SLAC-PUB-3394
July 1984
(M)

THE ESTIMATION OF SMOOTH CURVES*

Arthur Owen

Stanford Linear Accelerator Center

and

Department of Statistics
Stanford University
Stanford, CA 94305

Keywords: Transformations, Smoothing, Least Squares, Maximum Likelihood, Information, ACE, Projection Pursuit Regression, simultaneous confidence envelopes, bootstrap, non-parametric regression, smoothing splines, M-estimation, Method of Moments

Abstract

Smooth curves are often used to illustrate the relationship between two variables. They are also an important building block in many recent statistical models. A procedure to estimate such a curve is called a smoother.

This paper discusses currently available smoothers and introduces the class of maximum likelihood smoothers. A variety of other statistical techniques are shown to be applicable to the problem of smoothing, and some idea of the scope of models that can benefit from the use of smoothing is given.

(Submitted October 1983 to: The Canadian Journal of Statistics.)

---

* This work was supported by Department of Energy grant DE-AC03-76SF00515, by Office of Naval Research grant ONR-N00014-83-K-0340 and by U.S. Army Research Office grant DAAG29-82-K0056.

1 Introduction to Smoothers

1.1 Notation and motivation for smoothers

Smooth functions are a convenient way to illustrate the relationship between two random variables. The usual model is that

$$Y = S(X) + \epsilon$$

where the errors are i.i.d., and the explanatory variable X is either constant or independent of ϵ with a distribution that does not involve any of the parameters of the error distribution.

We use the data to estimate the above model obtaining the decomposition

$$y_i = \hat{S}(x_i) + r_i \qquad i = 1, \ldots, n$$

for the sample. The estimating procedure is called a smoother. A good smoother is one that captures 'just enough' of the non-linearity in the sample; capture too much and you are reproducing noise, capture too little and you might as well use a linear model.

Smooth functions of random variables play an important role in many non-linear statistical models and smoothers are a fundamental building block in the estimation of most such models. We often think of a smooth as the sample quantity corresponding to conditional expection. In such cases we will use $\mathcal{E}(\ Y|X\ )$ to denote the smooth of Y on X.

1.2 Examples of Commonly Used Smoothers

1.2.1 Histogram

The range of X is partitioned into k windows. In each window $\hat{S}$ is the average of the Y values corresponding to X's in the window. The choice of k and the partition is an important consideration. A common choice is to choose the windows so that they all have the same number of observations. The parameter

k governs the flexibility of the smoother. If $k = 1$ the smooth degenerates to a constant and if $k = n$ it merely reproduces the data. Large values of k tend to increase the variance of the fitted smooth while reducing its bias. All commonly used smoothers have some such tuning parameter to govern the smoothness of the output.

1.2.2 Running Average

A window is defined around each x_i. $\hat{S}(x_i)$ is the average of the y values in the i'th window. The window size is the tuning parameter. It is usually expressed as the number of observations in each window but can also be the length in the X direction of the windows.

1.2.3 Running Medians

As the name suggests the smooth takes the median value of the response variable in each window. The window length is the tuning parameter. Running median smoothers are discussed at length in Tukey (1977,Chapters 7,8).

1.2.4 Running Linear Fits

This is the same as the running average except that a line $L_i(x)$ is fit to each window and then $\hat{S}(x_i) = L_i(x_i)$. The lines are usually fit by least squares because fast updating formulae are available. For some tradeoff in computational cost a more robust line can be fit to each window.

1.2.5 Running Parametric Models

This further generalizes the running linear fit to running polynomials, running trigonometric polynomials or even more elaborate models. The tuning parameter is the combination of window size and the number of parameters applied to each window.

1.2.6 Spline Smoothers

The usual application of splines to smoothing is as follows: a subset of the data points are designated as knots and a p'th order polynomial is fit between

each pair of knots, subject to their being $p-1$ continuous derivatives at each knot. A histogram can be considered to be the simplest non-trivial spline.

The fitting is usually done by minimizing the sum of a goodness of fit criterion (such as sum of squared errors) and a penalty for nonsmoothness (such as the integral of the square of the second derivative of the estimated spline). The relative weight given to the goodness of fit criterion is the tuning parameter. Notice that the usual cubic interpolating spline is undesirable whenever the data are noisy. For a survey of the uses of splines in statistics see Wegman and Wright (1983).

1.2.7 Hybrid Smoothers

In a hybrid smoother a family of more basic smoothers $S_j, \quad j = 1, \ldots, k$ are applied to the data, and the smoother is estimated by amalgamating the basic smoothers. The basic smoothers usually differ from each other only in terms of the tuner. The choice of smoother is commonly made by cross-validation: each basic smoother is fit with the target point left out of its own window, the smoother that then has the minimum squared error is chosen. The idea is that the cross-validated squared error estimates the predictive squared error associated with the particular value of the tuner.

In the Supersmoother of Friedman and Stuetzle (1982) the basic smoothers are running linear fits, but the cross-validation is done locally. For each x_i the smoother is chosen that minimizes the squared errors near x_i. This allows the smoother to adapt to regional variations in the curvature of the underlying function and in the error variance – fine tuning, if you will. (The algorithm is also set up to favour larger intervals, unless there is a big difference in cross-validated squared error.)

Golub, Heath, and Wahba (1979) use cross-validation to choose the tuning parameter of a smoothing spline.

In the noncentral smoother of McDonald and Owen (1984) windows of various sizes are centered at each x_i. Other windows of various sizes are fit immediately

to the left and to the right of each point. The combination of the basic smoothers is done locally.

Hastie (1983) has proposed a method of combining any set of cross-validated basic smooths to obtain the linear combination of those smooths that minimizes an estimate of the mean squared error of such a combination. It considers the local squared errors and the local correlations between the smooths and there is an updating formula for these local weights.

1.3 What is Smooth?

It should be clear from the foregoing that the output of a smoother need not be smooth in any analytic senses (such as twice continuously differentiable). For example, the output of a histogram smoother is a step function. And yet, to be useful, the smoother cannot be arbitrarily flexible; $\hat{S}(x_i) \equiv y_i$ is like saying "the data is what the data is" and such tautologies are useless. At the other extreme a model such as linear regression can be thought of in simpler terms than as a smooth. A smoother should produce results between the trivial and tautological. Each smoother has its own definition of what is smooth.

1.4 Kernels and Weights

A kernel smoother is one which weights the observations in each window according to their position in that window. Usually the weights decrease as the distance between the point and the target point increases. The weight assigned to any given point depends on which window the point is currently being used in.

By contrast a weighted smoother is one in which each observation has a weight attached to it. This weight is often the inverse of some variance estimate for the point. If the i'th point were the average of n_i responses at x_i then it should get weight proportional to n_i.

Of course we can have weighted kernel smoothers. The weight attached to each point is then the product of the point-specific weight and the relevant kernel weight.

Having assigned weights to the points in the window, we use them by fitting weighted averages or weighted parametric models. In least squares estimation this is straightforward. In most other fitting schemes there is a sensible way to use the weights. With running medians the weights can be interpreted as point probabilities and we can take the median of the corresponding discrete distribution function. (For this we need positive weights that sum to unity over the window.) The smooth of Y on X with weights W will be denoted $\mathcal{E}(Y|X ; W)$.

1.5 Simple Smoothers

A smoother is said to be simple if it does not use point specific weights (kernels are o.k.) and estimates $\mathcal{E}(Y|X)$ in some sense. Simple smoothers are used as building blocks in more complicated smoothers. (Such smoothers need not be hybrid smoothers, nor need hybrid smoothers have simple basic smoothers.)

1.6 Properties of Smoothers

1.6.1 General

Here we discuss some of the commonly considered properties of smoothers. Most of the smoothers referred to above were designed to optimize one or more of the properties given below.

1.6.2 Pass Set

A smoother is said to pass the data if the smooth values equal the response values. The samples that a smoother would pass make up its pass set. (A smoother is tautological if it passes every data set.) Most smoothers will pass data in which the response is a constant for all X. Smoothers based on running linear fits will often pass linear functions. Running medians will pass monotone sequences. The noncentral smoother will pass piecewise linear (not necessarily continuous) functions subject to a condition on the size of the smallest piece.

1.6.3 Degree of Overfitting

In a sense the pass set reflects errors of type II that the smoother will not

make. We also want to guard against the possibility of type I errors – finding structure where none exists. The larger the tuning parameter the more unwanted structure will be found. One way to quantify the degree of overfitting is to generate white noise sequences of zero mean, smooth them, and record the sum of squared fits. The sampling distribution of the sum of squared fits will usually be positively skewed. In the case of standard normal noise and smoothers that regress the response on a p dimensional linear subspace this distribution is $\chi^2_{(p)}$. For more complicated smoothers the distribution is often well approximated by a chi-square or some other gamma distribution. In such cases the mean of the distribution of the sum of squared fits is a good indication of the degree of overfitting. We would not be very interested in a smoother with two or fewer degrees of overfitting, but we would want the degree of overfitting to be small compared to the sample size. For the super-smoother with the usual choice of window sizes the degree of overfitting is just over 3. (This depends slightly on the sample size which was 100 in this instance.) This is enough freedom to fit a 'bent line' and not much more.

1.6.4 Bias, Variance and Mean Squared Error

If the relevant moments exist, the bias, variance and mean squared error of a smoother can be defined, and are functions of X for any given error distribution. It is common for smoothers to have their greatest biases where the curvature of $S(\cdot)$ is largest compared to the scale of the errors. The tendency is for smoothers to 'fill in the valleys and erode the hills'.

In a hybrid smoother with local tuning, the bias is reduced by shortening the window size where the underlying function seems to have strong curvature, at the expense of increasing the local variance. The tradeoff is designed to minimize the mean squared error everywhere.

Twicing is a technique of Tukey (1977) for reducing the bias of a smoother. The smoother is applied to its own output. It fills valleys and erodes hills in its own output by an amount that we can compute. We then make an additive correction of that amount to the original smooth. This method can be shown to

correct for quadratic behaviour.

End effects are another source of bias and variance problems. Near the ends of the observed range of X the windows of most smoothers are shrunken and noncentral. This increases both bias and variance. The problem is particularly acute for running averages, and treating this problem was the motivation behind running linear fits.

If the set of X values doesn't have ends (an example is the unit circle) there are obviously no end effects. At the other extreme if the X values take values in a high dimensional space the end (or surface) effects can be considerable. For most of the smoothers treated here the X values come from an interval in the real line.

1.6.5 Equivariance

A smoother is affine equivariant if the smooth of $a + bY$ on X is a plus b times the smooth of Y on X, for any scalars a and b. Most smoothers are affine equivariant. If we are smoothing a function over a domain such as the unit circle we might impose a rotational equivariance constraint on the smoother.

1.6.6 Linearity

A smoother is linear if it can be written $\hat{S}(X) = C(X)Y$ where Y is the response vector and $C(X)$ is a matrix depending on X. Histograms and running least squares regressions are linear but running medians are not. Some hybrids are close to linear in that they always choose a combination of the basic smooths that is equivalent to a linear smoother. They are not linear if that choice is made using the observed response as is usually the case.

1.6.7 Idempotence

An idempotent smoother is one that passes its own output. They usually arise as projections onto some space of functions of X.

1.6.8 Consistency

A smoother $\hat{S}$ is consistent for a function S if the distance $d(\hat{S}, S)$ converges to zero as $n \to \infty$. The distance could be integrated squared difference or integrated absolute difference. The consistency is strong or weak if the convergence is almost sure or in probability, respectively. If the smoother can be expressed as a statistical functional of the joint empirical distribution function of X and Y, and that functional applied to the true joint distribution function of X and Y produces S^* with $d(S, S^*) = 0$ then then the smoother is Fisher consistent for S.

Consistency results are usually obtained by having the tuning parameter increase without bound as the number of sample points increases, but at a suitably slow rate. One often needs to make some assumption to the effect that S is well-behaved, such as measurability or continuity except at a finite number of points.

1.6.8 Robustness

In the sequel a number of smoothers are developed for use with specific distributional assumptions, and it is also shown that some common smoothers make implicit distributional assumptions. We would not want a smoother to go badly wrong if the distributional assumptions were violated.

Robustness can be built into a smoother either locally or globally. The local method is to use robust techniques at the lowest levels (e.g. the windows) of the smoother. One global approach, taken by Friedman and Stuetzle (1982) is to estimate a preliminary smooth in a robust way (running medians), 'reject' observations that are too far from the robust smooth, and smooth the other observations. Another global method designed to enhance robustness is the Cauchy maximum likelihood smoother of section 4. (It often happens that in extending a notion to smoothers there is a choice between local and global approaches.)

Robustness measures commonly used are the breakdown level and the influence function of a procedure. The influence of the Kaplin-Meier estimate (of the

probability of 'survival' beyond time $t > 0$) was derived by Reid (1981).

2 Least Squares Smoothers

2.1 Ordinary Least Squares Smoothers

Consider a location model $Y = f(X) + \epsilon$ where the errors are i.i.d. with mean zero and finite variance. We wish to find the function $f(\cdot)$ that minimizes $\mathcal{E}(\ (y - f(x))^2\ )$. The well known solution is $f(\cdot) = \mathcal{E}(\ Y|X = \cdot\ )$ and so we estimate $f(\cdot)$ with a simple smoother.

One way of arriving at this conclusion which extends to other situations is as follows. We start by assuming that partial differentiation with respect to each $f(x)$ commutes with the expectation above. We find the best value of f for any value of x by setting

$$\begin{aligned} 0 &= \frac{\partial}{\partial f(x)} \mathcal{E}(\ (Y - f(X))^2\ ) \\ &= 2\mathcal{E}(\ f(X) - \mathcal{E}(\ Y|X\ )\ ) \end{aligned}$$

which is satisfied by $\hat{f}(\cdot) = \mathcal{E}(\ Y|X = \cdot\ )$.

By using a criterion involving the expectation we can optimize over a large number of parameters. By expressing $\mathcal{E}(\ \cdot\ )$ as $\mathcal{E}(\ \mathcal{E}(\ \cdot|X\ )\ )$ we obtain a solution that can be estimated with a smoother. (Since we wanted a function of X we conditioned on X.) This simple-minded approach turns out to be surprisingly applicable. It is also possible to use other aspects of the distribution of the random variables, such as the median or maximum, but the expectation is the simplest choice.

2.2 Weighted Least Squares Smoothers

The setup here is as above except that we now wish to minimize $\mathcal{E}(\ w(X)(Y - f(x))^2\ )$ where $w(X) > 0$. Setting the partial derivatives to zero yields $0 = \mathcal{E}(\ w(X)(f(X) - \mathcal{E}(\ Y|X\ ))\ )$ which has solution $\hat{f}(\cdot) = \mathcal{E}(\ Y|X = \cdot\ )$. We are left wondering what happened to the weights–clearly they must enter somewhere.

We get them back by realizing that although we are using conditional ex-

pectation we don't have to use a simple smoother–a weighted smoother will also estimate conditional expectation. Our problem is identical to minimizing the ordinary expected sum of squares when $Y|X$ has variance $1/w(X)$, and so we should use weights proportional to $w(X)$. In a finite parameter linear regression context, using the wrong weights gives unbiased estimates that do not have minimum variance, and the same effect is plausible here. There should be some kind of Gauss-Markov theorem to the effect that the correct weights lead to best linear unbiased smoothers, although the class of smoothers to be considered will have to be carefully delineated.

2.3 Generalized Least Squares Smoothers

Suppose we wish to minimize $\mathcal{E}(\ (Y - f(X))'V^{-1}(Y - f(X))\ )$ where V is a positive definite (variance-covariance) matrix and Y and $f(X)$ are vectors of observed and fitted values respectively.

The local approach to this problem is to use running generalized least squares fits. If the window-level models are parametric regressions this is straightforward.

A global approach to the same problem is to premultiply the response vector by C, a matrix square root of V, smooth by a simple smoother and multiply the resulting smooth vector by C^{-1}.

3 Maximum Likelihood Smoothers

3.1 Overview

One would expect that the least squares smoothers would provide good results when the errors are Gaussian but not when the errors come from a heavy tailed distribution such as the Cauchy. We strengthen this notion by showing that the least squares smoother is the maximum likelihood smoother for Gaussian data.

Then we go on to derive the maximum likelihood smoothers for use against a variety of other distributions and along the way mention some techniques for solving the resulting likelihood equations. The likelihood equations for general

location, location-scale, and exponential family models are then given.

We turn to inference by introducing a notion of the Fisher information in a data set for the underlying smooth function. This information can be used to obtain bootstrap confidence envelopes for a smooth curve. Inverting the envelopes provides a measure of the significance of the estimated curve. Other large-sample inference methods are also generalized.

We conclude by considering the effects of the distribution of the explanatory variable X and showing how some non-linear models can be estimated by maximum likelihood.

As an historical note, the first maximum likelihood smoother to be used was the Box-Cox family of transformations (Box and Cox(1964)). Although they did not call their procedure a smoother, it estimated a variance-stabilizing transformation (smooth curve) by maximum likelihood from a one or two parameter family of transformations.

3.2 Some Examples

3.2.1 Gaussian Location

The Gaussian location model is the *E. Coli* of this subject. We have $Y|X \sim N(\mu(X), \sigma^2)$ where the observations are independent of each other and σ^2 is a constant. The likelihood is

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \mu(x_i))^2}$$

but it is clear that we can't just maximize the observed likelihood L as a function of $\mu(\cdot)$ since that collapses to the tautological solution $\hat{\mu}(x_i) = y_i \quad i = 1, \ldots, n$.

A more reasonable approach is to set the expected value of the score function to zero; that is to maximize the expected value of the log likelihood. This leads easily to the ordinary least squares smoother. We can estimate $\mu(\cdot)$ and σ jointly by this method, getting $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}(x_i))^2$.

3.2.2 Cauchy Location

We have the situation of section 3.2.1 except that the errors are independent standard Cauchy random variables. We would not want to use a least squares smoother on such data. It is also reasonable that the Cauchy maximum likelihood smoother would be robust against outlying values of Y.

The likelihood equations are

$$\mathcal{E}\left(\frac{Y - \mu(X)}{1 + (Y - \mu(X))^2} \right) = 0$$

which, just as in the single parameter case, cannot be expressed in closed form. (Notice that the expectation of the numerator does not exist, but that the expectation of the ratio does.) An iterative procedure is called for.

Notice that a sufficient condition for the likelihood equations is

$$\mu(x) = \frac{\mathcal{E}\left(\frac{Y}{1+(Y-\mu(X))^2} \mid X = x \right)}{\mathcal{E}\left(\frac{1}{1+(Y-\mu(X))^2} \mid X = x \right)}$$

and we can use this to estimate $\mu(\cdot)$ iteratively. We obtain $\mu^{(i+1)}(\cdot)$ by evaluating the right hand side of the condition above with $\mu^{(i)}(\cdot)$. When this produces no change in $\mu(\cdot)$, we have a solution to the likelihood equations. This algorithm is a direct generalization of one we might use to estimate the Cauchy location parameter in an i.i.d. sample. We could start the iteration with $\mu \equiv 0$. (An iterative estimation of location should be capable of starting with 0 and an iterative estimation of scale should be capable of starting with 1.)

3.2.3 Logistic Location Model

If one obtains the likelihood equations for $Y - \theta(x)$ having the standard logistic distribution, conditions on X and rearranges the result, one obtains

$$\theta(x) = \log 2\mathcal{E}\left(\frac{e^Y}{1 + e^{Y-\theta(X)}} \mid X = x \right)$$

which can be used in an iterative fashion to estimate $\theta(\cdot)$.

3.2.4 Binomial Probability

The model is that $Y \sim Bin(n(X), \theta(X))$ where $\theta(\cdot)$ is smooth, the Ys are independent, and every $n(X)$ is known. The likelihood equation is equivalent to

$$\mathcal{E}\left(\frac{Y - n(X)\theta(X)}{\theta(X)(1-\theta(X))} \right) = 0$$

which suggests using an iterative procedure based on smoothing Y/n on X with weights $\frac{n(X)}{\theta(X)(1-\theta(X))}$. The algorithm can be started with weights proportional to $n(X)$.

3.2.5 Poisson Intensity

Suppose we have independent observations from $Y \sim Poi(\lambda(X))$ where $\lambda(X)$ is smooth. The likelihood equation is

$$\mathcal{E}\left(\frac{Y}{\lambda(X)} - 1 \right) = 0$$

which suggests an iterative scheme

$$\lambda^{(i+1)}(\cdot) \leftarrow \mathcal{E}\left(Y|X = \cdot \,;\, 1/\lambda^{(i)}(X) \right)$$

starting with $\lambda^{(0)} \equiv 1$.

3.2.6 Non-Regular Cases

When the usual regularity conditions don't hold for the underlying model there is no reason to hope it will when the parameter is replaced by a smooth function of X. Consider the model in which the Y are independent Unif$[a(X), b(X)]$ where $a(x)$ and $b(x)$ are smooth. The likelihood is

$$\prod_{i=1}^{n} \frac{1}{b(x_i) - a(x_i))}$$

if all $y_i \in [a(x_i), b(x_i)]$ and zero otherwise. Expected scores won't work here; what is called for is the smallest possible intervals that are guaranteed to hold

all the data. We need to estimate something like $a(\cdot) = Max(Y|X = \cdot)$, $b(\cdot) = Min(Y|X = \cdot)$.

Another non-regular case is the double exponential location distribution. In the i.i.d. sample case, the maximum likelihood estimate is the median. (The non-regularity is that the derivative of the log-likelihood does not exist at the sample points.) The likelihood equations here reduce to

$$0 = \mathcal{E}(\ P(Y > \theta(x)) - P(Y < \theta(x))\ )$$

which is satisfied by $\theta(x) = MEDIAN(Y|\ X = x\ )$.

3.3 Location Maximum Likelihood Smoothers

We generalize maximum likelihood smoothers to location families $Y|X \sim$ ind $g(y - \theta(X))$ for a sufficiently regular density g. The likelihood equation is

$$\mathcal{E}(\ -\frac{g'(Y - \theta(X))}{g(Y - \theta(X))}\ ) = 0$$

which will often be estimable by an iterative method.

There is a striking similarity of the above to $-\mathcal{E}_{\theta=0}(g'(y_{(j)})/g(y_{(j)}))$ which is the optimal weight for the j'th order statistic in a rank test of $\theta = 0$ when θ is a constant. Since the optimal rank test for location shifts in a logistic distribution is the Wilcoxon test, perhaps it is fair to call the logistic maximum likelihood smoother the 'Wilcoxon Smoother'.

If it happens that $h(\cdot) = g'(\cdot)/g(\cdot)$ is invertible and vanishes at the origin (as it would for many symmetric unimodal densities) then for any trial smooth parameter $\theta(\cdot)$, there is a 'smooth residual' $Res(x) = \mathcal{E}(\ h(Y - \theta(X))|\ X = x\ )$ to guide us in improving $\theta(\cdot)$. If this residual is identically zero, then we have a solution to the likelihood equations. Otherwise we can update via $\theta^{(i+1)}(\cdot) \leftarrow \theta^{(i)}(\cdot) - h^{-1}(Res(\cdot))$.

3.4 Location Scale Maximum Likelihood Smoothers

3.4.1 Normal Case

In this model the conditional distribution of Y given $X = x$ is $N(\mu(x), \sigma^2(x))$, where $\mu(x)$ and $\sigma(x)$ are smooth and the Ys are independent of each other.

The likelihood equations are

$$0 = \mathcal{E}\left(\frac{\mu(X) - Y}{\sigma^2(X)} \right)$$
$$0 = \mathcal{E}\left(\frac{-1}{\sigma(X)} + \frac{(Y - \mu(X))^2}{\sigma^3(X)} \right)$$

which are to be solved jointly. We can estimate the model by alternating between $\mu(X) \leftarrow \mathcal{E}(Y|X; \frac{1}{\sigma^2(X)})$ and $\sigma^2(X) \leftarrow \mathcal{E}((Y - \mu(X))^2|X ; \frac{1}{\sigma^3(X)})$ which we can initialize by setting $\sigma^2 \equiv 1$.

Not only can the smoother adapt itself to heteroscedasticity, it provides an estimate of the local variance which may be of interest in its own right.

3.4.2 General Case

The general location-scale model $1/\sigma(x)g((y-\mu(x))/\sigma(x))$ has likelihood equations

$$0 = \mathcal{E}\left(-\frac{g'(Z)\frac{1}{\sigma(X)}}{g(Z)} \right)$$
$$0 = \mathcal{E}\left(-\frac{1}{\sigma(X)} - \frac{g'(Z)\frac{1}{\sigma^2(X)}}{g(Z)} \right)$$

where $Z = (Y - \mu(X))/\sigma(X)$, which will often be estimable by an iterative method.

With a scale model we can take $\mu \equiv 0$ in the second equation, and drop the first equation.

3.5 Location-Scale-Correlation Models

We might complicate the Gaussian location-scale model by introducing conditional covariances of the form $Cov(Y, Y') = \rho(x, x')\sigma(x)\sigma(x')$ where $\rho(\cdot, \cdot)$ is a smooth correlation function.

If ρ is known for all pairs of x values, then we can use a generalized least squares smoother (adapted to estimate the scale parameter as well). Otherwise we might consider alternating between the generalized least squares smoother and a method of estimating the correlations.

We take ρ to be a univariate function $\rho_0(d(x, x'))$ where $d(\cdot, \cdot)$ is a metric. (For example d could be 0 or 1 depending on whether its arguments were equal or not, or $d(x_i, x_j)$ could be $|i - j|$ for some ordering of the observations, or d could be the euclidean distance between its arguments.) We require $\rho_0(0) = 1$, and in general other conditions will need to be applied to guarantee that ρ is a bona fide correlation. (A sufficient condition on ρ_0 is that it is positive, convex and decreasing.)

For fixed $\mu(X)$ and $\sigma(X)$ the likelihood equation for ρ is

$$0 = \mathcal{E}\left(-\frac{1}{2}\frac{\frac{\partial}{\partial\rho_0(d)}|R|}{|R|} - \frac{1}{2}tr(EE')\frac{\partial}{\partial\rho_0(d)}tr(R^{-1}) \right)$$

where E is the vector of standardized residuals $(y - \mu(x))/\sigma(x)$ and R is the matrix of correlations.

If for a given correlation model the likelihood equations are intractable an intuitive estimator is

$$\rho(\cdot) = \mathcal{E}\left(E(x)E'(x) | d(x, x') = \cdot \right)$$

although this is likely to be less efficient than the likelihood equations, which I conjecture are more like a weighted version of the above. Of course the smoother used to estimate ρ must be constrained to always produce a valid correlation function. One way to do this is to estimate the correlation as a convex combination of some basic correlation functions such as 'symmetric triangular' functions.

3.6 Higher Moments

If one has a density model with coefficients for higher moments, then in principle one can solve the likelihood for those coefficients to estimate moving versions of skewness, kurtosis and so on. A cruder approach is to raise the residuals from a location-scale smooth to the power k and smooth the powers against X, but this does not correspond to maximum likelihood. A more systematic approach is discussed in the method of moments smoother in section 4.1.

3.7 Exponential Family Models

We now consider a one parameter exponential family with the parameter depending in a smooth way on X. That is the Y are independent with conditional density

$$exp\{a(\theta(x))b(y) + c(\theta(x)) + d(y)\}.$$

The likelihood equation is

$$0 = \mathcal{E}(\ a'(\theta(X))b(Y) + c'(\theta(X))\ )$$

which in many cases is simply solved.

If $a(\cdot)$ and $c(\cdot)$ have two continuous derivatives we can write

$$\begin{aligned} a'(z) &= \alpha(z) + \beta(z)z \\ c'(z) &= \gamma(z) + \delta(z)z, \text{ where} \\ \beta(z) &= a''(z); \alpha(z) = a'(z) - \beta(z)z \\ \delta(z) &= c''(z); \gamma(z) = c'(z) - \delta(z)z. \end{aligned}$$

The likelihood equation becomes

$$0 = \mathcal{E}(\ (\alpha(\theta(X)) + \beta(\theta(X))\theta(X))\mathcal{E}(\ b(Y)|X\ ) + \gamma(\theta(X)) + \delta(\theta(X))\theta(X)\ )$$

which we estimate iteratively via

$$\theta^{(i+1)}(\cdot) = \frac{-\alpha(\theta^{(i)}(\cdot))\mathcal{E}(\ b(Y)|X = \cdot\ ) - \gamma(\theta^{(i)}(\cdot))}{\beta(\theta^{(i)}(\cdot))\mathcal{E}(\ b(Y)|X = \cdot\ ) + \delta(\theta^{(i)}(\cdot))}.$$

This technique is called delinearization because it resembles complex demodulation. (It has a long history at S.L.A.C.) It amounts to a seperate first order Taylor's series approximation at each point of the curve. On any data set, the slopes and intercepts are obtained for each point in the sample.

If either $a'(\cdot)$ or $c'(\cdot)$ is invertible then we can obtain an updating equation for $\theta(\cdot)$ in the way used for the Cauchy and the Logistic models.

The generalization to multiparameter exponential families is straightforward. The smooth parameters might all depend on the same covariate X or there may be several covariates each with one or more smooth parameters depending on it.

3.8 Fisher Information

3.8.1 The Appropriate Notion

There are three ways to approach the Fisher information: the observed information, the expected value of the observed information, and the information at the maximum likelihood estimate. In the finite parameter case the first two are the same. In the smooth parameter case the first of these is tautological, the last is trivial and the other is useful.

To fix ideas we consider maximum likelihood smoothing in the Gaussian location model with unit variance. The score vector (a sample version of the 'score functional') has i'th element $e_i = y_i - \mu(x_i)$ and we want to estimate an underlying Fisher information surface for μ. We will only be able to estimate the surface over a finite grid of points, but can extend it by assuming that is smooth. The information, in direct analogy with the finite parameter case is $I(\mu(x), \mu(x')) = \mathcal{E}(\ E \times E'\ )$.

For any maximum likelihood estimate of $\mu(\cdot)$ the Fisher information at that estimate is 1 if $x = x'$ and 0 otherwise, because of the conditional independence of Y given X. That is the 'surface' is 0 over all but the diagonal of $\mathcal{X}^2$ where it is 1. This tells us nothing about dependecies among the estimated values of μ.

The observed Fisher information is the matrix with $\hat{E}(y_i, x_i)\,\hat{E}(y_j, x_j)$ as its

(i, j) element. If as is common, the X_i are distinct with probability 1, we will find that no matter how large the sample, the information estimate uses at most two observations at any point.

The expected value of the observed information is

$$\begin{aligned} I(\mu(X), \mu(X')) =& \mathcal{E}(\ (Y - \hat{\mu}(X))(Y' - \hat{\mu}(X'))\ ) \\ =& \mathcal{E}(\ \mathcal{E}(\ (Y - \hat{\mu}(X)(Y' - \hat{\mu}(X'))|X, X'\ )\ ) \end{aligned}$$

which we can estimate with a two-dimensional smoother using our estimate of $\mu(\cdot)$. This is just the observed information, smoothed.

If our smoother were a finite parameter regression we could obtain an estimate of the variance-covariance matrix for the parameters. From this we could obtain an estimate of the variance-covariance matrix of the vector $\hat{\theta}$ of point specific mean estimates. This matrix would be of order n, but its rank would in general be equal to the number of parameters in the regression. We could take a generalized inverse of this matrix as the information estimate for the vector. Without such a model, we start from the low rank estimate of the information matrix and take a generalized inverse as an estimate of the variance-covariance matrix of the vector of estimated means.

Call this information estimate $\hat{I}_\mu$. It is reasonable to expect that $\hat{\mu}(\cdot) - \mu(\cdot)$ will have an asymptotic normal distribution with mean zero and variance-covariance matrix $\hat{I}_\mu^-$, a generalized inverse of $\hat{I}_\mu$. (That is the μ estimates for any fixed set of observations should have an asymptotic covariance obtained from 'their components' of the information estimate as the number of subsequent observations tends to infinity.)

3.8.2 Bootstrap Confidence Envelopes

When the asymptotic normality referred to above holds, we can use the pivotal quantity $(\hat{I}_\mu^-)^{\frac{1}{2}}(\hat{\mu}(\cdot) - \mu(\cdot))$ to form confidence regions for μ. The distribution of this quantity does not depend on μ, but it does depend in a possibly complicated way on the smoothing algorithm used. The distribution of $\hat{\mu}(\cdot) - \mu(\cdot)$ is that of a vector of independent standard normal variables multiplied by a matrix

square root of a generalized inverse of the information estimate, and using this we can 'invert the pivot'. We generate a large number of normal vectors, multiply them by the matrix square root, and add $\hat{\mu}(\cdot)$. In this way we estimate a bootstrap confidence measure for $\mu(\cdot)$. Denote by $\mu^{*j}(\cdot)$ the j'th such bootstrap value. (For a discussion of the bootstrap see Efron (1982).) This is easier than resampling the data and smoothing the bootstrap samples a large number of times, although that option is always available if we suspect that the sample size is too small for asymptotic normality to hold.

We can obtain a (usually degenerate) confidence ellipsoid for the vector of μ values directly from the information matrix. However, we are more interested in a confidence envelope— two curves between which $\mu(\cdot)$ lies with a specified degree of confidence. Such an envelope corresponds to a rectangular confidence region for the vector, for which no convenient analytic expressions are available. It is reasonable to have the envelope's width at x proportional to the estimated standard error of $\hat{\mu}(x)$. In that case we need merely record for each bootstrap sample the largest ratio of the form $|\hat{\mu}(x_i) - \mu^{*j}(x_i)|/\hat{\sigma}(x_i)$ where $\sigma(x_i)$ is the estimate of the standard deviation of $\hat{\mu}(x_i)$ obtained from the diagonal of the information estimate. If $100(1-\alpha)$ percent of these suprema are less that k_α then $\hat{\mu}(\cdot) \pm k_\alpha \sigma(\cdot)$ is a $100(1-\alpha)$ percent central confidence envelope for $\mu(\cdot)$. One-sided envelopes can be obtained in a similar way.

If our model has two parameters of the same argument X then the joint values of the parameters can be thought of as a space curve or trajectory. We then can consider putting confidence tubes around the estimated trajectory. If the model has two parameters of different arguments (that are not related in a degenerate way) then the confidence region for the estimated surface is a four dimensional object. For any trajectory in the estimated surface the confidence region is a tube.

3.8.3 Significance Levels for Curves

3.8.3.1 Inverting Confidence Envelopes

Within most models with smooth curves are simpler models in which the curves are zero or linear. We need techniques for estimating the significance of the non-linearity of a given transformation.

A simple graphical approach is to invert the confidence envelopes discussed above. For example if we have a set of central α-level confidence envelopes for $\mu(\cdot)$, we can consider the greatest value of α for which the zero function is entirely within the confidence envelope. The significance of the non-linearity is obtained in a similar way: find the greatest value of α for which the confidence envelope contains a line. The generalization to the significance of the difference between $\mu(\cdot)$ and any other set of functions is similarly found, although the calculations could be horrendous if the set of functions is unusual. Reasonable sets of functions are: monotone functions, convex functions, sinusoids, exponentials, positive functions, and functions that are everywhere greater than a prespecified function. The latter two examples would best be evaluated using one-sided confidence envelopes.

In principle we could invert higher dimensional confidence regions to assess the significance of complicated relations between two or more smooth curves. In practice many statisticians will be understandably reluctant to invert any confidence region that cannot be displayed on a graphical output device.

3.8.3.2 Likelihood Ratio Methods

If we fit a p parameter model M_1 with a q parameter submodel M_2 in it, then we expect that the log likelihood ratio of M_1 to M_2 should be asymptotically $\chi^2_{(p-q)}$ when M_2 is true.

It is often clear how many parameters M_2 has, since M_2 typically specifies constancy, linearity, or some other finite parameter model. To generalize Wilks likelihood ratio test we need to generalize the notion of the number of parameters in a model. One approach to this problem is via the degrees of overfitting introduced in section 1.6.1.

3.8.3.3 F-Statistic Methods

If we have a good notion of the degrees of freedom of a smoother or model we can use it to mimic the classical F-tests.

If we constrain all the smooth functions in a model to be linear then it will generally be easy to count up the degrees of freedom; call it d_L. If our general model has degrees of freedom $d_G > d_L$ then we can set up a psuedo-ANOVA table:

SOURCE	'D. F.'	S. S.	M. S.
Linearity	d_L	SSF_L	SSF_L/d_L
Non-Linearity	$d_G - d_L$	$SSF_G - SSF_L$	$(SSF_G - SSF_L)/(d_G - d_L)$
Error	$N - d_G - 1$	SSE	$SSE/(N - d_G - 1)$
Total	$N - 1$	SST	$SST/(N - 1)$

where SSF refers to the sum of squared fits (about the mean) of the model corresponding to its subscript and SST is the total sum of squares of the response about its mean. The choice of F-test will depend as usual on whether the effects are considered to be random or fixed, a matter which leads to lively discussions even in the ordinary ANOVA formulation.

If the model contains several smooth curves, we can partition the sum of squares for non-linearity to see what the source of the non-linearity is. For an example of this in which two transformations have a non-zero interaction see Owen(1983a, pp15-17). To test in such situations one would have to obtain a value for the relevant interaction degrees of freedom.

There is as yet no Cochran's theorem for this problem.

3.9 The Distribution of X

Suppose that X is a (vector) random variable with density $h(x)$ and that conditionally on $X = x$, Y has density $g(y|X = x; \theta(\cdot))$, where $\theta(\cdot)$ is a smooth parameter. The overall likelihood is

$$L = \prod_{i=1}^{n} h(x_i)g(y_i|X_i = x_i; \theta(\cdot))$$

which is equivalent to the likelihoods considered above, provided that $h(x)$ does not involve $\theta(\cdot)$.

3.10 Maximum Likelihood Estimation of Nonlinear Models

3.10.1 Nonlinear Regression

We have a location model where $Y|X \sim g(Y - \sum_{j=1}^{k} \phi_j(X^j))$ and $X = (X_1, \ldots, X_k)'$ is the vector covariate.

We can estimate this model by starting with initial values for all the $\phi_j(\cdot)$ and updating them in sequence by solving the likelihood equations for one function at a time until convergence.

When the error distribution is Gaussian, this reduces to a special case of the ACE model of Breiman and Friedman (1982), and the resulting estimation is that of the ACE algorithm.

3.10.2 The ACE model

Here we have the location model of 3.10.1, but we apply it to $\theta(Y)$ instead of Y itself, where $\theta(\cdot)$ is smooth and is to be estimated along with the other parameters. Typically the ACE model is estimated with some sort of least squares smoother and so the error model is implicitly Gaussian.

If we maximize the likelihood over all smooth functions we get into trouble. The algorithm finds the global optimum in which all the transformations are identically zero (or constant).

Breiman and Friedman get out of this pitfall by imposing the constraints $\mathcal{E}(\ \theta(Y)\ ) = 0$ and $\mathcal{E}(\ \theta(Y)^2\ ) = 1$ on the parameter $\theta(\cdot)$ After each estimation of $\theta(\cdot)$ they rescale $\theta(\cdot)$ via

$$\theta(\cdot) \ \leftarrow \ \frac{\theta(\cdot) - \mathcal{E}(\ \theta(Y)\ )}{\sqrt{\mathcal{V}ar(\theta(Y))}}.$$

This corresponds to maximization of the expected log likelihood using two Lagrangian multipliers by an algorithm that estimates in sequence: $\theta(\cdot)$, the multipliers, and the $\phi_j(\cdot)$ until convergence.

The ACE algorithm also allows any of the smooth curves to be constrained to be monotone. This is done by fitting an unconstrained smooth to the cor-

responding variable (at each iteration) and then finding the closest monotone function to the result via an algorithm of Kruskal(1965). Since the closeness measure chosen is least squared error, this corresponds closely to constrained maximum likelihood equation. (For a discussion of this technique see Friedman and Tibshirani (1983).)

3.10.1 Projection Pursuit Regression

The projection pursuit regression model of Friedman and Stuetzle (1981) fits a location model

$$Y|X \sim \text{ind } g(Y - \sum_{m=1}^{M} f_m(\alpha_m' X))$$

where the $f_m(\cdot)$'s are smooth real valued functions, the α_m's are unit vectors, and X is a vector of covariates. They estimate the model by least squares. The smooth location parameter can be made arbitrarily close to the conditional expectation by including enough terms, whereas the ACE model as written is not so general.

3.10.4 Predictive ACE

The predictive ACE model of Friedman and Owen(1984) fits a model for Y with location $f(\sum_{j=1}^{k} \phi_j(X^j))$, where $f(\cdot)$ and the $\phi_j(\cdot)$ are smooth. The constraints imposed are $\mathcal{E}(\ \phi_j(X^j)\ ) = 0, \quad j = 1, \ldots, k$ and $\mathcal{E}(\ (\sum_{j=1}^{k} \phi_j(X^j))^2\ ) = 1$. As with the ACE model the constraints are imposed by adjusting each constrained item after each estimation of it. The estimation is by least squares.

3.10.5 Other Models

Recent work at SLAC involves extending models involving smooth curves to various special domains in statistics. Examples include Time Series (McDonald(1983), Owen(1983a)), Survival Analysis (Tibshirani(1982)) and Multivariate Statistics (Hastie (1983)).

No doubt similar work is going on at other sites as well.

4 Other Methods of Smoothing

4.0 Other Notions

Maximum likelihood and least squares are not by any means the only well developed areas of modern statistics. It seems that most other techniques of statistics are applicable to models involving smooth curves, and a few examples are given below.

4.1 Method of Moments Smoother

Suppose we wish to estimate the first k moving moments of Y, which depend on X in a smooth way, and we are willing to assume that there is an interval containing the origin over which the moment generating function of $Y|X = x$ exists for all x.

We begin by smoothing e^{tY} against X with a simple smoother. We do this for each t values on a grid near the origin.The result is an estimate of the conditional moment generating function of Y given X near the origin. We get the moments by estimating the derivatives of the moving moment generating function near the origin. (Our grid must have more than k points in it, or we will not be able to estimate the k'th derivative.)

Next we realize that $\mathcal{V}ar(e^{tY}|X = x) = \mathcal{E}(\ e^{2tY}|X = x\ ) - \mathcal{E}(\ e^{tY}|X = x\ )^2$ and so we can improve our initial estimate by updating with a weighted smooth (the weights being obtained from the initial estimate). This leads us to choose as our grid of t values a geometric progression with ratio 2 on either side of the origin. At the first updating stage we cannot update the smooth for the largest positive grid point, and at each subsequent stage one fewer positive grid point can be estimated. A similar situation holds for the negative grid points.

If we were extremely fastidious we would note the linear combination of the smooths that is to be used for each moment, put a quadratic loss function on the error of the moment vector and pick our updating weights accordingly. (The author is not this fastidious.)

It is generally difficult to estimate the higher moments of a distribution and moving higher moments must be still more difficult. While this method seems to provide an approach, it will be likely to require very large sample sizes to get reliable estimates of the moments. Since each moment is a linear combination of some estimates of the moment generating fuction, it will be relatively easy to obtain confidence envelopes to assess the reliability of the moment estimates. Note also that the moment estimates will be correlated.

If we are unwilling to assume that a moment generating function exists we can estimate a moving characteristic function. That is we smooth e^{itY} versus X for a grid of t values. Since the expected squared modulus of e^{itY} is always unity we need not bother with weighted estimates, so there may be no particular advantage to a geometric progression grid. We can obtain moving moments from the derivatives of the moving characteristic function.

Another, somewhat anachronistic, approach to this problem is to estimate a moving version of the Pearsonian family of frequency curves.

4.1 M-estimate Smoothers

The M-estimate of location generalizes the maximum likelihood estimation of location. (For a thorough discussion of M-estimation see Huber(1981).) The maximum likelihood estimate is that (constant) θ such that

$$\sum_{i=1}^{n} \frac{f'(y_i - \theta)}{f(y_i - \theta)} = 0$$

whereas for M-estimation $f'(\cdot)/f(\cdot)$ is replaced by a function $\psi(\cdot)$ that need not correspond to any particular density $f(\cdot)$. Much effort has been put into chosing $\psi(\cdot)$ to obtain robustness at as small as possible a cost in efficiency. For example if $\psi(\cdot)$ is bounded, outliers cannot dominate the estimation.

With maximum likelihood estimation we went from setting the score to zero to setting the expected score to zero to guard against drastic overfitting. If we adopt the same approach to M-estimation we get

$$\begin{aligned} 0 &= \mathcal{E}(\ \psi(Y - \theta(X))\ ) \\ &= \mathcal{E}(\ \mathcal{E}(\ \psi(Y - \theta(X)|X\ )\ ). \end{aligned}$$

For specific $\psi(\cdot)$ there may be a natural and easy sufficient condition for the above on which to base an iterative. For many of the commonly used ψ functions the updating formula $\theta^{(i+1)}(\cdot) \leftarrow \theta^{(i)}(\cdot) - \psi^{-1}(Res(\cdot))$ where $Res(\cdot)$ is the smooth residual $\mathcal{E}(\ \psi(Y - \theta^{(i)}(X))|\ X = \cdot\ )$ is applicable. When $\psi(\cdot)$ can be continuously differentiated, the delinearization technique introduced with the exponential family maximum likelihood smoother can be used. For example, Tukey's biweight (Huber, Chapter 3) is continuously differentiable.

M-estimate smoothers are extended to location-scale problems in the same way that maximum likelihood smoothers are.

4.3 Nonparametric Smoothers

Many of the common nonparametric statistical techniques are obtained by substituting ranks for observations in the formula for the commonly used parmetric technique. For example, the two-sample Wilcoxon test for a location shift is equivalent to the t-test applied to the ranks in the pooled sample of the observations in each sample.

The analogy for smoothers is to smooth the ranks of Y on X or its ranks. The resulting smooths can be translated back to their original metric by inverting the sample distribution functions (or a smooth estimate thereof).

4.4 Bayesian Smoothers

In the smooth location problem, we can put a prior distribution π on $\mu(\cdot)$ and maximize the expected log posterior. This would continue the analogy with the way we generelized maximum likelihood to smoothers. The prior will typically be a stochastic process, and we will want it to give rise to smooth $\mu(\cdot)$ with high probability. Possible priors are Brownian motion, a nonstationary ARIMA process (trending processes look more smooth), or polynomials in X with random variables for coefficients. Suppose that the prior distribution depends on a parameter θ. If θ is known, or well estimated, we solve for $\mu(\cdot)$ in

$$0 = \mathcal{E}\left(\frac{\frac{\partial}{\partial \mu(X)} \pi(\mu(X); \theta)}{\pi(\mu(X); \theta)} - \frac{g_0'(Y - \mu(X))}{g_0(Y - \mu(X))} \right)$$

and if θ is to be jointly estimated, we solve the above jointly with

$$0 = \mathcal{E}\left(\frac{\frac{\partial}{\partial\theta}\pi(\mu(X);\theta)}{\pi(\mu(X);\theta)} \right).$$

Bayesian smoothers have been used in LANDSAT image processing. There the parameter to estimate is a 'smooth map' of categories in the plane. When there are only a finite number of categories, the specification of the prior is simplified, although it is not trivial. For examples of such Bayesian smoothers see Switzer, Kowalik, and Lyon (1981), or Owen (1983b).

Decision Theoretic Smoothers

We can, of course, formulate smoothing as a decision problem, where the loss is some function $L(\mu(\cdot), \hat{\mu}(\cdot))$. The risk incurred when the true parameter is μ is $R(\hat{\mu}) = \mathcal{E}_\mu(L(\mu, \hat{\mu}))$ and notions of minimax smoothers and minimum Bayes risk smoothers are in sight.

5 Other Models

5.0 Other Problems

Up to now, we have concentrated on estimating location models. Many other statistical models can make use of smooth curves, and we illustrate some of them in this section.

5.1 Density Estimation

For most smoothers, there is a natural way to interpolate between the sample points. With such a smoother, we can smooth the order statistics $Y_{(i)}$ against $\frac{i-1/2}{n}$ to estimate the distribution function of Y, interpolate the smooth, and differentiate the result to estimate the density.

The estimation of multivariate densities by projection pursuit methods is treated in Friedman, Stuetzle and Schroeder (1981).

5.2 Sequential Methods with Smooth Curves

5.2.1 S.P.R.T.s

Suppose we have a model in which the distribution of Y depends on X through a smooth function $\theta(\cdot)$, and that the space of values for that function is partitioned into two sets ω and ω^c. To fix ideas, suppose that we know that $\theta(\cdot)$ is monotone and that ω contains the increasing functions.

The application of the sequential probability ratio test is straightforward. Having observed the data, we obtain the constrained maximum likelihood estimates of $\mu(\cdot)$ under the two competing hypotheses, and form the ratio of the maximized likelihood under ω^c to that under ω. If this ratio exceeds a prespecified value $B > 1$, we decide ω^c, if it is less than $A < 1$ we decide ω, and otherwise we take another observation. The values A and B are chosen to control the error probabilities of types I and II.

5.2.2 Sequential Estimation of a Curve

One could keep taking data until the curve $\theta(\cdot)$ is estimated with sufficient precision. One way to guage this precision is by the width of the family of confidence envelopes for $\theta(\cdot)$.

5.3 Stochastic Process Models

5.3.1 Estimating the Spectral Density

There is a large literature on the problem of estimating the spectral density of a stationary stochastic process. It is well known that the periodogram is a poor estimate in that it severely overfits. One approach to estimating the spectral density is to smooth the periodogram (see Anderson (1971,Ch. 9)). The most commonly used smoothers are kernel smoothers. Wegman and Wright (1983) discuss the use of smoothing splines to estimate the spectral density, and Palmer (1983) uses several methods including the Supersmoother.

5.3.2 Slowly Varying Poisson Intensity

Consider a point process in time with time-dependent intensity $\lambda(t)$. If we assume that $\lambda(\cdot)$ changes slowly with respect to the rate at which points are generated, then the waiting times are well approximated by exponential distributions.

The likelihood equations for $\lambda(\cdot)$ are

$$\begin{aligned} 0 &= \mathcal{E}\left(\frac{1}{\lambda(t_i)} - (t_i - t_{i-1}) \right) \\ &= \mathcal{E}\left(\frac{-\lambda(t_i)(t_i - t_{i-1}) + 1}{\lambda(t_i)} \right) \end{aligned}$$

which we can solve iteratively via:

$$\lambda^{(i+1)}(\cdot) \leftarrow \mathcal{E}\left(\frac{1}{t_i - t_{i-1}} \mid t_i ; \frac{t_i - t_{i-1}}{\lambda^{(i)}(t_i)} \right).$$

Notice that no matter how long we observe the process, we will only get a few observations near any value of t. Hence the smoother will not be consistent for anything. Parameters are being added at essentially the same rate as the data.

The situation improves when $\lambda(\cdot)$ is actually a function of t mod T for some period $T > 0$. Then the argument of λ is a phase angle, and the number of observations used to estimate each part of it tends to infinity as more data are observed. This setup could, for example, be a model for the rate at which calls arrive at a switch, with $T = 1$ day. We can add weekly, monthly and yearly cycles (and even a dummy variable for Christmas), in an obvious way.

This model is an easy adaptation of the one in McDonald (1983) for finding a periodic decomposition of a real valued time series. McDonald also has a way of estimating the period lengths jointly with the cyclical functions.

5.3.3 Evolving Markov Transition Probabilities

The point process model can be adapted in a straightforward way to Markov processes in which transitions out of a state occur at a rate which depends on t

(and also on the state), and the state to which a transition is made is selected at random according to the probabilities in a transition matrix whose elements are smooth functions of t.

6 Conclusions

This paper shows the variety of models and methods that can be used in conjunction with smooth curves. Experience with algorithms of the type considered here has shown that they typically have good convergence behaviour, although formal proofs are difficult. For example, Breiman and Friedman (1982) show that the ACE algorithm expressed in terms of the true conditional expectations converges in general to a unique (up to sign) set of functions. They also show that the algorithm when implemented with 'sample conditional expectations' (i.e. smooths) is weakly mean square consistent for the minimizing curves, when a mean square consistent smoother is used.

Current work is focussed on developing criteria by which to assess different methods for estimating a curve, finding sufficient regularity conditions for their use, and writing the necessary software. There is also work to do in checking the asymptotic results referred to.

7 Acknowledgements

I would like to thank the Stanford Linear Accelerator Center and the Natural Sciences and Engineering Research Council of Canada for making this work possible.

References

Anderson, T.W. (1971) The Statistical Analysis of Time Series, Wiley: New York.

Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations. J.R.S.S. B 26, 211-252

Breiman, L. and Friedman J.H. (1982). Estimating Optimal Transformations for Multiple Regression and Correlation. Dept. of Statistics, Stanford University, Tech. Report ORION 010.

Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling Techniques. S.I.A.M. Philadelphia, PA.

Friedman, J.H. and Owen, A.B. (1984) Predictive ACE: Estimating Optimal Transformations for Prediction. (in preparation).

Friedman, J.H. and Tibshirani, R. (1983) The Monotone Smoothing of Scatterplots. Dept. of Statistics, Stanford University, Tech. Report ORION 013.

Friedman, J.H. and Stuetzle, W. (1981) Projection Pursuit Regression. JASA. 76, 817-823.

Friedman, J.H. and Stuetzle, W. (1982). Smoothing of Scatterplots. Dept. of Statistics, Stanford University, Tech. Report ORION 003.

Friedman, J.H., Stuetzle, W. and Schroeder (1981). Projection Pursuit Density Estimation. Dept. of Statistics, Stanford University, Tech. Report ORION 002.

Hastie, T.J. (1983a) Personal Communication.

Hastie, T.J. (1983b) Principle Curves. Dept. of Statistics, Stanford University, Tech. Report ORION 024.

Huber, P.J. (1981) Robust Statistics. John Wiley and Sons Inc. New York.

Kruskal, J.B. (1965) Analysis of Factorial Experiments by Estimating Monotone Transformations of the Data. JRSS B, 27, 251-263.

McDonald, J.W. (1983) Periodic Smoothing of Time Series. Dept. of Statis-

tics, Stanford University, Tech. Report ORION 017.

McDonald, J.W., and Owen, A.B. (1984) Smoothing with Running Split Linear Fits, (Submitted to Technometrics)

Owen, A.B. (1983a) Optimal Transformations for Autoregressive Time Series Models. Department of Statistics, Stanford University, Tech. Report ORION 020.

Owen, A.B. (1983b) A Neighborhood-based Classifier for LANDSAT Data. (To appear in the Canadian Journal of Statistics.)

Palmer, E.T. (1983) Adaptive Spectral Smoothing. Statistics Center, M.I.T. Tech. Report No. ONR 30

Reid, N.M. (1981) Influence Functions for Censored Data., Annals of Statistics 9, 78-92.

Switzer, P., Kowalik, W.S. and Lyon, R.J.P. (1981) A Prior Probability Method for Smoothing Discriminant Analysis Classification Maps. Dept. of Statistics, Stanford University, Tech. Report 1.

Tibshirani, R. (1982) Censored Data Regression with Projection Pursuit. Dept. of Statistics, Stanford University, Tech. Report ORION 013.

Tukey, J.W. (1977) Exploratory Data Analysis. Addison-Wesley Publishing Company, Reading, Massachusetts.

Wegman, and Wright (1983) Splines in Statistics. JASA 78, pp 351-365.