The Annals of Statistics
1980, Vol. 8, No. 4, 695–715

SOME THEORY OF NONLINEAR SMOOTHERS

BY C. L. MALLOWS
Bell Laboratories

In recent years J. W. Tukey has introduced several algorithms for nonlinear smoothing of time series, but theoretical development has been slow owing to the extreme difficulty of obtaining analytical results even in the simplest cases. We present here some desiderata for smoothers generally, and give a framework within which some detailed comparisons between different smoothers can be made.

1. Introduction. A *time-series* $\mathbf{X}$ is a doubly-infinite sequence of real data $\{\cdots X_{-1}, X_0, X_1, \cdots\}$. A *smoother* S is an algorithm that operates on $\mathbf{X}$ to produce a new series $S(\mathbf{X})$. We write $S(\mathbf{X})_t$ for the resulting smoothed value at time t. We use the general term "smoother" to avoid confusion with "filter", which we reserve for a linear algorithm. We do not distinguish between smoothing and forecasting, and do not intend to imply that the image series is necessarily less rapidly varying than the original, though this will usually be a prime objective.

Some examples of smoothers of the type we shall be considering are:

(i) digital filters, as described in [4] and in Chapter 46 of [11], for example, moving averages such as

$$S(X)_t = \tfrac{1}{35}(-3X_{t-2} + 12X_{t-1} + 17X_t + 12X_{t+1} - 3X_{t+2}) \tag{1.1}$$

(which will reproduce a cubic polynomial exactly), and

(ii) recursive filters such as

$$S(\mathbf{X})_t = \alpha S(\mathbf{X})_{t-1} + (1 - \alpha)X_t, \tag{1.2}$$

("exponential smoothing");

(iii) the nonlinear smoothers introduced by Tukey [15], for example "3R" in which we define $S_0(\mathbf{X}) = \mathbf{X}$,

$$S_{k+1}(\mathbf{X})_t = \text{med}(S_k(\mathbf{X})_{t-1}, S_k(\mathbf{X})_t, S_k(\mathbf{X})_{t+1}) \qquad k = 0, 1,, \ldots \tag{1.3}$$

(iterated smoothing by 3-medians), and finally

$$S(\mathbf{X})_t = \lim_{k\to\infty} S_k(\mathbf{X})_t; \text{ and} \tag{1.4}$$

(iv) "53H twice" which is defined as follows: let

$$S_1(\mathbf{X})_t = \text{med}(X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}) \qquad (\text{"5"}) \tag{1.5}$$

$$S_2(\mathbf{X})_t = \text{med}(S_1(\mathbf{X})_{t-1}, S_1(\mathbf{X})_t, S_1(\mathbf{X})_{t+1}) \qquad (\text{"3"}) \tag{1.6}$$

$$S_3(\mathbf{X})_t = \tfrac{1}{4}S_2(\mathbf{X})_{t-1} + \tfrac{1}{2}S_2(\mathbf{X})_t + \tfrac{1}{4}S_2(\mathbf{X})_{t+1} \qquad (\text{"}H\text{"}) \tag{1.7}$$

Received November 1977; revised February 1979.
AMS 1970 *subject classifications.* Primary 62M10; secondary 62M15, 62G35, 60G35, 93E10.
Key words and phrases. Filters, time series, transfer function, robustness, resistance.

and finally ("twicing")

$$S(\mathbf{X})_t = S_3(\mathbf{X})_t + S_3(\mathbf{X} - S_3(\mathbf{X}))_t. \tag{1.8}$$

Before such a smoother can be applied to a finite stretch of data, rules must be given for handling the ends of the series. Here we shall ignore this complication, considering only the result of smoothing indefinitely long stretches of data, with a stationary probability specification.

Figure 1 shows the effect of these four smoothers (in (ii) α was taken to be 0.8) on a series used in [15], namely U.S. bituminous coal production (millions of net tons per year) for 1920–1968. The data appears as asterisks, the smooths as L, R, 3, 5 respectively. No special end rules were used, except that R was started with $S_{1920} = X_{1920}$, so that each smoother gives results over a different range of years. Several attributes of the smoothers can be noticed. For example R (with this value of α) is sluggish, slow to respond to changes; the other three are less so. L is strongly influenced by outlying observations, R less so; 3 and 5 give little weight to single outliers, but do respond to bursts of two or more. Consider for example 1932–3, where a pair of low values pulls each of L, 3, 5 down considerably; and 1956–7, where a pair of high values pulls up 3 and L, but not 5.

Nonlinear smoothers differ from the classical linear filters in two important ways; they are insensitive to the presence of occasional outliers in the data (they are "resistant"), and they are much less tractable analytically. The present work is the result of an attempt to find ways of describing the properties of such smoothers, so that objective comparisons and informed choices can be made.

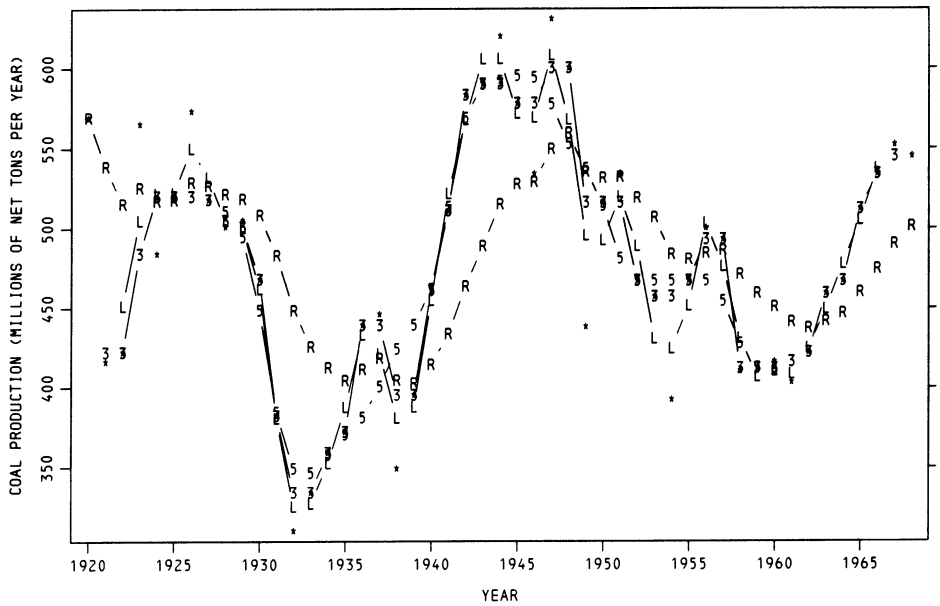

FIG. 1. Bituminous coal production in the U.S., 1920–1968.

In this search, we have drawn on the theory of robust estimation, as developed by Huber, Hampel and others (see, e.g., [1], [6], [9]), and have considered the relevance of the extensive theory of nonlinear systems (see, e.g., [10]). However, in neither field have we encountered concepts that bear directly on our concern. Much of the robustness literature deals with the estimation of one or at most a few parameters (an exception is [12]); also many of the theoretical developments are concerned with asymptotic behavior as the sample size increases indefinitely. In contrast, we are here concerned with algorithms that produce as many numbers in output as they use in input; also the arguments of $S(\mathbf{X})_t$ are not treated symmetrically, and it is unclear how an asymptotic theory can be relevant.

The main concern of nonlinear systems theory is the derivation of optimal or near-optimal algorithms for the detection or estimation of signals in noise, when the probability specification of the system is given (possibly to within some parameters to be estimated). In contrast, our emphasis is on describing the properties of a given smoothing algorithm, preferably in ways that will be as independent as possible of the specification of the system to which it is applied. Our reason for believing that such descriptions may be possible in principle, at least approximately and for certain smoothers, is the observation that a *linear* smoother (such as (i) or (ii) above) can be described by its coefficient sequence or "impulse response function" (i.e., $\frac{1}{35}(-3, 12, 17, 12, -3)$ for (1.1), $(1-\alpha)(1, \alpha, \alpha^2, \alpha^3 \cdots)$) for (1.2) or, equivalently, by its transfer function (see below). Further, *the transfer function does not depend on the probability specification.*

It is thus not unreasonable to hope that for some nonlinear smoothers, namely those that are "almost linear" in some sense, we may be able to define "coefficients" and "transfer functions" that are insensitive to changes in the specification of the process being smoothed.

Given a smoothing algorithm, we thus have the problem of defining quantities analogous to the coefficients and the transfer function, and of measuring the degree to which the smoother departs from linearity. We shall also be concerned with measuring its degree of resistance to outliers. Once these indices have been defined, we can proceed to search among the possible smoothers for ones with attractive combinations of properties.

Some basic results are expressed in Theorems 4.2–4.6 below. We find that when a nonlinear smoother is applied to a process of a certain type (stationary Gaussian plus independent noise), the resulting smoothed process can be decomposed into two parts, one part being a linear filtering of the Gaussian component, the other part being completely orthogonal to this and being nonzero by virtue of the presence of the added noise and the nonlinearity of the smoother. There is a corresponding decomposition of the spectrum of the output into a "linear" part and a residual "nonlinear" part. Subsequent linear operations act separately on the two components.

We can thus formulate the problem of designing a resistant nonlinear smoother in the following way. We must arrange that the transfer function of the linear

component has a suitable shape, while the nonlinear component has small spectrum (or simply small variance), over a suitably wide range of probability specifications, while at the same time maintaining suitable resistance properties.

We are not completely happy with our present formulation of the resistance requirement, which is discussed in Section 6.

Also, we cannot claim that the present formulation leads to excessively simple computations. We leave to another occasion a discussion of methods for Monte Carlo computation of the indices we propose; here we present (in Sections 5 and 6) merely some results relating to smoothers of a special type.

2. The smoothing problem. A natural starting-point for a theory of resistant smoothing of time series is the theory of robust estimation [1], [6], [9]. However, it is not clear that smoothing can usefully be formulated as an estimation problem. One approach would be to write $X_t = f_t(\theta) + \sigma Z_t$, $t = 1, \cdots, T$ where the nonrandom sequence $f = \{f_t(\theta)\}$ depends on an unknown (multidimensional) parameter θ, and $\mathbf{Z} = \{Z_t\}$ is an unobservable random sequence. The smoother S provides an estimate $S(\mathbf{X})_t$ of $f_t(\theta)$; in general, its mean square error will depend in detail on the choice of f, the values of t, θ, and σ, and the specification of Z. Some results of this kind have been obtained by Velleman [16], who took $f_t(\theta) = \sin \theta t$ and assumed $\{Z_t\}$ to be either independent Gaussian, independent Tukey-h ($Z_t = u_t \exp \frac{1}{2} h u_t^2$ with u_t Gaussian), or independent Gaussian with intermittent outliers. An extensive and ingenious Monte Carlo study was performed, in which some 17 smoothers were compared using 20 values of θ and three values of σ. However, it is not clear how results of this kind can be used to compare the performance of smoothers for other f's, or for the case where the noise sequence $\mathbf{Z}$ is correlated.

We take a different route, similar to that used in signal detection theory. First we give a heuristic development, starting from a simple identity. Suppose $Y_1, \cdots, Y_n, Z_1, \cdots, Z_n$ are independent random variables, $Y_1, \cdots, Y_n$ being Gaussian with mean 0, variance σ^2, and $Z_1, \cdots, Z_n$ having common distribution H (for the present, assume σ^2 and H known; we do *not* assume that H is centered). Suppose we observe $X_i = \mu + Y_i + Z_i$, $i = 1, \cdots, n$, and wish to estimate μ, using an estimate $S = S(X_1, \cdots, X_n)$ that satisfies

$$S(X_1 + c, \cdots, X_n + c) = S(X_1, \cdots, X_n) + c. \tag{2.1}$$

A natural measure of the performance of S is the mean square error $E(S - \mu)^2$. We have the following identity.

THEOREM 2.1. *Suppose* $X_i = \mu + Y_i + Z_i$ *with* $Y_i \sim N(0, \sigma^2)$, $Z_i \sim H$ *arbitrary* ($i = 1, \cdots, n$) *and with* $Y_1, \cdots, Y_n, Z_1, \cdots, Z_n$ *mutually independent. If* S *satisfies* (2.1) *and* $E(S - \mu)^2$ *is finite, then*

$$E(S - \mu)^2 = \frac{1}{n}\sigma^2 + E\big(S - \mu - \overline{Y}\big)^2$$

where $\overline{Y} = (1/n)\Sigma Y_i$.

PROOF. Writing $S - \mu = S - \mu - \bar{Y} + \bar{Y}$ and expanding the square, we have only to prove $E(\bar{Y}(S - \mu - \bar{Y})) = 0$. Using (2.1), this is $E(\bar{Y}S')$ where $S' = S(X_1 - \mu - \bar{Y}, \cdots, X_n - \mu - \bar{Y})$, which is a function only of $Y_1 - \bar{Y}, \cdots, Y_n - \bar{Y}$, $Z_1, \cdots, Z_n$. Using a standard property of the Gaussian distribution, $\bar{Y}$ and S' are independent, so $E(\bar{Y}S') = 0$.

This lemma shows that for the above specification, the problem of finding a good estimate of μ is equivalent to that of finding a good estimate of $\mu + \bar{Y}$.

This suggests the following formulation of the smoothing problem.

Basic specification.

$$X_t = \mu + Y_t + Z_t \tag{2.2}$$

where $\mathbf{Y} = \{Y_t\}$ is a zero-mean stationary Gaussian process (not completely deterministic) having covariance function $C = \{C_k\}(C_k = E(Y_t Y_{t+k}))$, and where $\mathbf{Z} = \{Z_t\}$ is a sequence of independent random variables having common distribution H. We make no assumptions on H.

Consider an arbitrary linear filter A, $A(\mathbf{Y})_t = \Sigma_j a_j Y_{t-j}$ and regard $S(\mathbf{X})_t$ as an estimate of $\mu + A(\mathbf{Y})_t$. We use as a measure of closeness the mean square error (assumed finite)

$$V(S, A) = E\big(S(\mathbf{X})_t - \mu - \Sigma_j a_j Y_{t-j}\big)^2. \tag{2.3}$$

Notice that we are not assuming that $\mathbf{Z}$ has finite moments. This specification reduces to that of the theorem when $\mathbf{Y}$ is an independent sequence and $a_j = 1/n, j = 1, \cdots, n$. We now introduce the *linear* filter S^L that minimizes $V(S, S^L)$. In general, S^L depends on μ, C, H (and S, of course). It turns out however, as we shall shortly show, that S^L has several very simple properties, which imply, amongst other things, that

$$V(S, A) = E\big(S^L(\mathbf{Y})_t - A(\mathbf{Y})_t\big)^2 + V(S, S^L). \tag{2.4}$$

Thus the adequacy of $S(\mathbf{X})_t$ as an estimator of $A(\mathbf{Y})_t$ (for given C) is known once S^L and $V(S, S^L)$ are known.

It may be worth reiterating that in our formulation we choose A to minimize $V(S, A)$ with S fixed, rather than the other way round. $S^L(\mathbf{Y})_t$ is the projection of $S(\mathbf{X})_t$ onto the manifold of linear functions of $\mathbf{Y}$, with respect to the usual L_2 product.

We shall call $S^L(\mathbf{Y})$ the "linear part" of $S(\mathbf{X})$, and define a residual series $\mathbf{R}$ by

$$S(\mathbf{X})_t = \mu + S^L(\mathbf{Y})_t + \mu_S + R_t \tag{2.5}$$

where $\mu_S = E(S(\mathbf{X})_t) - \mu$. In the next two sections we explore some properties of this representation in detail.

Notice that the basic specification (2.2) is flexible enough to cover a wide range of behavior. Models of special interest arise when C_1/C_0 is close to 1, since it is in these cases that realizations of $\mathbf{Y}$ appear smooth to the eye. If $C_0 = 1$ and C_1 is fixed, C_2 must lie in the interval $(2C_1^2 - 1, 1)$. At the lower limit, realizations of $\mathbf{Y}$

are exact sinusoids (with frequency $\cos^{-1} C_1$, amplitude Rayleigh with variance 2, phase uniform). The upper limit is less interesting; realizations of $\mathbf{Y}$ have $Y_{2t} = Y_0$, $Y_{2t+1} = Y_1$ with Y_0, Y_1 bivariate Gaussian (with correlation C_1). A more interesting case is the Markovian model, in which $C_k = C_1^{|k|}$. For $t = 0, 1, \cdots, T$, put $Y_t^* = (Y_t - Y_0)/(1 - C_1^2)^{\frac{1}{2}}$. Then as $C_1 \to 1$, the joint distribution of $Y_1^*, \cdots, Y_T^*$ conditional on $Y_0 = y_0$ approaches that of a standard Gaussian random walk, with $\{Y_{t+1}^* - Y_t^*\}$ approximately mutually independent.

3. Some assumptions. In the following, we apply a smoother S to a series $\mathbf{X}$ having the specification (2.2). Our aim is to establish the validity of the decomposition $S(\mathbf{X})_t = \mu + \mu_S + S^L(\mathbf{Y})_t + R_t$ obtained above, and to explore its properties. Before we can provide a rigorous development, we must make several assumptions regarding S, namely

A1. S is stationary: $S(T\mathbf{X}) = TS(\mathbf{X})$ where T is the shift operator ($(T\mathbf{X})_t = (\mathbf{X})_{t+1}$, $(T^{-1}\mathbf{X})_t = (\mathbf{X})_{t-1}$).

A2. S is location invariant: if $\mathbf{B}$ is a constant series ($(\mathbf{B})_t = b$), then $S(\mathbf{X} + \mathbf{B}) = S(\mathbf{X}) + \mathbf{B}$.

Notice that **A2** eliminates many of the linear filters that arise in signal-detection theory. Under the basic specification (2.2), if μ is known and Z is Gaussian with mean zero and variance V, the optimal (minimum mean square error) estimate $\hat{Y}_t$ of Y_t is linear, but fails to satisfy **A2**. For example, if $C_k = 0$ for $k \neq 0$, then $\mu + \hat{Y}_t$ is $\mu + (C_0/C_0 + V)(X_t - \mu)$. If μ is unknown, one obtains a minimum-variance invariant estimate of $\mu + Y_t$ by using the best invariant estimate of μ, which, apart from end-effects, is simply the mean of all observed X's. This filter satisfies **A2**, but is no longer stationary. Another approach is to assume that μ has a prior distribution; if this is Gaussian the optimal estimate of $\mu + Y_t$ is again linear in $\mathbf{X}$ but again fails to satisfy **A2**.

However, for the applications we have in mind, in fields such as econometrics, demography, and industrial production, Assumption **A2** seems appropriate. These are series that at least locally can reasonably be represented by a stationary model, but that are neither long enough nor stable enough for it to be reasonable to ignore the problem of estimating the local mean level.

Assumptions **A1**, **A2** imply that $S(\mathbf{B})_t = s_0 + b$ for some constant s_0; for convenience we assume

A3. S is centered; $S(\mathbf{0}) = \mathbf{0}$ where $\mathbf{0}$ is the constant zero series.

Notice that **A3** does not imply $\mu_S = 0$.

Some further technical assumptions will be needed; so far we have implicitly assumed that S is well defined for all $\mathbf{X}$, but insisting on this would lead to difficulties. One approach would be to require that S is well defined as a limit (in some sense) relative to the specification (2.2), but this is unattractive since it may turn out that details of the probability specification are important (we are after properties of S that are as independent as possible of the specification), and

furthermore we would be faced with having to justify numerous limiting arguments. We adopt the simple strategy of assuming

A4. $S(\mathbf{X})_t$ depends on only finitely many components of $\mathbf{X}$.

Assumption **A4** rules out recursive filters like (1.2) and smoothers defined as limits such as (1.4), but these definitions are only approximate (in practice every smoother requires a start-up rule) and little relevance is lost by excluding them. It is convenient to retain the fiction of a strictly stationary probability specification, however. We define the *span* $\operatorname{sp}(S)$ of S as being the smallest interval of indices such that $S(\mathbf{X})_t$ does depend only on $\{X_j: j - t \in \operatorname{sp}(S)\}$.

A second detail concerns the assumption

A5. Var $S(\mathbf{X})_t$ is finite, which we have used above in setting the stage for our decomposition (2.5). However several of the results to follow do not require this assumption; if $E|S(X)_t| < \infty$ we can define S^L to be the linear filter (satisfying **A1–A3**) that minimizes

$$E|S(X)_t - \mu - S^L(Y)_t|,$$

but if this fails we are in difficulty. We do not regard **A5** as being unduly restrictive, since it is satisfied by all moderately robust smoothers. Notice that we are not assuming that $\mathbf{Z}$ has finite moments. This is the only assumption we make that involves both the smoother S and the probability specification.

Another property enjoyed by all the smoothers considered in this paper is that for all real α, $S(\alpha\mathbf{X}) = \alpha S(\mathbf{X})$. This property is certainly desirable, from both the practical and theoretical points of view; without it the linear component S^L would depend on the scale of the process to which S is applied. However, it turns out that in our general development we do not need to make use of this assumption.

We give a formal definition of a linear smoother.

AL. A smoother S (satisfying **A4**) is *linear* if

$$S(\alpha\mathbf{X} + \beta\mathbf{X}') = \alpha S(\mathbf{X}) + \beta S(\mathbf{X}') \tag{3.1}$$

for all $\mathbf{X}$, $\mathbf{X}'$ and real α, β.

4. Properties of the decomposition (2.5). We start by establishing a basic result.

THEOREM 4.1. *If S satisfies* **A1**, **A4** *and* **AL**, *then it also satisfies* **A3**, *and*

$$S(\mathbf{X})_t = \sum_{j=-\infty}^{\infty} s_j X_{t-j} \tag{4.1}$$

for some sequence of coefficients $\{s_j\}$ (only finitely many of which are nonzero). Then also S satisfies **A2**, *if and only if* $\sum s_j = 1$.

PROOF. By **A1**, $S(\mathbf{0})_t$ is independent of t; writing $\mathbf{0} = \mathbf{0} + \mathbf{0}$ and using **AL** we find $S(\mathbf{0}) = \mathbf{0}$. Let $\boldsymbol{\Delta} = \{\Delta_t\}$ be the series with $\Delta_0 = 1$, $\Delta_t = 0$, $t \neq 0$. Define $s_t = S(\boldsymbol{\Delta})_t$. Suppose (by **A4**) $S(\mathbf{X})_t$ depends only on $\{x_j: t - a \leqslant j \leqslant t + b\}$, i.e., $\operatorname{sp}(S) = (-a, b)$. (Then s_t vanishes for $t < -b$, $t > a$). We define $\mathbf{X}^0$ by

$$\mathbf{X} = \mathbf{X}^0 + \sum_{j=t-a}^{t+b} X_j T^{-j}\boldsymbol{\Delta}.$$

Thus $(\mathbf{X}^0)_j = 0$ for $t - a \leqslant j \leqslant t + b$. Using **AL** repeatedly, we find

$$S(\mathbf{X})_t = S(\mathbf{X}^0)_t + \Sigma_{t-a}^{t+b} X_j S(T^{-j}\Delta)_t.$$

Now $S(\mathbf{X}^0)_t = S(\mathbf{0})_t = 0$, and by **A1**

$$S(T^{-j}\Delta)_t = T^{-j}S(\Delta)_t = S(\Delta)_{t-j} = s_{t-j}.$$

This establishes (4.1). Finally, by applying S to a constant series we find **A2** is equivalent to $\Sigma s_j = 1$.

REMARK. We could have chosen to define $s_j = S(\Delta)_{-j}$ and obtained $S^L(\mathbf{X})_t = \Sigma s_j X_{t+j}$, with s_j nonzero for $j \in \text{sp}(S)$, but this would lead to complications later, for example in (4.3) and Theorem 4.6 below. On balance, our actual choice leads to simpler results.

We now establish formally that a unique decomposition of the form (2.5) exists, prove the very attractive property that $S^L(\mathbf{Y})_t$ depends only on $\{Y_{t+j}, j \in \text{sp}(S)\}$, and obtain an explicit formula for S^L. *Throughout the rest of the paper,* $\mathbf{X}$ *is assumed to have the basic specification* (2.2).

THEOREM 4.2. *If* S *satisfies* **A1–A5**, *there is a unique* S^L *satisfying* **A4** *and* **AL** *that minimizes* $E(S(\mathbf{X})_t - \mu - S^L(\mathbf{Y})_t)^2$, *and* S^L *satisfies* **A1**, **A2**, **A3**. *Furthermore,* $\text{sp}(S^L) \subset \text{sp}(S)$, S^L *is determined by the equation*

$$E\big(Y_j(S^L(\mathbf{Y})_0 - S(\mathbf{X})_0)\big) = 0, \qquad j \in \text{sp}(S), \tag{4.2}$$

and its coefficients are given explicitly by (4.3) *below.*

PROOF. Fix t, and suppose $\text{sp}(S) = (-a, b)$. Choose any A, B with $-A < -a, B > b$. Write $(\mathbf{X}^{\text{in}})_{t+j} = (\mathbf{X})_{t+j} - a \leqslant j \leqslant b$, $= 0$ else; $(\mathbf{X}^{\text{out}})_{t+j} = (\mathbf{X})_{t+j} - (\mathbf{X}^{\text{in}})_{t+j}$, $-A \leqslant j \leqslant B$, $= 0$ else; and similarly for $\mathbf{Y}^{\text{in}}$, $\mathbf{Y}^{\text{out}}$, $\mathbf{Z}^{\text{in}}$, $\mathbf{Z}^{\text{out}}$. Then $S(\mathbf{X})_t = S(\mathbf{X}^{\text{in}})_t = \mu + S(\mathbf{Y}^{\text{in}} + \mathbf{Z}^{\text{in}})_t$ by **A2**.

Since $\mathbf{Y}$ is Gaussian, there are regression coefficients β_{jk} such that $W_{t+j} = Y_{t+j} - \Sigma_k \beta_{jk}(\mathbf{Y}^{\text{in}})_{t+k}$ is independent of $\mathbf{Y}^{\text{in}}$ for $-A \leqslant j \leqslant B$, and is zero for $-a \leqslant j \leqslant b$. We write $\mathbf{W}^{\text{out}}$ for the series having $(\mathbf{W}^{\text{out}})_{t+j} = W_{t+j}$, $-A \leqslant j \leqslant B$, $= 0$ else. Then for any linear filter S^L with $\text{sp}(S^L) \subset (-A, B)$ we have

$$\begin{aligned} S(\mathbf{X})_t - \mu - S^L(\mathbf{Y})_t &= S(\mathbf{Y}^{\text{in}} + \mathbf{Z}^{\text{in}})_t - S^L(\mathbf{Y}^{\text{in}} + \mathbf{Y}^{\text{out}})_t \\ &= S(\mathbf{Y}^{\text{in}} + \mathbf{Z}^{\text{in}})_t - S^L(\mathbf{Y}^{\text{in}} + (\mathbf{Y}^{\text{out}} - \mathbf{W}^{\text{out}}))_t - S^L(\mathbf{W}^{\text{out}})_t. \end{aligned}$$

Now $\mathbf{Y}^{\text{out}} - \mathbf{W}^{\text{out}}$ is a function only of $\mathbf{Y}^{\text{in}}$, so the term $S^L(\mathbf{W}^{\text{out}})$ is independent of the previous two terms. Hence $E(S(\mathbf{X})_t - \mu - S^L(\mathbf{Y})_t)^2$ is a minimum when $E(S^L(\mathbf{W}^{\text{out}})_t)^2 = 0$, i.e., when $S^L(W^{\text{out}})_t = 0$, i.e., when $\text{sp}(S^L) \subset \text{sp}(S)$.

Now suppose S^L has coefficients $\{s_j, -b \leqslant j \leqslant a\}$. Clearly $E(S(X)_t - \mu - \Sigma s_j Y_{t-j})^2$ is minimized when (4.2) holds, and the coefficients are independent of t by **A1**.

To show that S^L satisfies **A2**, it is convenient to put $m = a + b + 1$, and to write $\mathbf{s}$, $\mathbf{x}$, $\mathbf{y}$ for the $m \times 1$ vectors $\{s_{-j}\}$, $\{X_j\}$, $\{Y_j\}$ $(j = -a, \cdots, b)$ (notice the

reversed index in s!); $\mathbf{C}$ for the $m \times m$ matrix $E(\mathbf{y}\mathbf{y}^T)$; and $\mathbf{0}$, $\mathbf{1}$ for the vectors with $(\mathbf{0})_j = 0$, $(\mathbf{1})_j = 1$. $\mathbf{C}$ is nonsingular, since, by assumption, $\mathbf{Y}$ is not completely deterministic. Then (4.2) becomes

$$\mathbf{C}\mathbf{s} = E(\mathbf{y}S(\mathbf{x})) \tag{4.3}$$

where $S(\mathbf{x}) = S(\mathbf{X})_0$, so that

$$\mathbf{1}^T\mathbf{s} = E(\mathbf{1}^T\mathbf{C}^{-1}\mathbf{y}S(\mathbf{x})).$$

Put $\gamma = \mathbf{1}^T\mathbf{C}^{-1}\mathbf{1}, \bar{y} = \mathbf{1}^T\mathbf{C}^{-1}\mathbf{y}/\gamma$. Then (using **A2**)

$$\begin{aligned}\mathbf{1}^T\mathbf{s} &= \gamma E(\bar{y}S(\mathbf{x})) \\ &= \gamma E(\bar{y}(\bar{y} + S(\mathbf{x} - \bar{y}\mathbf{1}))).\end{aligned}$$

Now $E\bar{y}(\mathbf{y} - \bar{y}\mathbf{1}) = \mathbf{0}$, so that $\bar{y}$ is independent of $\mathbf{y} - \bar{y}\mathbf{1}$ (since $\mathbf{Y}$ is Gaussian!), and hence independent of $\mathbf{x} - \bar{y}\mathbf{1}$. Thus $\mathbf{1}^T\mathbf{s} = \gamma E(\bar{y}^2) = 1$ as desired.

The following result simplifies several subsequent calculations.

THEOREM 4.3. *The same coefficients are obtained if in* (4.3) $\mathbf{y}$ *is replaced by* y_+ *where* $y_+^T = (\mathbf{y}^T, \mathbf{y}_1^T)$, *where* $\mathbf{y}_1$ *contains arbitrarily many elements of* $\mathbf{Y}$ (*other than those in* $\mathbf{y}$), *with* $\mathbf{C}$ *and* $\mathbf{s}$ *being augmented to* $\mathbf{C}_+, \mathbf{s}_+$ *correspondingly. The extra elements of* $\mathbf{s}_+$ *are all found to be zero.*

PROOF. Write

$$\mathbf{C}_+ = E(\mathbf{y}_+\mathbf{y}_+^T) = \begin{pmatrix} \mathbf{C} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{A} \end{pmatrix}.$$

We verify that $(\mathbf{s}^T, \mathbf{0}^T)$ satisfies the augmented version of (4.3), i.e.,

$$\begin{aligned}\mathbf{C}\mathbf{s} &= E(\mathbf{y}S(\mathbf{x})) \\ \mathbf{B}\mathbf{s} &= E(\mathbf{y}_1S(x)).\end{aligned} \tag{4.4}$$

Now (4.4) merely repeats (4.3), while (4.5) requires

$$0 = E((\mathbf{y}_1 - \mathbf{B}\mathbf{C}^{-1}\mathbf{y})S(\mathbf{x})).$$

However,

$$E((\mathbf{y}_1 - \mathbf{B}\mathbf{C}^{-1}\mathbf{y})\mathbf{y}^T) = \mathbf{B} - \mathbf{B}\mathbf{C}^{-1}\mathbf{C} = 0$$

so that $\mathbf{y}_1 - \mathbf{B}\mathbf{C}^{-1}\mathbf{y}$ is independent of $\mathbf{y}$ and hence of $\mathbf{x}$. The result is established.

We now establish several properties of the "linear component" S^L and the corresponding decomposition

$$S(\mathbf{X}) = \mu + \mu_S + S^L(\mathbf{Y}) + \mathbf{R}.$$

THEOREM 4.4. *If* S *satisfies* **A1**–**A5** *and is linear, then* $S^L = S$.

PROOF. By Theorem 4.2, S^L exists. Suppose S has coefficients s^*; in the notation of the previous proof $S(\mathbf{X})_0 = \mathbf{x}^T\mathbf{s}^*$ and (4.3) becomes

$$\mathbf{C}\mathbf{s} = E(\mathbf{y}\mathbf{x}^T\mathbf{s}^*) = E(\mathbf{y}(\mathbf{y} + \mathbf{z})^T\mathbf{s}^*) = \mathbf{C}\mathbf{s}^*$$

and $\mathbf{s} = \mathbf{s}^*$ (since $\mathbf{C}$ is nonsingular).

Thus when S is linear, our decomposition states merely that $S(\mu + \mathbf{Y} + \mathbf{Z})_t = \mu + S(\mathbf{Y})_t + S(\mathbf{Z})_t$. Also $\mu_S = ES(\mathbf{Z})_t$, and the residual component is $R_t = S(\mathbf{Z})_t - \mu_S$. We return now to general (nonlinear) smoothers.

Since by assumption $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are stationary, so are $S(\mathbf{X})$ and $\mathbf{R}$; and since, by **A5**, Var $S(\mathbf{X})_t$ is finite, so is Var $(\mathbf{R})_t$. Thus we can define spectra F_Y, F_S, F_R such that (all integrals from $-\frac{1}{2}$ to $\frac{1}{2}$)

$$\begin{aligned} E(Y_t Y_{t+k}) &= \int e^{2\pi i k\omega}\, dF_Y(\omega) \\ E(S(\mathbf{X})_t - \mu - \mu_S)(S(\mathbf{X})_{t+k} - \mu - \mu_S) &= \int e^{2\pi i k\omega}\, dF_S(\omega) \\ E(R_t R_{t+k}) &= \int e^{2\pi i k\omega}\, dF_R(\omega) \end{aligned}$$

(see, for example, [2], Chapter 8). As is well known, if A is a linear filter with coefficients a_j, and $\mathbf{X}' = A(\mathbf{X})$, then $dF_{X'}(\omega) = |\hat{A}(\omega)|^2\, dF_X(\omega)$ where $\hat{A}(\omega)$ (the transfer function of A) is

$$\hat{A}(\omega) = \Sigma_{-\infty}^{\infty} a_j e^{-2\pi i j\omega}.$$

The transfer function gives a very useful description of a linear filter. We now prove an important property of our decomposition.

THEOREM 4.5. *If S satisfies* **A1–A5** *then for all t, u*

$$E(R_t Y_u) = 0 \tag{4.6}$$

and

$$dF_S(\omega) = |\hat{S}^L(\omega)|^2\, dF_Y(\omega) + dF_R(\omega). \tag{4.7}$$

PROOF. By **A1**, we may take $t = 0$. Let $\mathbf{y}_+$ be any vector whose elements are components of $\mathbf{Y}$, containing at least Y_u and all of the elements of $\mathbf{y}$. Then

$$E(\mathbf{y}_+ R_0) = E\big(\mathbf{y}_+ (S(\mathbf{X})_0 - \mu - \mu_S - \mathbf{y}_+^T \mathbf{s}_+)\big)$$

which vanishes by Theorem 4.3, so that $E(Y_u R_0) = 0$. Hence $E(S(\mathbf{X})_t - \mu - \mu_S)(S(\mathbf{X})_{t+k} - \mu - \mu_S) = E(S^L(\mathbf{Y})_t S^L(\mathbf{Y})_{t+k}) + E(R_t R_{t+k})$ and (4.6) follows.

The result (2.4) is a simple consequence of (4.6).

Notice that we have not shown $\mu_S = 0$; this can easily fail to be true, for example if $\mathbf{Z}$ is a constant (nonzero) series. Some centering assumption on $\mathbf{Z}$ would be required to ensure that μ_S vanishes.

It is most fortunate (and somewhat remarkable) that the result (4.7), which involves the serial structures of $S(\mathbf{X})$, $S^L(\mathbf{Y})$ and $\mathbf{R}$, is found to hold, when S^L was defined by the minimization of the single scalar quantity $E(S(\mathbf{X})_0 - \mu - S^L(\mathbf{Y})_0)^2$. The validity of (4.6) and (4.7) adds appreciably to the usefulness of our decomposition. A simple consequence follows.

THEOREM 4.6. *If A is a (finite) linear filter, S is a nonlinear smoother with linear component S^L, and $(AS)(\mathbf{X}) = A(S(\mathbf{X}))$, then $(AS)^L = A(S^L)$.*

PROOF. Suppose A, S^L, $(AS)^L$ have coefficients $\{a_j\}$, $\{s_j\}$, $\{p_j\}$ respectively. By Theorem 4.3, for suitably chosen $\mathbf{y}$ and $C = E(\mathbf{y}\mathbf{y}^T)$, the coefficients $\{p_j\}$ satisfy $\Sigma_m C_{t-m}\, p_{-m} = E(Y_t(AS)(\mathbf{X})_0) = E(Y_t \Sigma_j a_j S(\mathbf{X})_{-j}) = \Sigma_j a_j E(Y_{t+j} S(\mathbf{X})_0)$ (by stationarity) $= \Sigma_j a_j (Cs)_{t+j}$ (by Theorem 4.3 again) $= \Sigma_j \Sigma_k a_j C_{t+j-k} s_{-k} = \Sigma_m C_{t-m} \Sigma_j a_j s_{-m-j}$, so that $p_k = \Sigma a_j s_{k-j}$ as stated.

Some immediate consequences of this theorem are that $(AS)^L(\omega) = A(\omega)S^L(\omega)$ and $(AS)(\mathbf{X}) - (AS)^L(\mathbf{Y}) = A(\mathbf{R})$. These results add further interest to our decomposition of S into "linear" and "nonlinear" components; they show that as far as second-order properties are concerned, if the output of S is acted on only by linear operations, the two components can be thought of as being independent. This property holds out hope of greatly simplifying the task of designing a robust smoother; one uses first a nonlinear smoother to achieve the desired insensitivity to outliers, and follows it with a linear filter to achieve a desired transfer shape.

5. Some computations. In the proof of Theorem 4.2 we showed that, assuming **A1–A5** and the basic specification (2.2), the coefficients $\{s_j\}$ of the linear component S^L (defined in Theorem 4.2) of a nonlinear smoother S which is such that $S(\mathbf{X})_t$ is a function only of $\{X_{t-a}, \cdots, X_{t+b}\}$ for some a, b, can be obtained by solving the linear equations $\mathbf{Cs} = E(\mathbf{y}S(\mathbf{x}))$ where $\mathbf{s} = \{s_{-j}\}$, $\mathbf{y} = \{Y_j\}$, $\mathbf{x} = \{X_j\}(j = -a, \cdots, b)$, $\mathbf{C} = E(\mathbf{y}\mathbf{y}^T)$. The coefficients thus depend (in general) on the autocovariances of $\mathbf{Y}$ through order $a + b$ and the common distribution H of the elements of the noise component $\mathbf{Z}$. We now examine this dependence in a little more detail. Let Φ be the standard (independent) Gaussian measure on $R^{\text{sp}(S)}$.

THEOREM 5.1. *If S is differentiable* a.e. (Φ), *and for some* $\lambda > 0$, $e^{-\lambda\|x\|}S(x) \to 0$ *as* $\|x\| \to \infty$, *then*

$$s_{-j} = E\left(\frac{\partial S(x)}{\partial X_j}\right). \tag{5.1}$$

PROOF. Since $p(\mathbf{y}) = \text{const exp } -\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}$, we have

$$s_{-j} = \int \cdots \int S(\mathbf{x})\left(-\frac{\partial}{\partial y_j} p(\mathbf{y})\right)\Pi_i\, dy_i\, dH(z_i)$$

and the result follows on an integration by parts.

A simple subclass of nonlinear smoothers, for which we propose the name *selectors*, are those for which $S(\mathbf{X})_t$ is X_{t+j} for some j (depending on $\mathbf{X}$), i.e.,

$$S(\mathbf{X})_t = \Sigma_j I_j(T^t\mathbf{X})X_{t+j} \tag{5.2}$$

where each $I_j(\mathbf{X})$ takes values 0, 1 only, with $\Sigma_j I_j(\mathbf{X}) = 1$ for all $\mathbf{X}$. Examples are moving odd medians and iterates of these, such "3" and "53" (= S_1 in (1.3) above and S_2 in (1.5) above, respectively), but not "53H" (= (1.6) above) since this is not a selector. "3R" (= (1.4) above) is a selector, but strictly is not covered by our present theory since it does not have finite span.

THEOREM 5.2. *If S is a selector, the coefficients of S^L can be found from*

$$s_{-j} = P(S(\mathbf{x}) = X_j).$$

PROOF. From (5.2), $\partial S/\partial X_j = I_j(\mathbf{x})$ a.e. (Φ), and the result follows from Theorem 5.1.

This result shows that when $\mathbf{X}$ is an *independent* process with continuous marginal df, (for any H), the linear coefficients of some selectors, namely those for which $I_j(\mathbf{x})$ depends only on the relative ranks of $X_{t-a}, \cdots, X_{t+b}$, can be obtained by purely combinatorial methods, since in this case $P(S(\mathbf{x}) = X_j)$ is distribution-free. Some numerical results are collected in Table 1.

Table 1 gives the linear coefficients for six selectors of span $(-3, 3)$; the corresponding transfer functions are shown in Figure 2. Figure 3 gives estimates of the residual spectra, for the case $Z = 0$.

We remark that while the results in Table 1 have been derived assuming that $\mathbf{X}$ is an independent process, it is very plausible that the coefficients in Table 1 will apply to a good approximation whenever $X_t = M_t + Y_t + Z_t$ with $\mathbf{Y}$ independent (Gaussian) and $\mathbf{M}$ slowly-varying. (This model is adequate for many observed series.) Our formalism enables us to deal only where the case $\mathbf{M}$ is Gaussian. The following result for the smoother 3 ($= S_1$ in (1.3)) typifies what one can hope to be able to establish for this situation.

TABLE 1

Linear coefficients of some selectors when $\mathbf{X}$ *is independent.*

Selector*	Coefficients**
m	$1/m, 1/m, \cdots, 1/m$
3^2	$(2, 7, 12, 7, 2)/30 = (.\underline{400}, .233, .067)$
3^3	$(3, 10, 52, 80, 52, 10, 3)/210 = (.\underline{381}, .248, .048, .014)$
53	$(9, 30, 44, 44, 44, 30, 9)/210 = (.\underline{210}, .210, .143, .043)$
35	$(10, 24, 45, 52, 45, 24, 10)/210 = (.\underline{248}, .214, .114, .048)$
3^k, k large***	$(.\underline{383}, .244, .050, .011, .002, .001, \cdots)$
$W(5, 2)$	$(1, 1, 6, 1, 1)/10$
$W(7, 2)$	$(1, 1, 1, 15, 1, 1, 1)/21$
$W(7, 4)$	$(2, 2, 2, 9, 2, 2, 2)/21$
$7 = W(7, 6)$	$(3, 3, 3, 3, 3, 3, 3)/21$

*Notation: m denotes a median of (odd) span m. 53 denotes 5 followed by 3. 3^k denotes k iterates of 3. $W(m, 2a)$ denotes a "Winsorizing" smoother. Let $r(j)$ be the rank of X_j within $\mathbf{x}$. Then if $r(0)$ is not one of $(1, 2, \cdots, a, m + 1 - a, \cdots, m)$ we set $S(\mathbf{x}) = X_0$; if $r(0) < a$ we set $S(\mathbf{x}) = X_k$ where $r(k) = a$; if $r(0) > m + 1 - a$ we set $S(\mathbf{x}) = X_k$ where $r(k) = m + 1 - a$.

**The coefficients are given in decimal form starting with the middle one, which is underlined.

***For the selector $3R$, the coefficients and several distributional results have been obtained (in the independent case) by combinatorial methods, and will be reported elsewhere.

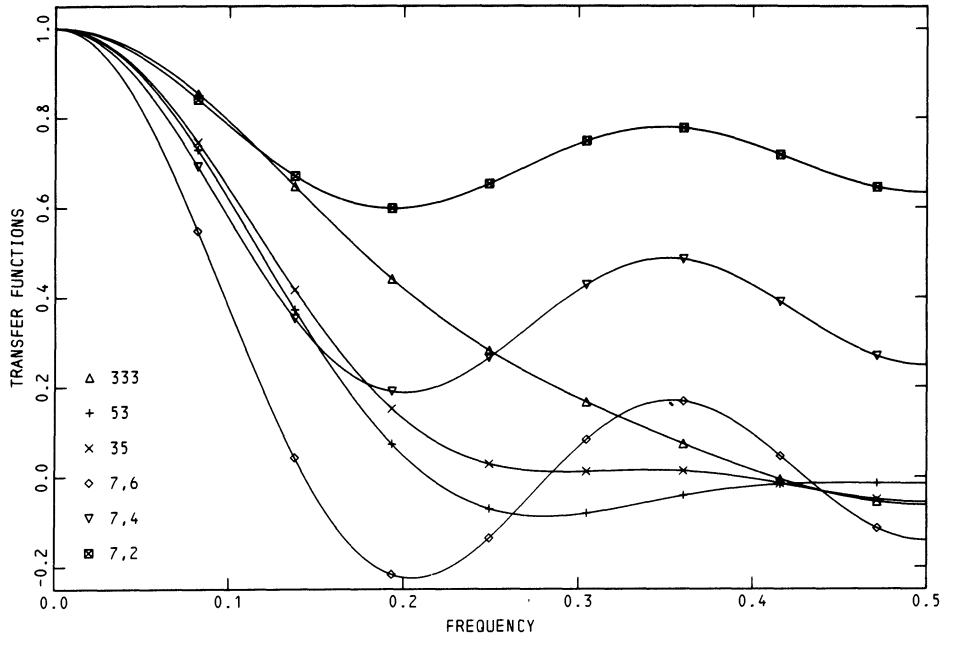

FIG. 2. Transfer functions of linear components of six smoothers.

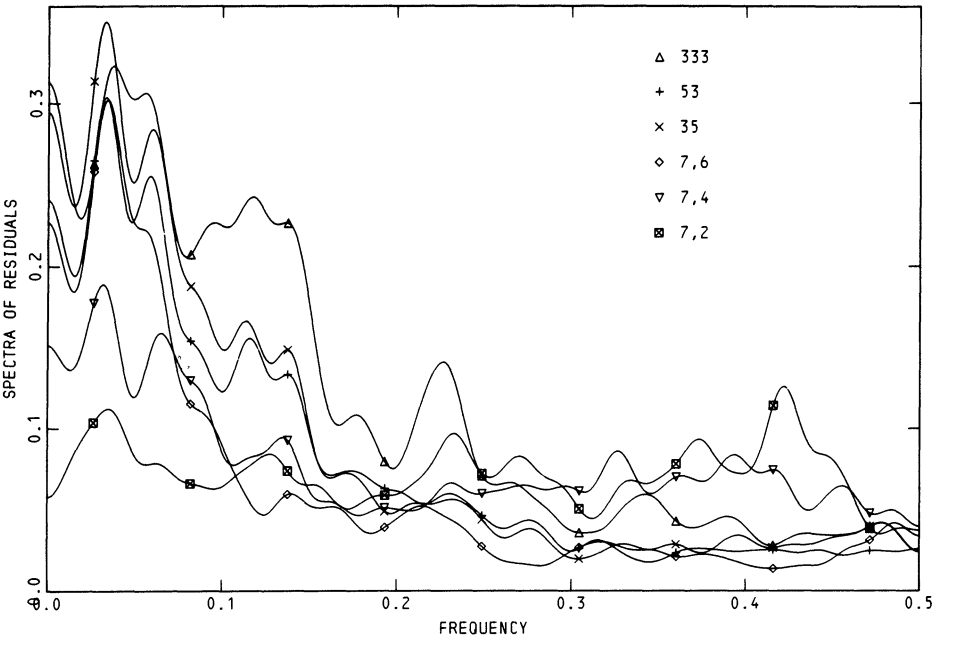

FIG. 3. Estimates of spectra of residual components of six smoothers.

THEOREM 5.3. *If, in the basic specification* (2.2), $\mathbf{Z} = \mathbf{0}$ w.p. 1 *and* $\mathbf{Y} = \mathbf{M} + \mathbf{U}$ *where* $\mathbf{M}$ *is Gaussian with covariance* C^M, $\mathbf{U}$ *is an independent white Gaussian process with* $\mathrm{Var}(U_t) = C_0^U$, *and if* $(C_0^M - C_1^M)/C_0^U$ *is small, then for the smoother "3" we have*

$$s_0 = \tfrac{1}{2} - s_1 = \tfrac{1}{2} - s_{-1} = \tfrac{1}{6} + \frac{1}{2\pi(3)^{\frac{1}{2}}} \frac{C_1^M - C_2^M}{C_0^U} + 0\left(\frac{C_0^M - C_1^M}{C_0^U}\right)^2.$$

Notice that we do not require C_0^M *to be small.*

PROOF. The proof is straightforward but a little tedious. By symmetry and Theorem 5.2, we need only consider $P(X_{-1} < X_0 < X_1)(= \frac{1}{2}s_0)$. We write this as

$$E_{M_{-1}, M_0, M_1} P(U_{-1} + M_{-1} < U_0 + M_0 < U_1 + M_1 | M_{-1}, M_0, M_1)$$
$$= \int dQ(a, b) \int_{x<y<z} \phi(x - a)\phi(y)\phi(z - b)\, dx\, dy\, dz$$

where Q is the joint measure of $(M_{-1} - M_0)/(C_0^U)^{\frac{1}{2}}$, and $(M_1 - M_0)/(C_0^U)^{\frac{1}{2}}$, and ϕ is the standard Gaussian density. We expand in powers of a and b, justifiably since the derivatives of ϕ are uniformly bounded, interchange the orders of integration, and finally evaluate several integrals of the form $\int_{x<y<z} \phi''(x)\phi(y)\phi(z)\, dx\, dy\, dz$, etc.

6. Resistance. In the previous sections we have introduced and discussed the representation $S(\mathbf{X}) = \mu + \mu_S + S^L(\mathbf{Y}) + \mathbf{R}$, which is valid (relative to our basic specification (2.2)) whenever assumptions **A1–A5** hold. In general, the coefficients of the "linear part" S^L, and the spectrum of the "residual" $\mathbf{R}$, depend on the parameters (C, H) of the probability specification. The problem of designing a nonlinear smoother can thus be formalized in the following way. First choose some linear filter A, representing the filter that would be desired if the specification were known to be exactly Gaussian. Choose also sets $\mathcal{C}$, $\mathcal{H}$ of interesting covariance and noise specifications, and metrics δ_1, δ_2 on the spaces of linear filters and residual spectra respectively. (Two simple choices are $\delta_1(A, B) = E(A(\mathbf{Y})_t - B(\mathbf{Y})_t)^2$, $\delta_2 = \mathrm{Var}(R_t) = \int f_R(w)\, dw$). Then search for a smoother S that makes $\delta_1(A, S^L)$ and $\delta_2(f_R)$ small for all (C, H) in $\mathcal{C} \times \mathcal{H}$.

An unappealing feature of this formulation is that it involves $\mathcal{C}$ and $\mathcal{H}$ explicitly. We now explore the possibility of saying something about the magnitude of the nonlinear component $\mathbf{R}$ for a nontrivial family of H's, under minimal assumptions. We search for an index with which to measure the degree to which a given smoother S is resistant to remote outliers, being guided by Hampel's observation [6] that in the location-estimation problem the "breakdown point" is a primary indicator of satisfactory behavior overall.

The breakdown point is defined in the following way [5]. Given a sequence of location-estimators $\{T_n\}$ and a probability distribution F, the breakdown point $B(T, F)$ is the largest β such that for all distributions G such that $\delta(F, G) \leqslant \beta$

(Prohorov metric), there is a compact set $K(\beta)$ such that $P_G(T_n \in K) \to 1$ as $n \to \infty$. Loosely speaking (and quoting Hampel [7]), "it is the smallest percentage of free contamination which can carry the value of the estimator over all bounds".

In attempting to extend this concept to the smoothing context, we encounter three difficulties. First, Hampel's definition concerns asymptotic behavior as the sample size increases, whereas it is not at all clear how a smoother can be (usefully) embedded in a sequence. Hampel's definition (in the location case) could be made n-specific (at the cost of additional complexity) by defining $\beta(n, \varepsilon)$ to be the largest β such that for some compact K, if $\delta(F, G) \leqslant \beta$ then $P_G(T_n \in K) > 1 - \varepsilon$.

A second difficulty is that application of a smoother to a segment $(X_0, \cdots, X_N)$ of an observed process produces not just one real estimate, but a sequence $(S_a, \cdots, S_{N-b})$ of smoothed values. Hampel's definition applies as it stands to the multidimensional case, but if the probability specification allows remote outliers to occur, one must expect that as N increases, the probability that one or more smoothed values will be an outlier will increase to unity. We can avoid this difficulty, relying on a stationarity assumption, by concentrating on $P(S_t$ is an outlier) for a single t (and perhaps on $P(S_t, S_{t+1}$ are both outliers), etc.).

A third difficulty is that Hampel's definition involves a metric on the space of probability specifications, and it is not immediately clear how to define a suitable metric on specifications of processes (however, see [14]).

The definition that follows presents our best attempt at avoiding these difficulties. We assume the basic specification (2.2), and write H in the form

$$H = (1 - p)H_0 + pH_1$$

with H_0, H_1 arbitrary, but with p fixed. Then the 'breakdown probability' is defined to be

$$B(p) = B(p; C, H_0) = \lim_{k \to \infty} \sup_{H_1} P(|S(\mathbf{X})_t| > k).$$

As the full notation indicates, in general this quantity may depend on C and H_0; however, for many smoothers of interest (including these mentioned in Section 1 above), $B(p)$ is independent of C and H_0. A search for necessary and sufficient conditions that this should happen has not been fruitful. Table 2 gives some explicit results.

The following observations can be made. First, in all these cases, $B(p)$ is independent of C and H_0, and is a polynomial in p, with leading term αp^β with α, β integral. The exponent seems to depend only on the maximally-trimming component of the smoother. Thus $\beta = 1$ for linear smoothers; $\beta = 2$ for $W(3, 2)$, $(=3)$, $W(5, 2)$, $W(7, 2)$, and for $3^k (k \geqslant 2)$; $\beta = 3$ for $W(5, 2)$, $W(7, 4)$ and all smoothers built out of 5 and components of lower exponent, etc. Second, the coefficient α depends on what components are included with the maximal-exponent one, and in what order. Concatenating components may increase or decrease α.

TABLE 2

Breakdown probabilities for some smoothers.

Smoothers* Linear, span $(-a, b)$	$B(p)$** $1-(1-p)^{a+b+1}$
m	$\sum_{j=\frac{1}{2}(m+1)}^{m} \binom{m}{j} p^j (1-p)^{m-j}$
3	$3p^2 - 2p^3$
5	$10p^3 - 15p^4 + 6p^5$
3^k, $k \geq 2$	$2p^2 + O(p^3)$
53	$7p^3 + O(p^4)$
35	$9p^3 + O(p^4)$
$53H$	$13p^3 + O(p^4)$
$53H$ twice***	$13p^3 + O(p^4)$
$5\bar{3}H$****	$34p^3 + O(p^4)$
$W(5, 2)$	$4p^2 + O(p^3)$
$W(7, 2)$	$6p^2 + O(p^3)$
$W(7, 4)$	$15p^3 + O(p^4)$
$7 = W(7, 6)$	$35p^4 + O(p^5)$

*The notation for selectors agrees with that in Table 1. "$53H$" is (1.7).

**In all these cases, $B(p)$ is independent of C and H_0.

***This result was obtained assuming that in the worst case H_1 would put all its probability on one side of the origin. This may not be correct.

****Here $\bar{3}$ denotes an arithmetic mean (a linear operation).

As an alternative to the formulation at the beginning of this section, the design problem can be posed in the following way. Given a linear filter A, interesting sets $\mathcal{C}$, $\mathcal{H}$, and a metric δ_1, find a smoother S that makes each of the criteria $\delta_1(A, S^L)$ and $B(p)$ small for all C, H in $\mathcal{C} \times \mathcal{H}$.

7. Further comments and some open questions. Clearly much detailed calculation will be necessary to arrive at an understanding of the trade-offs that are available among the three criteria $\delta_1(A, S^L)$, $\delta_2(f_R)$, $B(p)$. We have begun some numerical investigations using Monte Carlo methods to estimate these quantities, for a variety of smoothers and specifications. However, there are several open questions of a theoretical nature, on some of which some progress has been made.

(i) Throughout this paper we have assumed that the process $\mathbf{X}$ is stationary; yet it will be important to study the response of smoothers to various kinds of nonstationarity. A particularly simple and important case arises when the differenced series $\Delta X_t = X_{t+1} - X_t$ is stationary; thus when X_t is of the form $\Delta Y_t + \Delta Z_t$ with $\Delta\mathbf{Y}$ stationary Gaussian and $\mathbf{Z}$ independent noise, it appears that the previous development continues to apply, so that a decomposition of the form $S(\mathbf{X})_t = S^L(\mathbf{Y})_t + R_t$ where $\mathbf{R}$ is stationary with spectrum $f_R(\omega)$ can be defined, even though $\mathbf{X}$ and $\mathbf{Y}$ are not stationary and so do not have spectra. However, more complex kinds of nonstationarity will be harder to deal with.

(ii) Another direction in which the development needs to be extended is to consider noise specifications other than independent, with identical distributions. In practice, outliers often tend to occur in bursts, and smoothers that are otherwise quite similar may differ in their sensitivity to such outlier patterns.

(iii) Huber [8] introduced the class of M-estimators for the location problem. Given data

$$X_i = \mu + W_i \qquad i = 1, \cdots, n \tag{7.1}$$

with $W_1, \cdots, W_n$ independent with common distribution G, an estimate of μ is obtained as the minimizer of $\Sigma_i \rho(X_i - \mu)$ for some suitable function ρ; this is the maximum-likelihood estimate under the assumption $\ln G'(w) \propto \text{const.} - \rho(w)$. Analogously, an "$M$-smoother" can be defined as the minimizer of $\Sigma_j \rho_j(X_{t+j} - S_t)$ where, possibly, $\rho_j(x) = a_j \rho(x)$. Do smoothers of this type have any merits?

(iv) Huber established (for the location problem) several appealing properties of M-estimators, including an asymptotic minimax result of the following form. Writing $\psi = \rho'$, if G and ρ are symmetric about zero, the asymptotic variance of the above M-estimate is $1/nK(\psi, G)$ where $K(\psi, G) = (\int \psi' \, dG)^2 / \int \psi^2 \, dG$. Let $\mathcal{G}$ be a convex set of symmetric distributions such that at least one $G \in \mathcal{G}$ has finite Fisher information: $I(G) = \int (G''/G')^2 \, dG < \infty$. Then, if there is a $G_0 \in \mathcal{G}$, such that $I(G_0) \leqslant I(G)$ for all $G \in \mathcal{G}$, and if Ψ is a set of continuous skew symmetric functions containing $\psi_0 = -G_0''/G_0'$, then (ψ_0, G_0) is a saddlepoint of K, so that

$$K(\psi, G_0) \leqslant K(\psi_0, G_0) \leqslant K(\psi_0, G)$$

for all $\psi \in \Psi$ and all $G \in \mathcal{G}$. Thus G_0 is (asymptotically) a least favorable distribution in $\mathcal{G}$. Huber applied this result to the family

$$\mathcal{G}_\varepsilon = \{G: G = (1 - \varepsilon)\Phi + \varepsilon H\}$$

where Φ is the standard Gaussian distribution function, and H is arbitrary, (symmetric about zero), obtaining a very simple least-favorable G_0, with density const. $\exp - \rho_0(x)$ where $\rho_0(x) = \frac{1}{2}x^2$ for $|x| < k$, $= k|x| - \frac{1}{2}x^2$ for $|x| > k$, where k and ε are related by $(1 - \varepsilon)^{-1} = 2k^{-1}\phi(k) + 2\Phi(k) - 1$.

Our development in Section 2 suggests consideration of the same problem with $\mathcal{G}$ taken to be the set of distributions of random variables that can be written in the form $Y + Z$ where Y is standard Gaussian, and Z has a distribution in some symmetric (nonGaussian) family $\mathcal{H}$. A particularly appealing family is

$$\mathcal{H}_\varepsilon = \{H: H = (1 - \varepsilon)\delta_0 + \varepsilon F\}$$

where δ_0 assigns unit mass to $x = 0$, and F is arbitrary (symmetric). This specification differs from Huber's in that we consider *additive* noise, whereas he considers *replacing* noise. For us, when the true Y is not observed, which occurs with probability ε, it is because some nonzero realized Z has been added to Y; for Huber, when Y is not observed, which occurs with probability ε, it is because a Z is observed instead. This change might seem to be of little consequence, yet after much effort I have been unable to determine G_0 in this case. (B. F. Logan has been able to demonstrate that the worst-case F cannot have a continuous density).

(v) Several other attempts have been made to describe the properties of nonlinear smoothers, under a variety of specifications. For example, assuming **X** stationary, we have studied representations of the forms

$$S_t = \Sigma a_j X_{t+j} + R_t \tag{7.2}$$

$$S_t = \Sigma a_j f(X_{t+j}) + R_t \tag{7.3}$$

$$S_t = \Sigma f_j(X_{t+j}) + R_t \tag{7.4}$$

$$S_t = g(\Sigma a_j X_{t+j}) + R_t. \tag{7.5}$$

Although some of these lead to interesting and tractable calculations, they all have serious deficiencies. For example, (7.2) applies only when **X** is assumed to have finite variance, which is a very restrictive assumption. (7.5) leads to very complicated analysis.

(7.4), of which (7.3) is a special case, is more interesting, being related to a concept of central importance in the theory of robust estimation of location, namely Hampel's influence function, [6], [7]. This is defined in the following way. An estimator S (of the location parameter μ in the model (7.1)) is thought of as being a functional on the space of distribution functions, so that $S(G) = \mu$, the parameter of interest, and the estimate obtained from data $X_1, \cdots, X_n$ is $S(F_n)$ where F_n is the empiric distribution of the data:

$$F_n(x) = \frac{1}{n}\Sigma_{j=1}^{n} J_{x_j}(x).$$

where $J_y(x) = 0, 1$ for $x < y$, $x \geqslant y$. Then the influence function $S_1(G, x)$ of S at G is defined pointwise by

$$S_1(G, x) = \lim_{\varepsilon \downarrow 0}(S((1 - \varepsilon)G + \varepsilon J_x) - S(G))/\varepsilon. \tag{7.6}$$

The importance of this quantity arises from the relation, valid under certain weak conditions whenever $X_1 \ldots X_n$ are a random sample from G,

$$S(F_n) = S(G) + \Sigma_{j=1}^{n} S_1(G, X_j) + o_p(n^{-\frac{1}{2}}).$$

This expression shows that $S(F_n)$ can be approximated by a sum of independent terms.

To extend this concept to the smoothing context, we need a new definition to replace (7.6), since an asymptotic formulation is inappropriate. We appeal to a device of Hajek [3], who introduced a projection approximation $\hat{S}_n$ of a statistic $S_n = S(X_1, \cdots, X_n)$ by writing

$$\hat{S}_n = E(S_n) + \Sigma_{j=1}^{n} h_{jn}(X_j)$$

where

$$h_{jn}(x) = E(S_n | X_j = x) - E(S_n).$$

If S is a symmetric function of its arguments, as will usually be the case in the location problem, h_{jn} is independent of j. Hajek showed that when $X_1, \cdots, X_n$ are

independent with common distribution F, and $ES_n^2 < \infty$, then $E(S_n - \Sigma g_j(X_j))^2$ is minimized by taking $g_j = h_{jn}$, and

$$\text{Var } S_n = \text{Var } \hat{S}_n + E(S_n - \hat{S}_n)^2.$$

Under regularity conditions, which have not been worked out in detail, we will have

$$\lim_{n\to\infty} h_n(x) = S_1(F, x)$$

so that the function h_n is a finite-sample analogue of the influence function. (For some other analogues, see [13].)

This suggests that in the smoothing context, where S is not a symmetric function of its arguments, and these arguments are not independent, a representation of the form (7.4) might be useful, where the functions $\{f_j\}$ are dètermined to minimize the quantity $Q = E(S_t - \Sigma f_j(X_{t+j}))^2$. (Particularly elegant formulas result when $\mathbf{X}$ is assumed to be Gaussian).

However, this approach has a serious defect, which we explain by reference to a special case. Let S be the 3-median selector ($= S_1$ of (1.3)), and take $\mathbf{X}$ to be an independent Gaussian process with zero mean and unit variance. Thus we are in the location-parameter situation, and $f_{-1} = f_0 = f_1$, and we find that f_0 is an odd bounded monotonic function that approaches $\pi^{-\frac{1}{2}}$ as $x \to \infty$. Now if $\mathbf{X}$ is contaminated by long-tailed noise as in our basic specification (2.2), it can be shown that f_0 remains bounded, but that now $f_0(x) \to 0$ as $x \to \infty$. Thus f_0 responds very sensitively to changes in the specification, and the value of f_0 in the Gaussian case is not a good guide to the effect of the presence of outliers.

In the smoothing case the phenomenon just described increases in strength. Suppose now that $\mathbf{X}$ is a Gaussian Markov process with zero mean, unit variance, and parameter θ, so that $C_k = \theta^{|k|}$ for all k. Putting $q_j(x) = E(S_t | X_{t+j} = x)$, we see that Q is minimized when f_{-1}, f_0, f_1 satisfy

$$\begin{aligned} f_{-1}(x) + E(f_0(X_t)|X_{t-1} = x) + E(f_1(X_{t+1})|X_{t-1} = x) &= q_{-1}(x) \\ E(f_{-1}(X_{t-1})|X_t = x) + f_0(x) + E(f_1(X_{t+1})|X_t = x) &= q_0(x) \\ E(f_{-1}(X_{t-1})|X_{t+1} = x) + E(f_0(X_t)|X_{t+1} = x) + f_1(x) &= q_1(x). \end{aligned}$$

These equations imply $f_{-1} = f_1$. For x large, each of the left-hand expressions is approximately linear in x, being close to θx; it is thus very plausible that f_0 and f_1 are also approximately linear for large x. Assuming $f_0(x) \sim \lambda x$, $f_1(x) \sim \mu x$, we find $\lambda = \theta(1 - \theta)/(1 + \theta)$, $\mu = \theta/(1 + \theta)$. Now suppose $\mathbf{X}$ is contaminated by a small amount of additive long-tailed noise. Then for x large, the q's will increase much more slowly than before, so that the f's will also. Thus again the f's respond very sensitively to such changes in the specification. It is not clear whether anything useful can be salvaged from this approach.

(vi) An attempt has been made to study nonlinear smoothers analogous to those considered in this paper in the case of continuous-time processes. For example, if

$X(t)$ is a stationary Gaussian process one can define a "box-car median" smoother $S_a(t)$ as being the minimizer of $\int_{t-a}^{t+a}|X(u) - S_a(t)|\,du$; a "three-point median" is simply $\text{med}(X(t-a), X(t), X(t+a))$. Such smoothers have proved excessively difficult to handle, though some explicit results have been obtained for the box-car median of a Wiener process.

(vii) We remark that it is easy to define large numbers of nonlinear smoothers. Tukey [15] has proposed many, using a novel repertoire of elementary operations. By analogy with the location problem, an "L-smoother" can be defined by

$$S_t = \int x b(F_t(x))\,dF_t(x)$$

where b is some specified function and F_t is a weighted empirical distribution

$$F_t(x) = \Sigma a_j J(x - X_{t+j})$$

for some (positive) constants $\{a_j\}$ summing to 1. The Winsoring selectors of Table 1 are of this form. Another class is the "robustified discrete spline", obtained for example by minimizing $\Sigma_t \rho(X_t - S_t) + \beta \Sigma_t (S_{t+1} - S_t)^2$ for some suitable function ρ. Kleiner, Martin and Thomson [12] have had much success with robust autoregressive predictors of the form

$$S_t = \Sigma_j a_j S_{t-j} + c_t \psi\big((X_t - \Sigma a_j S_{t-j})/c_t\big)$$

when ψ is a bounded skew-symmetric function, and where $(a_1, \cdots, a_k)$ and $\{c_t\}$ are estimated from the observed series.

In the present work we have not attempted to extend this list of smoothers, but have tried to develop methods for comparing the properties of given smoothers. Only when suitable performance criteria have been developed does it become appropriate to search for optimal smoothers and to ask how close to optimal are various simple ones.

Acknowledgments. Grateful thanks for their helpful comments are due to S. P. Lloyd, B. F. Logan, Y. Vardi, P. Velleman, and numerous other patient souls.

REFERENCES

[1] BICKEL, P. J. (1976). Another look at robustness: a review of reviews and some new developments. *Scand. J. Statist.* **3** 145–168.

[2] COX, D. R. and MILLER, H. D. (1965). *The Theory of Stochastic Processes*. Wiley, New York.

[3] HAJEK, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. Statist.* **39** 325–346.

[4] HAMMING, R. W. (1977). *Digital Filters*. Prentice-Hall, Englewood Cliffs, N.J.

[5] HAMPEL, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42** 1887–1896.

[6] HAMPEL, F. R. (1973). Robust estimation: a condensed partial survey. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **27** 87–104.

[7] HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393.

[8] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101.

[9] HUBER, P. J. (1972). Robust statistics: a review. *Ann. Math. Statist.* **43** 1041–1067.

[10] JAZWINSKI, A. H. (1970). *Stochastic Processes and Filtering Theory*. Academic, N.Y.

[11] KENDALL, M. G. and STUART, A. (1966). *The Advanced Theory of Statistics, Vol.* 3. Hafner, New York.
[12] KLEINER, B., MARTIN, R. D. and THOMSON, D. J. (1979). Robust estimation of power spectra. *J. Roy. Statist. Soc. Ser. B.* **41** 313–338 (Discussion, 338–351).
[13] MALLOWS, C. L. (1975). On some topics in robustness. Presented at IMS meeting, Rochester, N.Y.
[14] PAPANTONI-KAZAKOS, P. (1977). Some performance criteria incorporating data dependence in robust estimation. Unpublished manuscript.
[15] TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, Mass.
[16] VELLEMAN, P. (1975). Robust non-linear data smoothing. Tech. Rep. No. 89, Series 2, Dept. Statist., Princeton Univ.

BELL LABORATORIES
600 MOUNTAIN AVE.
MURRAY HILL, N.J. 07974