Robust Estimation of Power Spectra

B. KLEINER, R. D. MARTIN†
Bell Laboratories, Murray Hill, New Jersey 07974

and D. J. THOMSON
Bell Laboratories, Whippany, New Jersey 07981

[Read before the Royal Statistical Society at a meeting organized by the Research Section on Wednesday, May 9th, 1979, Professor J. F. C. KINGMAN in the Chair]

SUMMARY

Outliers in time series can play havoc with conventional power spectrum density estimation techniques when the spectrum features of interest consist of one or more peaks of relatively low amplitude. Furthermore, the outliers need not be large relative to the scale of the process in order to have considerable impact on the estimate; they only need to be large relative to the scale of the innovations process. Following a tutorial review of the central notions of robustness and the principles guiding conventional spectrum estimation, methods are presented for computing robust spectral densities. These are quite insensitive to such outliers and produce reliable estimates reflecting the structure of the core portion of the data. Two distinct methods are described. Both are based on robust prewhitening and on weighting the resulting residuals. They differ only in the way prewhitening is done: the first method computes the prewhitened values iteratively through a filtering operation; the second fits a robust autoregressive representation to the data by iterated weighted least squares. Each approach can be regarded as the solution to an intuitively appealing minimization problem, where the solution is exact in one case, approximate in the other. Both procedures have proved to be successful over a wide variety of data sets as will be demonstrated with real and simulated data.

Keywords: ROBUST PREWHITENING; M-ESTIMATES; GENERALIZED M-ESTIMATES; ROBUST AUTO-REGRESSION; ROBUST FILTERING; FOURIER TRANSFORM; POWER SPECTRUM

1. INTRODUCTION

In practice, statistical models only rarely describe a given situation exactly. One particularly important kind of inadequacy stems from the fact that the errors are usually assumed to have a Gaussian distribution. But actual data frequently contain a small fraction of unusual data points, or "outliers", which are not consistent with the Gaussian assumption. Such data could in principle be modelled by distributions which have a nearly Gaussian shape in the central region and tails which are considerably "heavier" or "thicker" than those of a Gaussian distribution corresponding to the central part. But accurate estimation of the tails of such distributions requires sample sizes which are much larger than one normally encounters in statistical practice. Thus, one will usually not be able to use fully efficient estimates, because the exact structure of a heavy-tailed nearly Gaussian density is not known.

On the other hand, it is often argued that the careful statistician will examine the data, discover the outliers, take appropriate ad hoc remedial action (such as removing or down-weighting questionable data points) and then apply conventional least squares or Gaussian maximum-likelihood procedures. The presumption here is that the ensuing inferences are more or less correct. Responsible and serious investigators do indeed often proceed this way with considerable success. For some outstanding examples see, for instance, Daniel and Wood (1971).

† R. D. Martin is with the Departments of Electrical Engineering and Biostatistics, University of Washington, Seattle, and a consultant to Bell Laboratories, Murray Hill, New Jersey.
However, there are serious shortcomings to this approach. First, the *ad hoc* nature of the procedures obviates the possibility of providing precise distributional characterizations of the final estimates (such as estimator efficiency under both Gaussian and alternative distributions). Second, the amount of effort, time and insight required to spot all but the most obvious outliers is often quite considerable (see again Daniel and Wood, 1971) and often the sheer bulk of data to be processed or real time requirements essentially rule out careful *ad hoc* approaches.

What is desired are statistical procedures whose performance may be evaluated for various distributions and which remain quite good for a broad class of underlying distributions (including the Gaussian and heavy-tailed nearby alternatives) but which are not necessarily best for any one of them. Such procedures are called *robust*.

There now exists a large and rapidly growing statistical literature dealing with robust estimation. A large portion of this literature treats location and linear regression models with independent and identically distributed errors. A few contributions deal with robust estimation of covariance matrices. Research on robust estimation in the time series context has lagged behind, and understandably so, in view of the increased difficulties imposed by dependency between the data points. Furthermore, in spectrum analysis it is sometimes claimed that, although outliers in time series may frequently occur, conventional methods of estimating power spectra have enough inherent robustness to make special methods unnecessary. This claim is based on the assumptions that one is usually interested primarily in the shape of a power spectrum and that the shape is quite insensitive to all but extremely bad contamination. This may be true for spectra consisting primarily of a narrow band or low frequency component with little else of interest. As we will show below, this is not the case when one is looking for comparatively low amplitude peaks in a wide-range spectrum, a situation frequently encountered in practice. We will also show that outliers in the original series need not be extremely large with respect to the scale of the observed process before they can obscure important features of the corresponding power spectrum. They only need to be large compared to the scale of the prediction residuals (sometimes called the innovations process) and may, therefore, be easily overlooked.

In the next section, we review some of the central notions of robustness and discuss theoretical and computational aspects of M-estimates. This appealing class of estimates was introduced by Huber for location (1964) and regression (1973) and is motivated by maximum-likelihood procedures. Section 3 deals with the lack of robustness of both parametric and non-parametric estimates of the spectrum. Section 4 recalls the conventional motivation for prewhitening techniques and proposes a general approach for robustly estimating spectral densities via robust prewhitening combined with the downweighting of outliers in the whitened process.

In Sections 5 and 6, two particular tactical approaches are suggested for carrying out the above strategy. Both procedures are modifications of the ordinary M-estimate approach to regression, and can be regarded as (perhaps approximate) solutions to an intuitively appealing minimization problem. Both procedures have proved to be successful over a wide variety of data sets with real and simulated data, as we demonstrate in Section 7.

### 2. Robustness Concepts and M-Estimates

To provide motivation and focus for the remainder of the paper this section summarizes some basic notions and background material on robustness. In addition robust maximum-likelihood type estimates (M-estimates) and generalized M-estimates (GM-estimates) for regression are discussed. The presentation is of necessity brief, but missing details are available in the references cited.

#### 2.1. Robustness Concepts

The loose definition of a robust estimate given in the previous section can be generalized. Suppose that $F$ is a nominal or conjectured distribution of the observations and $\{G_j\}$ is a family
of potential true observation distributions where each \( G \) is not too far from \( F \) in some appropriate metric. Then an estimate \( T \) is \textit{robust} at \( F \) if its performance is good over the family \( \{F\} \cup \{G\} \). The definition of the previous section and the focus of interest in this paper corresponds to a Gaussian \( F \) and heavy-tailed \( G \)'s. Although this is the case which has received the most attention in the literature, other situations may be of interest too. For example, Hogg (1974) has suggested that more attention should be devoted to short-tailed distributions and Hampel (1968) has discussed some possibilities where \( F \) is some non-Gaussian parametric model, such as the Gamma distribution.

Whether or not we limit ourselves to the nominally Gaussian \( F \) with heavy-tailed alternatives, as we shall do for the remainder of this paper, more precise notions of robustness are useful. The papers by Tukey (1960), Huber (1964) and Hampel (1971) provide three important concepts of robustness, namely \textit{efficiency} robustness, \textit{min–max} robustness and \textit{qualitative} robustness (see also Box, 1953, who introduced the term "robust" to the field of statistics).

An efficiency robust estimate is an estimate whose efficiency is high over \( \{F\} \cup \{G\} \). While the Cramer–Rao bound is usually adequate for asymptotic comparisons of efficiencies, in finite samples other standards of comparison, such as the best known estimate in a given class, may be appropriate (see, for example, Andrews \textit{et al.}, 1972), with suitably high efficiencies being somewhat arbitrarily taken as 90–95 per cent. In practice only a finite number of distributions can be considered and Tukey (1975) has emphasized “tri-efficiency”, selecting the Gaussian, a moderately contaminated normal and the Cauchy distribution to comprise \( \{G_1, G_2, G_3\} \).

Huber’s min–max robust location estimates minimize the maximum asymptotic variance over uncountably infinite families of symmetric distributions. These estimates also yield good, though not best, efficiency robustness.

Hampel’s qualitative robustness is a rather natural equicontinuity requirement. Suppose that \( F_0 \) and \( F \) are two distributions for independent and identically distributed observations, \( \{T_n\} \) a sequence of estimators indexed by sample size \( n \), and \( H(T_n, F_0), H(T_n, F) \) the distributions of \( T_n \) under \( F_0, F \) respectively. Then \( \{T_n\} \) is qualitatively robust at \( F_0 \) if the sequence of mappings \( F \rightarrow H(T_n, F) \) is equicontinuous at \( F \) in appropriate metrics. Hampel’s (1971) results do not apply in the time series setting and suitable alternatives are needed. One possibility has been recently proposed by Kazakos and Gray (1979).

Estimates which possess all three of the above robustness properties are quite attractive. However, the applied statistician might well be quite happy with just efficiency robustness and qualitative robustness.

### 2.2. The Lack of Robustness of Least Squares

Least squares estimates are notoriously lacking in robustness. Consideration of the sample mean \( \bar{y} \), for instance, reveals that the value of \( \bar{y} \) can be affected greatly by the presence of a single outlier. To use Tukey’s (1975) data-oriented term, the sample mean lacks “resistance”. Simple alternatives such as the sample median and trimmed means, as well as other modern robust estimates, are resistant.

The lack of efficiency robustness of \( \bar{y} \) towards contaminated normal models of the form \( f = (1-\gamma) N(0,1) + \gamma N(0,\sigma^2) \) was pointed out by Tukey (1960), and Hampel (1968, 1971) showed that \( \bar{y} \) is not qualitatively robust. The sample mean does possess certain min–max properties (Mallows, 1975), but since \( \bar{y} \) is neither efficiency robust nor qualitatively robust these are of no particular interest from the robustness point of view.

Similar comments apply to least squares estimates of coefficients in linear regression models. However, the regression situation is considerably more complicated than the location problem. One reason for this is discussed in the next section.

### 2.3. Regression M-estimates

There are by now quite a number of robust alternatives to least squares regression, including some which enjoy all three of the robustness properties we have mentioned. In fact
the rate at which new estimates are being proposed for some estimation problems may well exceed our ability to sort out the better estimates and understand the issues involved. For an indication of the wide variety of estimates that have been proposed for regression see, for example, Hill’s recent (1977) Ph.D. thesis.

Despite the large number of proposed estimates for some problems, there has been almost no attention devoted to the robust estimation of power spectra. Thus it seems appropriate to determine those situations for which the conventional procedures are not robust, and to devise robust procedures for estimating spectral densities as needed.

Since both classes of estimates we will propose for robust estimation of power spectra make use of modifications of ordinary M-estimates for regression (Huber, 1973), these estimates are briefly described here.

Consider the usual linear model

\[ Y_i = \sum_{j=1}^{p} x_{ij} \beta_j + e_i, \quad i = 1, \ldots, n. \]

An M-estimate \( \hat{\beta} \) is a solution of the minimization problem

\[ \min_{\hat{\beta}} \sum_{i=1}^{n} \rho \left( \frac{y_i - x_i^T \hat{\beta}}{\hat{\delta}} \right), \] (2.1)

where \( x_i^T = (x_{i1}, \ldots, x_{ip}) \), \( \beta^T = (\beta_1', \ldots, \beta_p') \), \( \rho \) is a symmetric loss function and \( \hat{\delta} \) is some robust estimate of the scale of the residuals. Equivalently if \( \psi(t) = \rho'(t) \) the estimate \( \hat{\beta} \) is determined by

\[ \sum_{i=1}^{n} x_i \psi \left( \frac{y_i - x_i^T \hat{\beta}}{\hat{\delta}} \right) = 0. \] (2.2)

The choice \( \rho(t) = t^2/2 \) gives \( \psi(t) = t \) and \( \hat{\beta} = \hat{\beta}_{LS} \), the usual least-squares estimate, whereas the choice \( \rho(t) = |t| \) yields \( \psi(t) = \text{sign}(t) \) and \( \hat{\beta} \) as the regression analogue of the median (often called the \( L_1 \)-estimate or the least absolute residuals estimate). In each of these cases \( \hat{\beta} \) does not depend on \( \hat{\delta} \). However, for other choices of robustifying \( \rho(t) \) it is necessary to compute \( \hat{\delta} \) to ensure scale invariance for \( \hat{\beta} \). Probably the best method of doing this is to follow Huber’s (1973, 1977) proposal to solve (2.2) simultaneously with the equation

\[ \sum_{i=1}^{n} \psi^2 \left( \frac{y_i - x_i^T \hat{\beta}}{\hat{\delta}} \right) = (n-p) A. \] (2.3)

The constant \( A \) is chosen so that \( \hat{\delta} \) converges to \( \delta \) when the \( e_i \) have a normal distribution with mean zero and variance \( \sigma^2 \).

Two popular choices for \( \psi \) are Huber’s psi-function

\[ \psi_H(t) = \begin{cases} t, & |t| < 1, \\ \text{sign}(t), & |t| \geq 1, \end{cases} \] (2.4)

and Tukey’s “redescending” bisquare function

\[ \psi_{BS}(t) = \begin{cases} t(1-t^2)^3, & |t| < 1, \\ 0, & |t| \geq 1. \end{cases} \] (2.5)

It is known that \( \psi_{BS} \) yields higher efficiencies than \( \psi_H \) for extremely heavy-tailed distributions (Andrews et al., 1972).
Another form of $\psi$, which was used implicitly by statisticians and engineers long before psi-functions entered the statistical literature, is the hard rejection rule. It rejects all points which are more than $3\sigma$ away from the centre of the data, i.e.

$$\psi_R(t) = \begin{cases} t, & |t| < 3, \\ 0, & |t| \geq 3. \end{cases}$$

Due to its discontinuities however, $\psi_R$ should not be used in robust regression.

For monotone $\psi$, Yohai and Maronna (1977) have established the existence and uniqueness of solutions of (2.2) and (2.3) under certain assumptions, but non-monotone $\psi$ such as $\psi_{BS}$ give rise to multiple solutions and consequently require special computational strategies.

The estimating equation (2.2) can be regarded as a weighted-least-squares (WLS) equation

$$\sum_{i=1}^{n} x_i w_i (y_i - x_i^T \hat{\beta}) = 0, \quad (2.6)$$

where the weights $w_i = w((y_i - x_i^T \hat{\beta})/s)$, with $w(t) = \psi(t)/t$, are data-dependent. For $\psi_H$, $\psi_{BS}$, or any other good $\psi$, large standardized residuals

$$r_i = \frac{y_i - x_i^T \hat{\beta}}{s}$$

produce small weights (of the order $r^{-1}$ for $\psi_H$ and weights identically zero for $\psi_{BS}$ when $|r| > 1$). Thus a small fraction of outliers will not have much influence on the estimate.

In addition to providing an intuitively appealing equivalent form of the M-estimate at the solution point, the WLS form suggests how to compute M-estimates iteratively using existing weighted-least-squares computer programs. Write (2.3) as well as (2.2) in WLS form and iterate both equations to approximate convergence, starting with $\hat{\beta}_{LS}$ and the robust initial scale estimate $\hat{s}_{0} = \text{median}(|y_i - x_i^T \hat{\beta}_{LS}|)/0.6745$.

This iterated-weighted-least squares (IWLS) algorithm works well in practice provided $\psi$ is monotone. A strategy to find the "correct" root of (2.2)-(2.3) for redescending $\psi$ (such as $\psi_{BS}$) which works well in practice is as follows: first obtain an IWLS solution $(\hat{\beta}, \hat{s})$ using a monotone $\psi$, and then compute an IWLS solution for a redescending $\psi$ by using $(\hat{\beta}, \hat{s})$ as a starting point and iterating only once or twice. Note that more than two iterations can be dangerous since the IWLS procedure need not converge for non-monotone $\psi$. Some other computational algorithms for solving (2.2)-(2.3) are discussed by Huber (1977).

2.4. Other Robustness Tools

Hampel (1971, 1974) also introduced the breakdown point and the influence curve, both of which provide useful indicators or robustness (the former being global in nature while the latter is local).

The breakdown point of an estimate is essentially the largest fraction of the data, over all combinations for each fraction, which can be taken to $\pm \infty$ without taking the (asymptotic) value of the estimate to the boundary of the parameter space. In location problems the natural boundary points are $\pm \infty$. For estimating a correlation the points $\pm 1$ might appear to be the natural boundary points; however, under the hypothesis that the true correlation is non-zero, it would be reasonable to include the origin as well as $\pm 1$ in the boundary set.

The influence curve measures the effect on an estimate of adding a data point $x$ in a limiting manner as the sample size $n \to \infty$. Given a distribution function $F$ and a statistic $T$ which is itself a function of a distribution function in the neighbourhood of $F$, the influence curve of the statistic $T$ at $F$ is defined as

$$IC(x; T, F) = \lim_{\varepsilon \downarrow 0} \frac{T((1-\varepsilon)F + \varepsilon \delta_x) - T(F)}{\varepsilon},$$
Hampel (1974) argues that a highly robust estimate should have the following two properties (among others): (i) the highest possible breakdown point ($\frac{1}{2}$ is an upper bound) and (ii) a bounded and continuous influence curve which "redescends" to zero.

Consider the special location estimate case of (2.2)

$$\sum_{i=1}^{n} \psi \left( \frac{x_i - \hat{\mu}}{\hat{s}} \right) = 0. \quad (2.7)$$

Unlike in the general regression case, it is possible here to obtain an adequate robust estimate directly from the data. It turns out (Hampel, 1971) that the resulting $\hat{\mu}$ has a breakdown point of $\frac{1}{2}$ provided $\hat{s}$ has a breakdown point of $\frac{1}{2}$ (which is the case if, for example, $\hat{s} = \text{median} \{|y_i - \text{median} \{y_i\}|\}$). By way of comparison the sample mean has a breakdown point of zero, and an $\alpha$-trimmed mean has breakdown point $\alpha$.

The influence curve $\text{IC}(\psi; \hat{\mu}, F)$ turns out to be equal to $\psi((y_i - \mu)/s)$ except for a scale factor depending on $F$, where $\mu$ and $s$ are the asymptotic values of $\hat{\mu}$ and $\hat{s}$. Thus for location $M$-estimates continuity and boundedness of $\psi$ will result in the influence curve being continuous and bounded. This in turn yields qualitative robustness (which is essentially a weak-continuity property) for $\hat{\mu}$. The sample mean lacks qualitative robustness because it has an unbounded influence curve. In the general regression problem the situation is more complicated and additional steps are needed to obtain bounded and continuous influence curves.

### 2.5. GM-estimates for Regression

If the possibility of contamination in the $x_{ij}$ is ignored and only contamination of the $y_i$ is allowed, then the breakdown point of the regression $M$-estimate is $\frac{1}{2}$, as long as $\psi$ is bounded. However, allowing for the realistic possibility of contamination or gross errors in the $x_{ij}$ yields a breakdown point of zero, a clearly unsatisfactory situation.

Computation of the influence curve for a regression $M$-estimate gives

$$\text{IC}(\psi; \hat{\beta}, F) = Bx \psi \left( \frac{y - x^T \beta}{s} \right) \quad (2.8)$$

where $x^T = (x_1, ..., x_p)$, $\beta$ and $s$ are the asymptotic values of $\hat{\beta}$ and $\hat{s}$, and $B$ is a $p \times p$ matrix which depends upon the distribution $F$ of $(y, x^T)$. Even for monotone and bounded $\psi$ this influence curve remains unbounded as a function of $x$, for each fixed $y$. Thus outliers in the independent variable can have a considerable influence on the estimate.

In order to deal with this difficulty Hampel (1974) and Mallows (1975) have proposed the use of bounded influence regression estimates. Such estimates might naturally be called generalized $M$-estimates (GM-estimates). One version of GM-estimate is obtained by solving

$$\sum_{i=1}^{n} W(x_i) \psi \left( \frac{y_i - x_i^T \hat{\beta}_GM}{\hat{s}} \right) = 0, \quad (2.9)$$

where $W(x_i)$ is a continuous scalar-valued weight function chosen so that $W(x_i) x_i$ is bounded. The scale estimate $\hat{s}$ would be obtained by simultaneously solving (2.3) along with (2.9).

The influence curve for $\hat{\beta}_{GM}$ is then

$$\text{IC}(\psi; \hat{\beta}_{GM}, F) = BW(x) x \psi \left( \frac{y - x^T \beta}{s} \right), \quad (2.10)$$

which is a bounded and continuous function of $(y, x^T)$. A GM-estimate having such an influence curve is qualitatively robust. A good choice for $W(x)$ is

$$W(x) = W(p^{-1} x^T C^{-1} x), \quad (2.11)$$
where \( W \) is a non-negative scalar-valued weight function which is monotone decreasing on \([0, \infty)\), and \( \hat{C} \) is a robust estimate of the covariance matrix for the \( x_i \). The breakdown point of the resulting GM-estimate is positive, and is actually \( \frac{1}{2} \) if \( \hat{C} \) has a breakdown point of \( \frac{1}{2} \).

Favourable small-sample efficiency robustness properties of GM-estimates have been reported by Denby and Larsen (1977) and Hill (1977). The use of GM-estimates for fitting autoregressions is discussed in Section 6.

### 3. The Power Spectrum and the Need to Estimate it Robustly

The simplest and most commonly used representation for a stationary time series \( \{y_k\} \) is the second-order characterization in terms of its mean, \( \mu = E y_k \), and its autocovariance sequence \( C(l) = \text{cov} [y_k, y_{k+l}], \ l = 0, 1, 2, \ldots \). Although \( \mu \) and \( \{C(l) : \ l = 0, 1, 2, \ldots \} \) provide a complete characterization for the Gaussian case, this is not true for an arbitrary stationary series, for which specification of all finite-order joint distribution functions is required. Since the estimation of higher order properties of a time series is rather cumbersome, most investigators content themselves with estimating the second-order properties such as the mean and the covariance \( C(l) \) for a finite number of lags \( l = 0, 1, \ldots, L \) or equivalently the spectrum as defined below.

#### 3.1. The Non-parametric Spectrum

Any second-order stationary process \( \{y_k\} \) has the spectral representation

\[
y_k = \int_{-\frac{1}{2}}^{\frac{1}{2}} \exp (i2\pi f) dZ(f),
\]

where \( Z(f), -\frac{1}{2} \leq f \leq \frac{1}{2} \), has orthogonal increments. The process \( Z(f) \) defines a monotone increasing function \( F \) through

\[
F(f) = E|Z(f)|^2, \quad dF(f) = E|dZ(f)|^2, \quad F(-\frac{1}{2}) = 0, \quad F(\frac{1}{2}) = \sigma^2 = C(0).
\]

The function \( F(f) \) is called the spectral distribution function, and, when it exists, its derivative \( S(f) = F'(f) \) is called the spectral density function of \( \{y_k\} \). Other commonly used terms for \( S(f) \) are power spectral density, spectrum and power spectrum.

It follows from (3.1) and the orthogonal increments property of \( Z(f) \) that

\[
C(l) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \exp (i2\pi lf) S(f) df, \quad l = 0, \pm 1, \ldots
\]

when the spectral density \( S(f) \) exists. Thus the covariances \( C(l), l = 0, \pm 1, \ldots \), are the Fourier coefficients of \( S(f) \) and so

\[
S(f) = \sum_{l=-\infty}^{\infty} C(l) \exp (-i2\pi lf),
\]

which is periodic, with period unity. Many textbooks (especially those oriented towards application) define the spectral density directly by (3.4) without ever mentioning the spectral decomposition (3.1). However, the latter is quite basic and in essence says that any stationary time series can be interpreted as the limit of finite sums of sinusoids of the form

\[
A_l \cos (2\pi f_l t + \Phi_l)
\]

over all frequencies \( f = f_l \). The amplitudes \( A_l = A(f_l) \) and phases \( \Phi_l = \Phi(f_l) \) are uncorrelated random variables with \( S(f) \approx E A^2(f_l) \) for \( f_l \approx f \neq 0 \). The \( A_l \) and \( \Phi_l \) can be related to the increments \( Z(f_l+\Delta f/2) - Z(f_l-\Delta f/2) \) in a simple way. For details see Doob (1953).
3.2. The Autoregressive Spectrum

There are three major time domain models for stationary time series: the moving average model, the autoregressive model, and the mixed autoregressive moving average model. While the finite mixed model often provides the most parsimonious representation and has therefore been used extensively in time domain modelling (see Box and Jenkins, 1970), the finite autoregressive model has received by far the most attention in the context of spectrum estimation. One reason for this is the computational attractiveness of the associated linear least-squares estimation problem, where the presence of moving average terms leads to non-linear problems. Furthermore, under fairly reasonable assumptions (a continuous and non-zero spectrum) a purely non-deterministic stationary time series has the infinite order autoregressive representation

\[ x_k = \sum_{j=1}^{\infty} \phi_j x_{k-j} + \varepsilon_k, \]  

(3.5)

where \( \sum_{j=1}^{\infty} \phi_j^2 < \infty \) and the \( \varepsilon_k \) are uncorrelated random variables with mean 0 and finite variance \( \sigma^2_\varepsilon \). The \( \varepsilon_k \) were labelled the innovations process by Wold (1954). The spectrum is of the form

\[ S(f) = \sigma^2_\varepsilon |H(f)|^2 \]  

(3.6)

with

\[ H(f) = 1 - \sum_{j=1}^{\infty} \phi_j \exp(i2\pi fj). \]

For any process \( \{x_k\} \) with a continuous spectrum \( S(f) \) there exists a finite-order autoregressive process whose spectrum \( \tilde{S}(f) \) approximates \( S(f) \) arbitrarily well uniformly in \( f \) (this is essentially the Weierstrass theorem for approximating a continuous function on a closed interval with polynomials—in this case trigonometric polynomials).

3.3. Non-parametric Estimation of the Spectrum

The most frequently used method of estimating the spectrum today is non-parametric and is based on smoothing the periodogram. Given the observed time series \( x_k, 1 \leq k \leq n \), the basic steps in constructing the estimate are as follows. A data window or taper \( d_k, 1 \leq k \leq n \), which slopes gently towards zero at both ends, is used to form the modified data \( \tilde{x}_k = d_k x_k \). The purpose of this tapering is to reduce "leakage" (see, for example, Bloomfield, 1976) associated with the finiteness of any actual set of data. Then the discrete Fourier transform of \( \tilde{x}_k \)

\[ X(f_l) = \sum_{k=1}^{n} \tilde{x}_k \exp\left(-\frac{i2\pi kl}{n}\right), \quad f_l = \frac{l}{n}, \quad l = 0, 1, \ldots, \left[\frac{n}{2}\right], \]  

(3.7)

is computed using the fast Fourier transformation, and the periodogram

\[ \tilde{S}_x(f_l) = n^{-1} |X(f_l)|^2 \]  

(3.8)

is formed. Since the periodogram is not a consistent estimate, it is smoothed to get a respectable estimate of the power spectrum:

\[ S_\tilde{x}(f_l) = \sum_{j=-L}^{L} w_j \tilde{S}_x(f_{l+j}), \quad w_j = w_{-j}, \quad \sum_{j=-L}^{L} w_j = 1. \]  

(3.9)

It is also possible to reduce variability by averaging periodograms computed from different (often overlapping) time segments (see Welch, 1967).
A good spectrum analysis by the above method usually requires computing several estimates based on different amounts of smoothing in order to exploit the trade-offs between bias and variance. The leakage problem can be alleviated by either applying one of the better tapering functions such as the cosine window, the Parzen window or the prolate spheroidal window (see Thomson, 1977), or through prewhitening (discussed in Section 4).

Further details concerning the above “direct” method (by way of contrast with the prefast Fourier transform approach based on using covariance estimates $\hat{C}(l)$, $l = 0, 1, \ldots, L$ in (3.4)) may be found, for example, in the introductory text by Bloomfield (1976).

### 3.4. The Parametric Autoregressive Estimate

In recent years there has been considerable interest in estimating the spectrum by fitting a finite-order autoregressive model, an approach heavily based on the approximation property mentioned in Section 3.2. Thus least-squares estimates $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$, $\hat{\sigma}_{\hat{\phi}}^2$ are obtained from the data and $S(f)$ is estimated by the autoregressive spectral density estimate

$$\hat{S}(f) = \frac{\hat{\sigma}_{\hat{\phi}}^2}{1 - \left| \sum_{j=1}^{p} \hat{\phi}_j \exp(i2\pi f j) \right|^2}, \quad (3.10)$$

computed at a grid of frequencies.

For the autoregressive estimate there is of course the problem of choosing a good approximating order $p$, which is analogous to choosing the amount of smoothing in the smoothed-periodogram approach.

Two commonly used and reasonable order-selection methods are Akaike’s (1969) final prediction error criterion and Parzen’s (1974) criterion autoregressive transfer function (CAT).

In Akaike’s stopping rule the order $\hat{p}$ of the autoregression is the value of $p$ which minimizes the final prediction error (FPE)

$$\text{FPE}(p) = \frac{n + p + 1}{n - p - 1} \hat{s}_p^2, \quad (3.11)$$

where $\hat{s}_p^2$ is an estimate of the residual variance of the fitted autoregressive process of order $p$.

The FPE stopping rule is an approximate version of the AIC maximum likelihood criterion of Akaike (1974) under the assumption of a Gaussian process.

Parzen’s criterion chooses the order $\hat{p}$ which minimizes

$$\text{CAT}(p) = 1 - \frac{s_{\infty}^2}{\hat{s}_p^2} + \frac{p}{n},$$

where

$$s_{\infty}^2 = \exp \left[ \frac{1}{n} \sum_{l=0}^{n-1} \ln \{ S(f_l) \} \right], \quad (3.12)$$

is an estimate of the residual variance of the infinite-order autoregressive fit and $s_p^2$ is the estimated residual variance of autoregressive fit of order $p$.

Although Parzen’s and Akaike’s criteria are philosophically different, Parzen’s experience (1974), as well as ours, show that the two criteria yield very similar results in practice.

### 3.5. Time Series Outliers and the Need for Robust Spectrum Estimates

Most of the work on robust estimation to date has been concerned with distributional robustness for problems where the observations are independent and identically distributed. In time series analysis realistic modelling of outliers poses a more serious challenge because...
the observations are no longer independent and outliers can occur under a wide variety of circumstances or be due to very different mechanisms. In the face of these difficulties the problem is to specify some simple but reasonably representative outlier models which are tractable, and which lead to the construction of good robust estimates.

Fox (1972) introduced two distinct kinds of outliers for autoregressive time series: innovations outliers (IO) and additive outliers (AO). Their probabilistic models, in their simplest forms, can be described in terms of a single heavy-tailed univariate distribution. The IO model is of the form

\[ x_k = \phi_1 x_{k-1} + \ldots + \phi_p x_{k-p} + \epsilon_k, \tag{3.13} \]

where the distribution \( F_\epsilon \) of the innovations \( \epsilon_k \) is heavy-tailed. The AO model is

\[ y_k = x_k + \nu_k, \tag{3.14} \]

where \( x_k \) is a Gaussian \( p \)th-order autoregression and the \( \nu_k \) are independent identically distributed with common distribution \( F_\nu = (1-\gamma)\delta_0 + \gamma H \); \( \delta_0 \) is the degenerate distribution having all its mass at the origin and \( H \) is a heavy-tailed symmetric distribution with mean 0 and variance \( \sigma^2 \). Thus \( P(\nu_k = 0) = 1-\gamma \) and, if \( \gamma \) is small, the \( x_k \) are observed perfectly most of the time. As we will argue below and show by a striking example in Section 7, \( H \) need not be heavy-tailed relative to the scale of the process in order to have a serious impact on spectral density estimates.

The above models are obviously rather simple and it is easy to imagine several modifications which might result in more realistic outlier generating mechanisms. Of these perhaps the most important would be to drop the assumption that the \( \nu_k \) are independent while maintaining \( P(\nu_k = 0) = 1-\gamma \) with \( \gamma \) small. This would account for patchy outliers which occur rather frequently in practice. On the other hand, even the simplest outlier models raise serious enough difficulties and questions to warrant considerable study.

Innovations outliers do not pose serious problems for spectrum estimation when the shape of the spectrum is of primary interest. Only the scaling factor requires a robust estimate. For large sample autoregressive spectral density estimates of the form (3.10) this follows from the fact that the asymptotic covariance matrix of the least squares estimates of the autoregressive coefficients \( \hat{\phi}_{1:p} \) is independent of \( F_\epsilon \). This was pointed out by Whittle (1962) who hastened to add that the associated \( \hat{\sigma}_\epsilon^2 \) did not enjoy this robustness property. For smoothed-periodogram estimates the claim may be verified by checking the bias and variance expressions for the periodogram.

The real problem for spectral analysis then are additive outliers. First consider the mean and variance of the periodogram estimate based on the observed data \( y_1, \ldots, y_n \).

For the special case of a single fixed additive outlier (i.e. \( \nu_k = 0 \forall k \neq j, \nu_j = A \)) the periodogram \( \hat{S}_\nu(f) \) of \( y_k \)

\[ \hat{S}_\nu(f) = \hat{S}(f) + \frac{A^2}{n} + \frac{A}{n} 2 \text{Re}\left( \frac{X(f)}{f} \right) \]

is raised by \( \frac{A^2}{n} \) and retains its shape only if the oscillatory term \( \frac{A^2}{n} 2 \text{Re}\left( \frac{X(f)}{f} \right) \) is considerably smaller than \( \hat{S}(f) \) for all frequencies \( f \). This means that if the spectrum varies over a wide range (e.g. if the spectrum contains a large portion of its total power in a narrow frequency band and has small but important peaks elsewhere), even relatively minor outliers can cause serious problems, because their effect, although small compared to the total power, may be large relative to the power in small peaks.

Two single outliers will give rise to oscillatory behaviour in the low-amplitude portion of the spectrum. Example 4 in Section 7 illustrates this point as well as the fact that outliers can be quite small and still have serious consequences. More than two outliers will give rise to more complicated behaviour which is best described in terms of the mean and variance of the spectrum estimate.
If \( E\nu_k = E\nu_k^2 = 0 \) and \( E\nu_k^2 < \infty \) then straightforward computation gives
\[
E\hat{S}_0(f) = E\hat{S}_x(f) + E\hat{S}_i(f), \quad \var\hat{S}_0(f) = \var\hat{S}_x(f) + \var\hat{S}_i(f) + 2E\hat{S}_x(f)\hat{S}_i(f) \quad (3.16)
\]
for the periodogram \( \hat{S}_0(f) \) based on the \( y_i \). For i.i.d. \( \nu_k \), \( E\hat{S}_i(f) = \sigma_i^2 \) and \( \var\hat{S}_i(f) = \sigma_i^4 \). Thus \( \hat{S}_0(f) \) can be badly biased and also have a seriously inflated variance. This can be the case even for \( \sigma_i^2 < \sigma_x^2 \) if \( \hat{S}_x(f) \) is small in the region of interest (as it often is).

The bias and increased variability in the periodogram will result in a corresponding distortion and increased variability in the smoothed periodogram as an estimate of the power spectrum.

Under the additive outlier model (3.14) the least squares estimates \( \hat{\Phi}_{LS} \) computed from the observed series \( y_k \) will be biased. This bias of \( \hat{\Phi}_{LS} \) can be disastrous for large \( \nu_k \)'s, even if only a small fraction of the \( \nu_k \)'s are non-zero. For example, when \( x_k \) is of order 1 with parameter \( \phi \), \( \hat{\Phi}_{LS} \to \phi_0 \) almost surely with
\[
\phi_0 = \frac{Ey_iy_{i+1}}{Ey_i^2} = \frac{Ex_i x_{i+1} + E\nu_i \nu_{i+1}}{\sigma_x^2 + \sigma_i^2}, \quad (3.17)
\]
where \( \sigma_x^2 = \gamma \sigma^2 \). Therefore the asymptotic bias of \( \hat{\Phi}_{LS} \) for independent \( \nu_k \) is
\[
B_{LS} = - \phi \frac{\sigma_x^2}{\sigma_x^2 + \sigma_i^2} \quad (3.18)
\]
which only vanishes for \( \phi = 0 \) or \( \sigma_x^2/\sigma_i^2 \to \infty \), but can be quite substantial even for moderate amounts of contamination. For instance if \( \gamma = 0.1 \) and \( \sigma^2/\sigma_x^2 = 10 \) the bias is 50 per cent. The variance of \( \hat{\Phi}_{LS} \) is also seriously inflated. Similar results can be obtained for higher order processes \( p \) and there is a corresponding impact on the autoregressive spectrum estimate.

We have seen that the use of either the smoothed periodogram approach or the autoregressive method to estimate the power spectrum can yield poor results when the data contain additive outliers. Furthermore, additive outliers in the original series need not be extremely large with respect to the scale of the observed process to obscure important features of the corresponding power spectrum. They only need to be large compared to the scale of the innovations process and may therefore be easily overlooked.

In the following sections we discuss alternative ways of estimating power spectra which are not easily disturbed by such outliers.

4. ROBUST PREWHITENING

4.1. Prewhitening

Prewhitening is an all-too-often neglected technique for reducing the dynamic range of the spectrum by filtering the data so as to transform the original time series into a series whose spectrum is nearly flat (or white). Its main advantage is that smoothed periodogram estimates based on the whitened series do not suffer nearly as much from leakage as do estimates based on the original data. Thus fine structure in the spectrum, such as two close low-amplitude peaks, is more likely to be spotted in such an estimate.

The need for prewhitening along with its utility and interpretation was discussed by Blackman and Tukey (1958). In a later article Tukey (1967) says: "If low frequencies are \( 10^4 \), \( 10^4 \) or \( 10^4 \) times as active as high ones, a not infrequent phenomenon in physical situations, even a fairly good window is too leaky for comfortable use. The cure is not to go in for fancier windows, but rather to preprocess the data towards a flatter spectrum, to analyse this 'prewhitened' series, and then to adjust its estimated spectrum estimates for the easily computable effects of preprocessing."

The smoothed periodogram and autoregressive spectral density estimates are limiting versions of the prewhitening approach. The former amounts to no prewhitening and the
latter corresponds to total prewhitening if the process is truly a finite-order autoregression with known order $p$. Since this is an unlikely event in practice, the autoregressive approach is almost always an approximation. Once we openly admit the approximate nature of the finite-order autoregressive spectral density, we should be rather uncomfortable with the use of $\hat{\Theta}_p^2$ in the numerator of the estimate (3.6), because the approximation may very well miss important low-amplitude details.

Thus we prefer an autoregression prewhitened spectral density estimate having the general character suggested by Tukey:

$$S(f) = \frac{\tilde{S}_r(f)}{[\tilde{H}_p(f)]^2}, \quad (4.1)$$

where $\tilde{S}_r(f)$ is a smoothed periodogram estimate based on tapered versions of the prediction residuals $r_k = y_k - \sum_{j=1}^p \hat{\phi}_j y_{k-j}$ and

$$\tilde{H}_p(f) = 1 - \hat{\phi}_1 \exp(i2\pi f) - \hat{\phi}_2 \exp(i2\pi 2f) \cdots - \hat{\phi}_p \exp(i2\pi pf). \quad (4.2)$$

The main aspects of prewhitening which have changed since Blackman and Tukey (1958) are (i) the willingness to fit autoregressive models with orders much higher than two or three (using modern numerical algorithms for the least squares fitting procedure orders of 20, 30 or higher may even be appropriate for some applications; Thomson, 1977), and (ii) the availability of reasonable order selection criteria such as $\text{AIC}(p)$ and $\text{CAT}(p)$.

It is not entirely clear to us at this point how to use order selection rules in computing the hybrid parametric/non-parametric estimate (4.1). Though our preference is for orders somewhat lower than those obtained with a minimum $\text{AIC}(p)$ or $\text{CAT}(p)$ criterion, providing the data contain no outliers, we will not belabour the point here since precise estimation of the order seems relatively unimportant. What is important is that the order be high enough so that the residual spectrum is relatively flat.

### 4.2. Robust Prewhitening

When a time series contains outliers, the advantages of prewhitening as a technique of estimating the spectral density function become even more apparent.

As pointed out in Section 3, it is not sufficient to trim or delete only those time series outliers which are large relative to the scale of the process, but one must also contend with outliers which are only large relative to the scale $\sigma$ of the innovations or prediction error process. Since prewhitening procedures involve the computation of linear predictor coefficients and associated prediction residuals, they provide a natural framework for dealing with such outliers in a manner that will result in a robust spectral density estimate.

Our basic proposal is to robustify both the numerator $\tilde{S}_r(f)$ and the denominator $\tilde{H}_p(f)$ in (4.1). Thus we shall make use of (i) robust estimates $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$ of autoregressive/linear predictor parameters, and (ii) a procedure for obtaining prediction residuals which are free of outliers but are otherwise relatively unaltered—these “clean” residuals are then used in computing $\tilde{S}_r(f)$. In the next two sections we describe two distinct methods for carrying out this idea.

### 5. Estimates based on Robust Filtering

#### 5.1. The Estimation Procedure

For simplicity let us assume that the observed values $y_k$ have mean zero (translation equivariant versions can be obtained in a straightforward manner). Then the iterative procedure is initiated by fitting (by least squares) the autoregression

$$y_k = \sum_{j=1}^p \hat{\phi}_j y_{k-j} + \epsilon_k \quad (5.1)$$
to the data $y_k$, $k = 1, 2, \ldots, n$, where $p$ is chosen according to one of the stopping rules described in Section 3.4. The estimate $\hat{\Phi}_T = (\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p)$ is then used in a robust filtering algorithm which operates on the original observations $y_k$ to produce $\{\hat{x}_k\}$, a robust estimate of the underlying process $\{x_k\}$.

The robust filtering algorithm is

$$\hat{x}_k = \hat{x}_{k-1} + c\hat{\psi}\left(\frac{y_k - \hat{x}_{k-1}}{\hat{\delta}}\right),$$

(5.2)

where $\hat{x}_{T-1} = (\hat{x}_{k-1}, \hat{x}_{k-2}, \ldots, \hat{x}_{k-p})$ and $\hat{\delta}^2$ is an estimate of the innovations variance $\sigma_x^2$ of $x_k$.

$c$ is the constant chosen to obtain high efficiencies for both $\gamma = P(\nu_k \neq 0) = 0$ and $\gamma > 0$ with $\gamma$ not too large.

For the first iteration $\hat{s}$ is the positive square root of a robust variance estimate based on the prediction residuals $\{e_k\}$ in (5.1). When the $y_k$ contain outliers none of the initial estimates $\Phi$, $\hat{x}_0$ or $\hat{x}_k$ is necessarily very good. However, these estimates can be improved through iteration. Robustly prewhitened residuals $\{r_k\}$ are obtained by

$$r_k = x_k - (\hat{\phi}_1 x_{k-1} + \hat{\phi}_2 x_{k-2} + \ldots + \hat{\phi}_p x_{k-p}) = c\hat{\psi}\left(\frac{y_k - \hat{x}_{k-1}}{\hat{\delta}}\right).$$

(5.3)

and these residuals are used to compute a new estimate $\hat{s}$. The values $\hat{x}_k$ are inserted in place of the $y_k$ into the autoregressive model (5.1) to compute new estimates $\hat{\Phi}$ and then $\hat{s}$ and $\hat{\phi}$ are used in (5.2) to get new $\hat{x}_k$ etc. The whole process is iterated until $\hat{\phi}$ shows only small relative changes. Experience indicates that in practice at most three to four iterations are required for data with not more than 10 per cent gross outliers.

Finally $\hat{s}(f)$ is computed from the $\{r_k\}$ and $\hat{\phi}$ is inserted in (4.2) to give the robust spectrum estimate (4.1). For computational details see Thomson (1977).

If $\{x_k\}$ is a finite autoregressive process, $\Phi$ in (5.2) is known and $c\hat{\delta}$ is replaced by an appropriate fixed value, the above recursion is a member of a class of robust filters studied by Masreliez and Martin (1977). These filters are robust in the sense that their mean-squared error is bounded over a broad family of distributions including the Gaussian and heavy-tailed alternatives. For the special case of first-order autoregression this bound has been shown to be quite close to the minimum attainable mean-squared error over a wide range of parameter values.

### 5.2. An Optimality Property of $\hat{x}_k$

The algorithm described above is actually an approximate solution to an intuitively appealing minimization problem. Replace $\Phi$ by $\Phi'$ and denote the estimate (5.2) by $\hat{x}_k(\Phi')$. Then we define an estimate $\Phi$ as a solution of the minimization problem

$$\min_{\phi', \delta'} \left\{ \sum_{k=p+1}^n \rho\left(\frac{y_k - \hat{x}_{k-1}((\phi') \phi')}{\hat{\delta}}\right) + (n-p)\ln \hat{\delta}' \right\}.$$  

(5.4)

Since the solution is a stationary point we have with $\psi = \rho'$

$$\sum_{k=p+1}^n \left[ \frac{\partial \hat{x}_{k-1}(\Phi')}{\partial \phi_l} \right] \psi\left[\frac{y_k - \hat{x}_{k-1}((\phi') \phi')}{\hat{\delta}}\right] = 0$$

(5.5)

with

$$(G_{k-1}(\Phi))_{ji} = \left. \frac{\partial \hat{x}_{k-1}((\phi') \phi')}{\partial \phi_l} \right|_{\phi' = \hat{\phi}} = \delta_{jl}, \quad j, l = 1, \ldots, p,$$

and

$$\frac{1}{n-p} \sum_{k=p+1}^n \frac{y_k - \hat{x}_{k-1}((\phi') \phi')}{\hat{\delta}} \psi\left(\frac{y_k - \hat{x}_{k-1}((\phi') \phi')}{\hat{\delta}}\right) = 1.$$  

(5.7)
If we assume that \( G_{k-1}(\phi) \phi \) is small compared with \( \hat{\phi}_{k-1}(\Phi) \) for all \( k = p + 1, \ldots, n \) and all \( \Phi \) corresponding to a stationary process (heuristic arguments and Monte Carlo results indicate that this is a reasonable assumption), then (5.5) reduces to

\[
\sum_{k=p+1}^{n} \hat{\phi}_{k-1}(\Phi) \psi \left( \frac{y_k - \hat{y}_{k-1}^T(\Phi) \phi}{\delta} \right) = 0. \tag{5.8}
\]

Thus the estimate \( \phi \) is the result of regressing the observed values \( \{y_k\} \) robustly on robust estimates \( \{\hat{x}_k\} \) of the unobservable values \( \{x_k\} \). This class of estimates appears to be a simple and natural analogue of M-estimates for ordinary regression which is appropriate for time series with additive outliers.

It is now easy to see that (5.8) is equivalent to the estimation procedure described in the previous section (with the exception of the scale parameter \( s \), which is estimated differently). Just rearrange (5.2) as

\[
\hat{x}_k - \hat{x}_{k-1}^T \Phi = c \delta \psi \left( \frac{y_k - \hat{y}_{k-1}^T \phi}{c \delta} \right) \tag{5.9}
\]

and use the fact that the \( \phi \) satisfy equations of the Yule–Walker form obtained by replacing \( x_k \)'s with \( \hat{x}_k \)'s (note that from now on \( \phi \) arguments of \( \hat{x}_k(\phi) \) are suppressed for convenience):

\[
\sum_{k=p+1}^{n} \hat{x}_{k-1}(\hat{x}_k - \hat{x}_{k-1}^T \Phi) = c \delta \sum_{k=p+1}^{n} \hat{x}_{k-1} \psi \left( \frac{y_k - \hat{y}_{k-1}^T \phi}{c \delta} \right) = 0. \tag{5.10}
\]

Thus robust filtering is approximately equivalent to iteratively computing the estimates (5.8) to convergence, while being much simpler computationally.

The price paid for using the robust filter estimate when \( x_k \) is Gaussian and \( \nu_k = 0 \) is increased variability and some bias. These are obviously difficult to evaluate analytically in view of the complexity of the estimator; some limited Monte Carlo experience indicates that the loss in mean-squared-error efficiency will be small in the Gaussian situation. Some rough variational arguments described in Appendix A indicate that the bias in the spectrum is small and is proportional to the true spectrum.

6. ROBUST AUTOREGRESSIVE SPECTRUM ESTIMATION

This approach obtains estimates of the spectral density by first robustly fitting an autoregressive model to the data and then operating on the resulting residuals.

For simplicity let us again assume that the observed values \( y_k \) have mean zero. Then the basic idea behind our estimation procedure for the robust autoregressive coefficients \( \Phi_T = (\phi_1, \ldots, \phi_p) \) is a generalization of Huber's (1973) M-estimates for regression so that the summands of the estimating equations are bounded and continuous functions of the data. This in turn results in an influence function having the same properties, the desirability of which has been stressed by Hampel (1974) (see Section 2.4).

This is done by introducing a weighting function \( W \) into equations (2.2) and (2.3) so that \( \Phi \) and \( \delta \) are solutions of

\[
\sum_{i=p+1}^{n} W(y_{i-1}) y_{i-1} \psi \left( \frac{y_i - y_{i-1}^T \Phi}{\delta} \right) = 0, \tag{6.1}
\]

\[
\frac{1}{n-2p} \sum_{i=p+1}^{n} W(y_{i-1}) \psi^2 \left( \frac{y_i - y_{i-1}^T \Phi}{\delta} \right) = A, \tag{6.2}
\]

where

\[
W(y_{i}) = W((y_{i}^T \hat{\Phi}^{-1} y_{i})/|p|). \tag{6.3}
\]
with \( W(\cdot) \) a non-negative continuous function such that \( W(y_i)y_i \) is bounded (cf. Section 2.5). The \( p \times p \) matrix \( \hat{C}^{-1} \) is a robust estimate of the inverse of the covariance matrix \( C \) of the underlying process \( \{x_k\} \); it has to be estimated from the observed data \( y_k \). The constant \( A \) is chosen so that \( \hat{s} \) is an asymptotically unbiased estimate of the standard deviation \( \sigma \) when the data are Gaussian (some specific values are given by Martin and Zeh, 1978). Equations (6.1) and (6.2) can be solved by using iterated weighted least squares (IWLS); their solutions are called generalized M-estimates (GM-estimates).

It can be shown that GM-estimates are qualitatively robust and have positive breakdown point under reasonable assumptions. It is somewhat disappointing that the breakdown point is bounded above by \( 1/(p+1) \) under the AO model (3.14); it can be higher if the outliers \( v_k \) are not independent.

For \( W(y_i) = 1 \) the GM-estimates reduce to the autoregressive version of ordinary M-estimates for regression. These are however quite non-robust toward additive outliers having asymptotic biases nearly as large as those of least squares estimates (Denby and Martin, 1979).

Computational details concerning the IWLS procedure and the estimate of \( \hat{C}^{-1} \) can be found in Martin and Zeh (1978). This reference also gives examples of applications of GM-estimates to real data.

To compute the robust spectral density (4.1) the GM-estimates \( \hat{\phi} \) are inserted into (4.2). The residuals \( r_i \) used to compute the numerator \( S_r(f) \) can be obtained either by

\[
r_i = W(y_{i-1}) s \phi \left( \frac{y_i - y_{i-1}^T \hat{\phi}}{s} \right)
\]

(6.4)

or by inserting \( \hat{\phi} \) into (5.2) and computing the residuals according to (5.3). The latter version appears to give somewhat better results.

7. Examples

Four examples are shown here, two with real data, two with artificial data.

For all examples exactly the same computational procedures are used and all parameters are kept fixed at the same values. This is done to demonstrate that the procedures can be used for quite different types of data without the need for constantly adjusting parameters. Of course this does not preclude the use of specific knowledge about a particular set of data to fine-tune the procedures and parameters for even better results.

All examples show a basic set of four to five curves (sometimes augmented by further figures) which includes:

1. At least half of the raw data in the time domain.
2. The non-robust classical periodogram-based power spectrum (always shown as a dashed line): The time series, minus its mean, is tapered by the prolate spheroidal window with parameter \( c = 4\pi \) (see Thomson, 1977); then its discrete Fourier transform is computed at 1,025 (regardless of sample size) equidistant frequencies between 0 and 0.5 (the Nyquist frequency), the Fourier transform is squared and then smoothed over 61 neighbouring values by modified Parzen weights (Cleveland and Parzen, 1975). Every eighth point is plotted.
3. The power spectrum resulting from robust filtering as described in Section 5 (always shown as a solid line): The influence function used here in (5.2) was motivated by the extreme value distribution and has the redescending form

\[
\psi_{ev}(x) = x \exp \left[ -\exp \{q(\|x\| - q)\} \right].
\]

(7.1)

This function represents a middle ground between \( \psi_{BS} \) (which it resembles for \( q \) around 1-75) and the hard rejection rule (which it approaches for \( q > 5 \)). It stays close to the line
Robust Estimation of Power Spectra

The power spectrum from the robust autoregressive approach (always shown as a solid line):
The GM-estimates of the autoregressive coefficients $\Phi$ are computed as described in Section 6. The robust estimates $\Phi$ are inserted into the robust filter (5.3) using $\psi_{en}$ with the same values of $q$ as in (3) and an estimate of the innovations scale $s$ from (6.2); $S_\lambda(f)$ is computed the same way as in the periodogram-based method. Experience shows that the final spectrum is not too dependent on the order of the autoregressive fit; since we have not yet been able to find a very reliable robust stopping rule, we always picked order 7. This is a compromise between a too low order, which might not represent an adequate dynamic range for some data sets of interest, and a too high order which could overfit and would lower the breakdown point of the GM-estimation procedure. Order 7 followed by robust filtering proved to be adequate even for spectra where max $S(f)/\min S(f)$ was as large as $10^8$; for most cases orders $2-3$ would suffice.

Example 1
This is an AO-like artificial example composed of the following three autoregressive processes:

\begin{align*}
  u_k &= 0.975 u_{k-1} + e_k, \\
  w_k &= 0.95 w_{k-1} - 0.9 w_{k-2} + \eta_k, \\
  z_k &= 0.33 z_{k-1} - 0.9 z_{k-2} + \zeta_k
\end{align*}

with $e_k, \eta_k, \zeta_k \sim N(0,1)$. $u_k$ denotes a process with a large amount of power at very low frequencies, $w_k$ constitutes a narrow band component with a peak at 0.17 and a bandwidth of 0.1, and $z_k$ has a peak at 0.22 with a bandwidth of 0.1. We standardize $u_k, w_k$ and $z_k$ to unit variance, compute

\[ y_k = \sqrt{(75)} u_k + w_k + z_k, \quad k = 1, 2, \ldots, n, \]

and then standardize $y_k$ to unit variance. Fig. 1A shows the first 512 points of a realization of length 1,024 of $y_k$ while Fig. 1B shows the first 512 points of the same realization, but with noise from $0.9 \delta_0 + 0.1 N(0, 10)$. Fig. 1C shows the spectra (on a logarithmic scale) of the theoretical spectrum (dotted line), the periodogram-based spectrum and the spectrum resulting from the robust filtering approach. Fig. 1D shows the same spectra with the robust filtering spectrum replaced by the robust autoregressive spectrum. Both robust spectrum estimates follow the theoretical spectrum reasonably well while the periodogram-based spectrum does not bear any resemblance to the true spectrum. Fig. 1E shows the spectrum based on the transfer function of the least squares autoregressive model of order 7 (dashed line), the spectrum based on the robust autoregressive fit of order 7 (solid line) and the theoretical spectrum (dotted line). Here, as in many other instances, the autoregressive spectrum is not able to reproduce the shape of the theoretical spectrum and one would have to go to much higher orders of autoregression (which would lower the breakdown point of the GM-estimates considerably). Note, however, the large gain in range of the robust spectrum over the least squares method.

Example 2
This synthetic example is taken from Brillinger (1973) and consists of 512 values of white noise with five adjacent outliers close to the centre of the data set. Fig. 2A shows the data in
Fig. 1A. Sample path of AR(1) + 2AR(2) (no noise added).

Fig. 1B. Sample path of AR(1) + 2AR(2) + added noise.

Fig. 1C. AR(1) + 2AR(2) + added noise. Periodogram-based spectrum (dashed), robust filtering spectrum (solid) and theoretical spectrum (dotted).

Flo. 1A. Sample path of AR(1) + 2AR(2) (no noise added).

Flo. 1B. Sample path of AR(1) + 2AR(2) + added noise.
Fig. 1D. AR(1)+2AR(2)+added noise. Periodogram-based spectrum (dashed), robust autoregressive spectrum (solid) and theoretical spectrum (dotted).

Fig. 1E. AR(1)+2AR(2)+ added noise. Spectra from transfer functions of autoregressive model of order 7: least squares (dashed) and robust (solid). The theoretical spectrum is shown as a dotted line.
the time domain. Fig. 2B shows the theoretical spectrum without the outliers (dotted), the periodogram-based spectrum and the robust filter spectrum. Only one iteration was used here in the robust filtering approach, because the stopping rule chose an autoregressive representation of order 0. In Fig. 2C the robust autoregressive spectrum replaces the robust filter spectrum. The periodogram-based spectrum is heavily influenced by the five outliers, while both robust techniques produce more or less flat (i.e. white) spectra, as does Brillinger's technique. However, while his technique would break down if, for instance, confronted with two sets of outliers not close to each other, the techniques described here would still work in that case.

Example 3

This geophysical data set consists of 900 adjusted voltage measurements taken every 2 sec from a relative ionosphere opacity meter stationed at the Siple observation post in Antarctica. This data set not only has a relatively large number of isolated outliers (due to electrical discharges caused by snow blowing across the antenna), but also a very large transient signal resulting from calibration. Fig. 3A shows the 900 data points in the time domain; the first outlier is not shown completely (it runs as low as -300 over two data points); the other outliers are around -250. The transient signal due to calibration creates enormous outliers; note the break in the axis. Fig. 3B shows the periodogram-based spectrum and the robust filtering spectrum; Fig. 3C the periodogram-based spectrum and the robust autoregressive spectrum. While the periodogram-based spectrum shows nothing but the effects of the calibration period, both robust techniques show several distinctive peaks at frequencies 0·07, 0·18, 0·20, 0·24 and 0·29. These peaks could be due to periodic modulation of the ionospheric conductivity, which could be caused by hydromagnetic waves and particle precipitation from the magnetosphere into the ionosphere (Lanzerotti et al., 1977).

Example 4

This telecommunications example consists of 1,000 measurements of diameter distortions of a dielectric-lined waveguide, a copper-plated circular steel tube capable of carrying over 200,000 simultaneous telephone conversations. Here relatively minor outliers, barely detectable on a plot of the original data, hide important features of a spectrum with a wide dynamic range. Fig. 4A shows the data in the time domain, with part of the curve enlarged. The two outliers are close together shortly after 2 m and are both due to a malfunctioning of the recording instrument. These outliers are extremely small on the scale of the original
Fig. 2B. White noise with five outliers. Periodogram-based spectrum (dashed), robust filtering spectrum (solid) and theoretical spectrum (dotted).

Fig. 2C. White noise with five outliers. Periodogram-based spectrum (dashed), robust autoregressive spectrum (solid) and theoretical spectrum (dotted).
process; however, since the ratio of the prediction variance/process variance for the auto-
regressive fit of order 7 is \( \approx 10^{-6} \), they are very large on the scale of the innovations and
distort the power spectrum accordingly. Fig. 4B shows first differences of the data; here the
outliers do not appear to be negligible. Fig. 4C shows the periodogram-based spectrum and
the spectrum based on robust filtering. Only one iteration was used because, due to round-off
errors on our computer (a Honeywell 6078 with 36 bits per word), further iteration would not
have yielded significantly better results. Fig. 4D shows the periodogram-based spectrum and

![Time series graph](image)

**Fig. 3A.** Relative ionosphere opacity measurements.

![Log spectrum graph](image)

**Fig. 3B.** Relative ionosphere opacity measurements. Periodogram-based spectrum (dashed) and robust
filtering spectrum (solid).

![Log spectrum graph](image)

**Fig. 3C.** Relative ionosphere opacity measurements. Periodogram-based spectrum (dashed) and robust
autoregressive spectrum (solid).
FlO. 4B. First differences of waveguide data.

FlO. 4C. Diameter distortions of dielectric lined waveguide. Periodogram-based spectrum (dashed) and robust filtering spectrum (solid).

FlO. 4A. Diameter distortions of dielectric lined waveguide.

FlO. 4B. First differences of waveguide data.
the robust autoregressive spectrum. Above frequency 0·1, the periodogram-based spectrum just reflects the effects of the two outliers, while the two robust spectra exhibit a much wider range and show several distinctive peaks. For example, the peak at 0·074 is twice the basic frequency of the straightener applied to the steel tubes housing the waveguide (see Boyd and Thomson, 1976); the peak at 0·136 is twice the basic frequency of the extruder from which the plastic interior lining of the waveguide flows (the basic frequencies of both the straightener and the extruder do generally not manifest themselves in the spectrum of diameter distortions, but could be seen in the spectrum of the curvature). The relatively minor peak at 0·111 is at three times the basic frequency of the straightener, the large peak at 0·185 at five times, while the broad hump around 0·265 consists of contributions at seven times the basic frequency of the straightener and at four times the basic frequency of the extruder. The peak at 0·445 is at twelve times the straightener's fundamental frequency.

8. SUMMARY AND CONCLUSIONS

Two classes of robust estimates of the power spectrum density have been introduced: a robust filtering procedure and a robust autoregressive approach. Each method provides solutions of an intuitively appealing minimization problem, where these solutions are exact in the autoregressive case, approximate for robust filtering. Both methods make use of robust prewhitening and weighting of prediction residuals in an essential way. The procedures have proved to be successful over a wide variety of data sets.

The robust filtering approach and the robust autoregressive procedure only differ in the way "good" autoregressive coefficients for prewhitening are computed: robust filtering
computes its filtered values (and their autoregressive representation) iteratively, while the autoregressive technique gets the autoregressive coefficients by iteratively weighted least squares. For small \( p \) both procedures use roughly the same amount of core and computer time and give equally good results.

Both estimates have rather obvious and appealing resistance properties as evidenced by the examples in Section 7. Their robustness properties have, however, not yet been evaluated in detail. This will be a challenging task for several reasons. First, the results of Hampel (1971) on qualitative robustness do not apply to the general time series case and new theory in this area will be needed. Second, although finite sample mean-squared-error efficiency might be investigated by Monte Carlo techniques, computing asymptotic efficiencies requires knowledge of the asymptotic distributions of our hybrid estimates, which remain to be determined. It might be worth noting that the GM-estimates \( \hat{\phi}_{GM} \) have a positive breakdown point and a purely asymptotic form of qualitative robustness (Martin, 1979). We also believe this to be true for the robust filtering estimates \( \hat{\phi} \) and \( \{ \hat{x}_k \} \), but have not yet given a formal proof. Thus it is expected that the resulting robust prewhitened estimates have positive breakdown points and at least an asymptotic form of qualitative robustness.

There is of course always the question if simpler and faster methods could not do the job just as well. Earlier (Kleiner et al., 1976) we tried to use non-linear smoothers such as local M-estimates to get rid of outliers before computing the spectrum; this failed because these smoothers usually not only get rid of the outliers but of interesting higher frequencies as well. As mentioned previously, there is generally no need for sophisticated robust techniques if the spectrum is reasonably flat. However, robust techniques may be needed in lightly coloured situations and are definitely a necessity if the spectrum has a dynamic range of more than three or four decades. Such spectra are commonplace in physical sciences such as telecommunications (where wide range spectra arise, for example, in transmitting signals by cable, waveguide or optical fibres) and geophysics (e.g. in measuring phenomena of the magnetosphere); they also occur in the fields of medicine, where Dumermuth (1971) showed spectra of electroencephalograms ranging over six to seven decades, and even economics, where, for instance, the spectrum of the Federal Reserve Board Index of Industrial Production ranges over five decades (Nerlove, 1964).

**ACKNOWLEDGEMENT**

This research was supported in part by NSF grant ENG–76–00504.

**REFERENCES**

Akaike, H. (1969). Fitting autoregressive models for predictions. *Ann. Inst. Statist. Math.*, 21, 243–247.

—— (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control*, AC-19, 716–722.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972). *Robust Estimates of Location*. Princeton: Princeton University Press.

Blackman, R. B. and Tukey, J. W. (1958). *The Measurement of Power Spectra*. New York: Dover.

Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction*. New York: Wiley.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, 40, 318–335.

Box, G. E. P. and Jenkins, G. M. (1970). *Times Series Analysis: Forecasting and Control*. San Francisco: Holden Day.

Boyd, R. G., Jr. and Thomson, D. J. (1976). Geometric requirements for fabrication of WT4 waveguide quality tubing. *IEE Proceedings of Conference on mm Waveguide Systems*, pp. 26–29.

Brillinger, D. R. (1973). A power spectral estimate which is insensitive to transients. *Technometrics*, 15, 559–562.

Cleveland, W. S. and Parzen, E. (1975). The estimation of coherence, frequency response and envelope delay. *Technometrics*, 17, 167–172.

Daniel, C. and Wood, F. S. (1971). *Fitting Equations to Data*. New York: Wiley.

Denby, L. and Larsen, W. (1977). Robust regression estimators compared via Monte Carlo. *Comm. in Statist.*, A6, 355–362.
DENBY, L. and MARTIN, R. D. (1979). Robust estimation of the first-order autoregressive parameter. J. Amer. Statist. Ass., 74, 140-146.

DOOR, J. L. (1953). Stochastic Processes. New York: Wiley.

DUMERMUTH, G. (1971). Electronic data processing in pediatric EEG research. Neuropädiatrie, 4, 349-374.

FOX, A. J. (1972). Outliers in time series. J. R. Statist. Soc. B, 34, 350-363.

HAMPEL, F. (1968). Contributions to the theory of robust estimation. Ph.D. dissertation, Dept. of Statist., University of California, Berkeley.

--- (1971). A general qualitative definition of robustness. Ann. Math. Statist., 42, 1887-1896.

--- (1974). The influence curve and its role in robust estimation. J. Amer. Statist. Ass., 69, 383-393.

HILL, R. (1977). Robust regression when there are outliers in the carriers. Ph.D. dissertation, Dept. of Statistics, Harvard University.

HOGG, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. J. Amer. Statist. Ass., 69, 909-923.

HUBER, P. J. (1964). Robust estimation of a location parameter. Ann. Math. Statist., 35, 73-101.

--- (1973). Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Statist., 1, 799-821.

--- (1977). Robust Statistical Procedures. Philadelphia: Society for Industrial and Applied Mathematics.

KAZAKOS, P. and GRAY, R. M. (1979). Robustness of estimators on stationary observations. Annals Prob. (to appear).

KLEINER, B., MARTIN, R. D. and THOMSON, D. J. (1976). Three approaches towards making power spectra less vulnerable to outliers. Proc. Business and Economic Statistics Section of the Amer. Statist. Ass., 386-391.

LANZEROTTI, L. J., MACLENNAN, C. G., EVANS, C. (1977). Measurements of magnetic fluctuations and ionosphere continuity changes at Siple Station. Antarctic Journal of the U.S., 12, 186-189.

MALLOWS, C. L. (1975). On some topics in robustness. Bell Laboratories Technical Memorandum.

MARTIN, R. D. (1979). Robust estimation for time series. In Robustness in Statistics (R. L. Launer and G. N. Wilkinson, eds). New York: Academic Press.

MARTIN, R. D. and ZEH, J. (1978). Robust generalized M-estimates for autoregressive parameters: small sample behaviour and applications. Technical Report 214, Dept. of Electrical Engineering, University of Washington.

MASRELIEZ, C. J. and MARTIN, R. D. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. IEEE Trans. on Autom. Control, AC-20, 107-110.

NERLOVE, M. (1964). Spectral analysis of seasonal adjustment procedures. Econometrica, 32, 241-286.

PARZEN, E. (1974). Some recent advances in time series modelling. IEEE Transactions on Autom. Control, AC-19, 723-730.

THOMSON, D. J. (1977). Spectrum estimation techniques for characterization and development of WT4 waveguide. Bell System Tech. J., 56, 1769-1815, 1983-2005.

TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In: Contributions to Probability and Statistics (I. Olkin, ed.). Stanford, Ca: Stanford University Press.

--- (1967). An introduction to the calculations of numerical spectrum analysis. In: Proceedings of the Advanced Seminar on Spectral Analysis of Time Series (B. Harris, ed.), pp. 25-46. New York: Wiley.

--- (1975). Usable resistant/robust techniques of analysis. In: Proc. First ERDA Symposium, Los Alamos, New Mexico, pp. 1-31.

WELCH, P. D. (1967). The use of FFT for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio and Electroacoustics, AU-15, 70-73.

WHITTLE, P. (1962). Gaussian estimation in stationary time series. Bull. Int. Stat. Inst., 39, 105-129.

WOLD, H. (1954). A Study in the Analysis of Stationary Time Series. Stockholm: Almqvist & Wiksell.

YOHAI, V. J. and MARONNA, R. A. (1977). Asymptotic behavior of least-squares estimates for autoregressive processes with infinite variances. Ann. Statist., 5, 554-560.

APPENDIX A

We give a simple perturbation argument which, for the uncontaminated case, indicates that the bias in the spectrum introduced by using the robust filtering procedure is small and is proportional to the true spectrum.

Let

\[ x_k = x_k + \delta_k, \quad \psi(t) = t + \lambda(t) \]  

(A.1)

and assume that the original process \( x_k \) can be represented by a finite-order autoregressive process

\[ x_k = \sum_{j=1}^{p} \phi_j x_{k-j} + \epsilon_k. \]  

(A.2)
Discussion of the Paper by Drs Kleiner, Martin and Thomson

With this notation the autocovariances \( \{ \hat{\rho} \} \) from which we can compute the spectrum \( \hat{S}(f) \) of \( \hat{x}_k \) are of the form

\[
\hat{\rho}_j = \rho_j + 2Ex_k+i\delta_k + E\delta_{k+j}\delta_k.
\]  

(A.3)

If the scale of the process \( x_k \) is much larger than that of the errors, the dominant error term in (A.3) is \( 2Ex_k+i\delta_k \).

The most useful measure for characterization of a spectrum density estimate is the relative error \( E\hat{S}(f)/S(f) \); this can be obtained not only from the process \( \{ x \} \) and \( \{ \hat{x} \} \), but also by prewhitening both the data \( \{ x \} \) and the error process \( \{ \delta \} \) using the autoregressive coefficients \( \phi_1, \phi_2, \ldots, \phi_p \). Prewhtening \( \{ x \} \) in this way results in \( \{ \hat{x} \} \), while for the error sequence one obtains

\[
\eta_k = \delta_k - \sum_{j=1}^{P} \phi_j \delta_{k-j}.
\]  

(A.4)

Before computing \( E\hat{\varepsilon}_{k+j}\eta_k \), we note that from (A.1) and (5.2) we obtain for \( \nu_k \equiv 0 \)

\[
\delta_k = cs\lambda \left( x_k - \hat{x}_k, \frac{T}{cs} \phi \right)
\]  

(A.5)

and from (A.2)

\[
\delta_k = cs\lambda \left[ \frac{1}{cs} \left( \varepsilon_k + \sum_{j=1}^{P} (x_{k-j} \phi_j - \hat{x}_{k-j} \hat{\phi}_j) \right) \right]
\]  

(A.6)

if we assume our estimate of the autoregressive order to be correct.

Since no contamination, i.e. \( \nu_k \equiv 0 \), assumed, the sum in (A.6) will be rather small and a Taylor series expansion around \( \varepsilon_k/cs \) gives

\[
\delta_k \approx cs\lambda \left( \frac{\varepsilon_k}{cs} \right) + cs\lambda \left( \frac{\varepsilon_k}{cs} \right) \sum_{k=1}^{P} (x_{k-j} \phi_j - \hat{x}_{k-j} \hat{\phi}_j).
\]  

(A.7)

\( E\varepsilon_{k+j}\eta_k \) can now be computed easily, since \( \{ \varepsilon \} \) is an i.i.d. process, \( \delta_k \) (and therefore \( \eta_k \)) does not contain any function of \( \varepsilon \) after time \( k \) and since \( E\varepsilon_k \hat{x}_{k-j} = 0 \), we have

\[
E\varepsilon_{k+j}\eta_k = 0 \quad \text{for all } j > 0, \quad E\varepsilon_k \eta_k = csE \left( \varepsilon_k \lambda \left( \frac{\varepsilon_k}{cs} \right) \right).
\]  

(A.8)

Hence the relative bias of the spectrum is nearly constant:

\[
\frac{\text{Bias } \hat{S}(f)}{S(f)} = 2csE \left( \varepsilon_k \lambda \left( \frac{\varepsilon_k}{cs} \right) \right).
\]  

(A.9)

DISCUSSION OF THE PAPER BY DRs KLEINER, MARTIN AND THOMSON

Professor J. Durbin (London School of Economics): I am happy to welcome this excellent paper from several points of view. Taking first the specific problem of spectrum estimation with which the authors have been concerned, they have demonstrated convincingly that contamination due to outliers in the time domain can cause substantial bias in any frequency domain analysis throughout the whole range of the spectrum. The examples the authors have produced are very dramatic and convincing. The modification of standard robustness procedures that they have developed and applied to this problem seem to me to have dealt with it satisfactorily. I can claim no great expertise or experience in spectral analysis, but it seems to me that their basic technique is the right way to estimate the spectral density. The authors have been able to draw on the extensive experience of Bell Telephone Laboratories, both in spectral estimation and in robustness. They have gone a long way towards producing a definitive treatment of this problem.