large. But, at least,

the estimate has nice asymptotic sampling properties, as follows. Put

$$z_j = (n_j - np_j)/(np_j).$$

Then $\sum z_j = 0$, $E(z_j) = 0$, and $E(z_j^2) = (1-p_j)/(np_j)$.

For large n the relative entropy then is

$$\begin{aligned} E(\bar{f},\hat{f}) &= \int\log\left|\frac{\bar{f}}{\hat{f}}\right|\bar{f}\,dx \\ &= -\sum_1^k \log(1+z_j)\,p_j \\ &\approx \frac{1}{2}\sum_1^k z_j^2 \approx E(\bar{f},\hat{f}). \end{aligned}$$

For large n, this is distributed like

$$\frac{k-1}{2n}\chi^2_{k-1}$$

Hence, if $1 \ll k \ll n$, we have for one-dimensional densities

$$E(\bar{f},\hat{f}) \approx \frac{k-2}{2n} \pm \frac{\sqrt{2(k-1)}}{2n}. \tag{11.4}$$

If this estimate is applied to the marginal distributions of a p-dimensional density, $E(\bar{f}_n,\hat{f}_n)$ is asymptotically distribution free, its distribution depends only on the sample size n and the number k of bins. Clearly, this distribution freeness is absent for $E^*(\bar{f},\hat{f})$, the asymptotic distribution of the latter quantity needs to be investigated.

A third approach, probably the most rational one, would be to minimize the expected (asymptotic) mean square error at each x, that is, to minimize the sum

$$E\left|\frac{\bar{f}-\hat{f}}{\hat{f}}\right|^2 + \left|\frac{\bar{f}-\hat{f}}{\hat{f}}\right|^2$$

at each x. Unfortunately, it is difficult to turn this into a computationally feasible estimate.

As already mentioned, it may be wise to modify the definition of $E(f,g)$ so as to make it less sensitive to tail effects. An alternative possibility is to replace the factor f/g by $(c+f)/(c+g)$, where $c>0$ is a small constant depending on the sample size. In detail, we

propose to replace $E(f,g)$ by

$$E_\varepsilon(f,g) = \int \left| \frac{\varepsilon+g}{\varepsilon+f} - 1 - \log \frac{\varepsilon+g}{\varepsilon+f} \right| f \, dx; \tag{11.5}$$

note that $E_0 = E$ if f dominates g, and that $E_\varepsilon(f,g) > 0$ for $f \neq g$ (this follows easily from Lemma 7.2).

The problems just discussed are related to Stein estimation: we are estimating a point in a high dimensional (function) space, and it pays to decrease the total error by biasing the estimate. A detailed theoretical investigation is still outstanding.

We now return to a discussion of the error terms in (11.1). We need crude, order-of-magnitude, estimates; the random error term has already been discussed (we admit independence) in formulas (11.3) and (11.4). We only discuss the simplest case, namely kernel smoothers with a normal kernel; then the estimates $\hat{f}_\sigma$ are marginals of a genuine density $\hat{f}$, and the marginal densities satisfy a simple partial differential equation, as follows.

Lemma 11.1. Let f be a probability density on the real line, and let f_σ be the convolution of f with $N(0,\sigma^2)$. Then f_σ has bounded derivatives of all orders, and

$$\frac{\partial}{\partial(\sigma^2)} f_\sigma(x) = \frac{1}{2} \frac{\partial^2}{\partial x^2} f_\sigma(x), \quad \sigma > 0.$$

Proof. Essentially, the lemma states some well known properties of the heat equation. A simple probabilistic proof goes as follows. We first note that the characteristic function of f_σ is

$$\phi_\sigma(t) = \exp(-\tfrac{1}{2}\sigma^2 t^2)\phi(t),$$

where ϕ is the characteristic function of f, hence $t^k \phi_\sigma$ is absolutely integrable for all k if $\sigma > 0$, and thus all derivatives of f_σ exist and are bounded. Let $\tau^2 = \sigma^2 + \Delta\sigma^2$, and let Z be normal $N(0,\Delta\sigma^2)$. Then

$$f_\tau(x) = E(f_\sigma(x+Z)),$$

and if we take a Taylor expansion with remainder term inside the expectation, we obtain that

$$\frac{f_\tau(x) - f_\sigma(x)}{\Delta\sigma^2} = \frac{1}{2} f_\sigma''(x) + o(\Delta\sigma),$$

which proves the assertion. ∎

If f itself is smooth, the approximation is valid already for $\sigma = 0$, and we obtain that

$$f_\sigma(x) \approx f(x) + \frac{\sigma^2}{2} f''(x),$$

hence, according to (11.1),

$$E(f, f_\sigma) \approx \frac{\sigma^4}{8} \int \left| \frac{f''}{f} \right|^2 f \, dx$$

The approximation error $E^*(g,\hat{f})$ is an observable quantity, so there is nothing to estimate. But we note that $\hat{f}$ is not estimating f itself, but a smoothed version thereof, so the following proposition about the relative entropies of smoothed densities is of some interest.

Proposition 11.2. Let f and g be probability densities on the real line, and let f_σ and g_σ be their respective convolutions with $N(0,\sigma^2)$. Then, for $\sigma > 0$,

$$\frac{d}{d(\sigma^2)} E_{SK}(f_\sigma) = \frac{1}{2} \int \left| \frac{f_\sigma'}{f_\sigma} \right|^2 f_\sigma \, dx$$

and

$$\frac{d}{d(\sigma^2)} E(f_\sigma, g_\sigma) = -\frac{1}{2} \int \left| \frac{f_\sigma'}{f_\sigma} - \frac{g_\sigma'}{g_\sigma} \right|^2 f_\sigma \, dx,$$

(i.e. the derivatives are proportional to Fisher information and relative Fisher information respectively).

Proof. The proof is straightforward: use Lemma 11.1 to differentiate the defining expressions for entropy and relative entropy under the integral sign, and then integrate the result by parts. ∎

References

P.Billingsley (1968), *Convergence of Probability Measures*, Wiley, New York.

L.I.Bonera, D.Kendall and I.Stefanov (1971), Spline Transformations: Three New Diagnostics Aids for the Statistical Data Analyst (with discussion), *J.Royal Statist. Soc.*, B 33, 1-70.

Z.Chen and G.Li (1981), Robust principal components and dispersion matrices via projection pursuit, *Tech. Rep.*, Dept. of Statistics, Harvard University.

B.Efron (1977), Bootstrap methods: another look at the jackknife, *Ann. Statist.*, 7, 1-26.

D.Freedman and P.Diaconis (1980), On the histogram as a density estimator, *Tech. Rep. No. 103*, Dept. of Statistics, Stanford University.

J.H.Friedman and W.Stuetzle (1981), Projection pursuit methods for data analysis, SLAC PUB 2768.

J.H.Friedman and J.W.Tukey (1974), A projection pursuit algorithm for exploratory data analysis, *IEEE Trans. Computers*, C 23, 881-889.

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

P.J.Huber (1981b), Projection Pursuit, *Tech. Rep.*, Dept. of Statistics, Harvard University.

B.Silverman (1981), Using kernel density estimates to investigate multimodality, *J. Royal Stat. Soc.* (to appear).

C.J.Stone, (1975), Adaptive maximum likelihood estimators of a location parameter, *Ann. Stat.*, 3, 267-284.

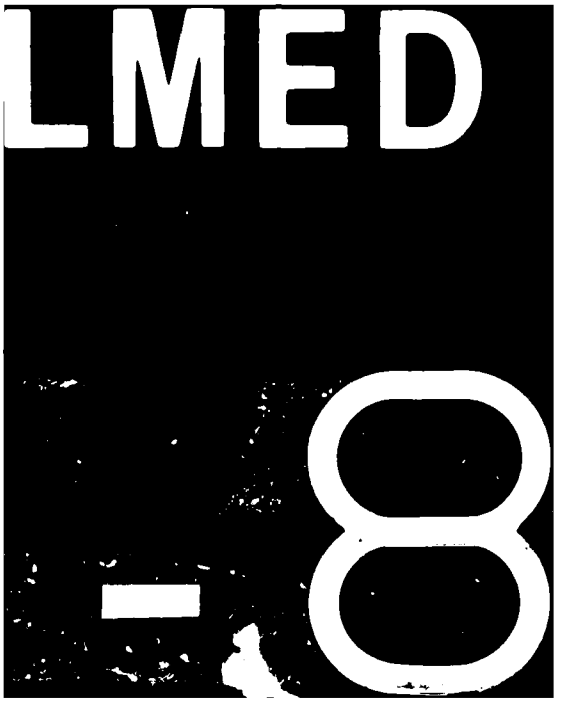
R.A.Tapia and J.R.Thompson (1978), *Nonparametric Probability Density Estimation*, Johns Hopkins Univ. Press, Baltimore.

P.A.Tukey and J.W.Tukey (1980), Methods for direct and indirect graphic display for data sets in 3 and more dimensions, *Proc. of the Sheffield Conference*, ed. by V.Barnett, Wiley (to be published).

G.Wahba and S.Wold (1975), A completely automatic french curve: Fitting spline functions by

cross-validation, *Comm. in Statistics*, 4, 1-17.

E.J.Wegman (1972), Nonparametric probability density estimation I. A summary of available methods, *Technometrics*, 14, 533-546.


LMED
8