ROBUST METHODS FOR ARIMA MODELS

BY

R. DOUGLAS MARTIN, ALEXANDER SAMAROV,
AND WALTER VANDAELE

TECHNICAL REPORT NO. 21
MARCH 1982

THIS RESEARCH IS PARTIALLY SUPPORTED
BY NATIONAL SCIENCE FOUNDATION
GRANT NO. SOC78-09474

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON
98195
ROBUST METHODS FOR ARIMA MODELSa

by

R. Douglas Martin1, Alexander Samarov2,

Walter Vandaele3

1. Introduction

"As everybody knows, Box and Jenkins have revolutionized the world of forecasting, but all the Box-Jenkins packages known to me make no provisions for outliers and contamination of various kinds"

J. Durbin (1979, p. 339)

ARIMA models have worked successfully for analyzing and forecasting many time series. However, there have also been frequently occurring cases where problems in model building were encountered because the data contained atypical observations, or outliers, and/or missing values. For example, outliers may result because of special promotions, strikes, or recording errors. Missing values result because certain data were lost or never recorded. In this paper we will primarily analyze outlier problems. However, as shown in Section II, the methodology can easily handle missing data problems.

The robustness point of view adopted here is similar to that of Huber (1977). The nominal statistical model is assumed to be a parametric Gaussian

---

aFor presentation at the ASA-CENSUS-NBER Conference on Applied Time Series Analysis of Economic Data, Washington, D.C., October 13-14, 1981. We thank H. Roberts, T.W. Anderson, and many other participants of the Conference for their stimulating discussions.

1University of Washington, Department of Statistics.

2M.I.T., Sloan School of Management, Center for Computational Research in Economics and Management Science.

3M.I.T., Sloan School of Management, Center for Computational Research in Economics and Management Science and Harvard University, Department of Economics. Research supported by NSF SOC78-09474, DOE EX-76-A-01-2295, and NSF MCS-7726902. We wish to thank B. Dunsmuir, J. Durbin, E. Kuh, M. Watson, and R. Welsch for helpful discussions. Many helpful comments were also received when an earlier version of this paper was presented at the Conference on Model Selection, University of Florida, Gainesville, Florida, April 1980. We also thank M. DuPree for editorial help and K. Glennon for typing this paper.
ARIMA model, i.e., the error term in the model is assumed normally and independently distributed with mean zero and constant variance. However, reality dictates that one allow for the possibility of "small" non-Gaussian deviations from the nominal model which result in time series realizations containing outliers.

The presence of outliers in a time series model violates the basic homogeneity assumption of the data. The existence of even one or two extreme outliers can: (1) seriously distort the sample autocorrelation and partial autocorrelation function leading towards incorrect model identification, (2) substantially affect the parameter estimates since an extreme outlier will have a relatively large contribution towards the residual sum of squares, and (3) seriously impede the construction of forecasts based upon these historical data.

In practice outliers have often been adjusted using such ad hoc techniques as eyeballing, or curve-fitting by using the same months in other years.\(^1\) Although such techniques may be suitable in some situations, there are many cases in which they are not, as they require a priori designations of which points are outliers, which is not always easy. Also curve fitting methods may not fully exploit the information in the data, e.g., they may neglect the correlations with neighboring points. Therefore, we need a practical statistical method which more fully exploits the historical data contents in dealing with outliers.

The objectives of this paper are:

1. To construct robust/resistant ARIMA parameter estimates and forecasts which provide insurance against undue reliance on small segments of the data, and, as such, protection against certain model failures.

\(^1\)An exception is the work of Brubacher (1974). However, Brubacher's method still requires a priori specification of which data points to treat as outliers.
2. To use these robust estimates to detect outlier points conditional on the model contemplated. This also may lead to a reevaluation of the model.

This paper is arranged as follows. Section 2 describes the additive outlier model; Section 3 outlines our basic approach; Section 4 formulates the robust maximum likelihood type estimates; Section 5 presents the state variable representation and robust filter technique; Section 6 describes an iterative robust estimation procedure for minimization of the robustified objective function under certain simplifying assumptions; Section 7 discusses problems of starting values; Section 8 presents the model identification and selection; Section 9 shows how to generate robust forecasts; Section 10 presents some empirical results and in Section 11, we outline some topics for further research.
2. The Additive Outlier Model

Research on robust estimation in the time series context has lagged behind studies in robust location and robust regression estimation, and understandably so, in view of the increased difficulties imposed by dependencies between the data points.

There are few published works on outliers in time series. Fox (1972) basically defines two types of outliers which might occur in time series data and considers a special characterization of outliers in time series. In the context of a $p$-th order autoregressive model he derives a likelihood criterion that a particular data point is an outlier. See also Barnett and Lewis (1978, Chapter 7) for a discussion of the work of Fox. More recent work on outliers in time series include Tiao and Guttman (1975), Martin and Jong (1976), Martin and Zeh (1978), Abraham and Box (1979), Martin (1979a), Denby and Martin (1979), and Martin (1980b), and de Alba and Van Ryzin (1980).

In the face of the difficulties imposed by the dependencies between the data points, it seems imperative to begin by specifying simple outlier generating models. This is a task made difficult by the wide variety of time series data types. Using Fox's (1972) terminology, one can distinguish two particular outlier types in time series.

One possibility is that of isolated gross-error outliers which might, e.g., be due to various recording errors, i.e., errors superimposed on an otherwise reasonable realization of the time series process. These outliers could be completely isolated data points or could exhibit a particular local behavior not related to the behavior of the remainder of the series. Another possibility is that the character of the outlier is consistent with the sample path by virtue of the innovations or white noise series having a heavy tailed distribution.
While admitting the limitations of the above categorization, it seems quite desirable to capture some of the essence of the above kinds of behavior with the following two outlier models.

The first kind of gross-error outliers in time series models can be represented with the following additive outliers (AO) model. Let \( x_t \) be represented by an autoregressive-integrated moving average (ARIMA) model

\[
\phi(B)w_t = \mu + \theta(B)\varepsilon_t \quad t = 1, 2, \ldots, n
\]

(2.1)

with

\[
w_t = (1 - B)^d x_t
\]

where \( d \) is the order of consecutive differencing and \( \phi(B) \) and \( \theta(B) \) are polynomials in the backward shift operator \( B \) of degrees \( p \) and \( q \) respectively. In (2.1) \( \mu \) represents the mean of the stationary process \( w \). We subsequently refer to the model (2.1) as an ARIMA\((p, d, q)\) process.\(^1\) The error term \( \varepsilon_t \) in (2.1) is normally distributed with mean zero, \( E(\varepsilon_t) = 0 \), constant variance, \( E(\varepsilon_t^2) = \sigma^2 \), and zero correlation over time, \( E(\varepsilon_t \varepsilon_{t'}) = 0 \), \( t \neq t' \).

The sign convention of the ARIMA\((p,d,q)\) polynomials in (2.1) is specified as in the following two polynomials

\[
\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p
\]

\[
\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q.
\]

\(^1\)For expositional purposes we assume that there are no seasonal parameters or seasonal differencing in the model (2.1). Quite often economic data are nonstationary and contain seasonal parameters. However, the methodology presented below is applicable to situations including differencing and seasonality. In that case model (2.1) would be written as

\[
\phi(B)\phi(B^s)w_t = \mu + \theta(B)\theta(B^s)\varepsilon
\]

with

\[
w_t = (1 - B)^d (1 - B^s)^D x_t
\]

where \( s \) is the seasonal periodicity, \( D \) is the order of seasonal differencing, and \( \phi(B^s) \) and \( \theta(B^s) \) are polynomials in \( B^s \) representing the seasonal autoregressive and moving average parameters, respectively.
It is also assumed that all the roots of \( \phi(B) = 0 \) and of \( \theta(B) = 0 \) lie outside the unit circle. Under the gross-error AO model, (2.1) represents the 'core' of the model, by which we mean that for most of the times \( t \) we will be able to observe \( x_t \) exactly. However, in a few instances \( x_t \) will be contaminated with gross-errors denoted by \( v_t \). Therefore we represent the measurement equation as

\[
y_t = x_t + v_t.
\]  

(2.2)

The additive outliers model specifically assumes that \( v_t \) represents gross-error situations with the fraction of additive outliers, \( \gamma \), positive and small i.e., \( \text{Prob}(v_t = 0) = 1 - \gamma \). Typically \( \gamma \) has often been observed to be in the range .01 to .2. Dependency structure for the \( v_t \) could be specified in order to model patchy outliers, and i.i.d. \( v_t \) could be used to model outliers which are typically isolated.

The second model, the innovation outlier (IO) model, is obtained when \( v_t = 0 \) for all \( t \), so that the \( x_t \) process is observed perfectly, but with \( \epsilon_t \) having a heavy-tailed nonnormal distribution.

In this paper we will concentrate on the AO model, although the method proposed can also be used as a diagnostic tool for IO situations. It is indeed known that the least squares estimates are less affected by innovation outliers than by additive outliers. For innovation outliers the least squares estimates are consistent, but not asymptotically efficient, see Whittle (1953), Martin (1979a). Additive outliers can cause more serious problems as the least squares estimate is not even consistent and a small percentage of additive outliers can create large biases.
3. The Basic Approach

The basic procedure fits in the framework of the following general robustifying process (see Huber, 1979, 1981). Starting with initial parameter estimates:

a) "Clean" the data using a procedure which pulls the outliers toward their fitted values. This is accomplished via a robust filtering algorithm.

b) Fit an ARIMA nonlinear least squares algorithm to the cleaned data to obtain new parameter estimates.

c) Iterate a) and b) until convergence.

The robust filtering algorithm is of the form

\[ \hat{x}_t = \hat{x}_{t|t-1} + s_t \psi \left( \frac{y_t - \hat{y}_{t|t-1}}{s_t} \right) \]  

(3.1)

where \( \hat{x}_{t|t-1} \) and \( \hat{y}_{t|t-1} \) are the one-step-ahead (robust) predictors using data up to and including period \( t - 1 \) of \( x_t \) and \( y_t \), respectively, and in fact for our model, equations (2.1) and (2.2) yield \( \hat{x}_{t|t-1} = \hat{y}_{t|t-1} \); \( s_t \) is a data-dependent scale measure for the residuals \( y_t - \hat{y}_{t|t-1} \). Details are given in Sections 4, 5, and 6.

Two psi-functions which are (among others) likely candidates for use here are the hard rejection function

\[ \psi_{HR}(r) = \begin{cases} 
    r & |r| \leq b \\
    0 & |r| > b 
\end{cases} \]  

(3.2)

and the Hampel (1974) two-part redescending function

\[ \psi_{HA}(r) = \begin{cases} 
    r & |r| \leq a \\
    \frac{a}{b-a} (b-r) & a < r \leq b \\
    -\frac{a}{b-a} (b+r) & -b \leq r < -a \\
    0 & |r| > b 
\end{cases} \]  

(3.3)

Figures 1 and 2 show these functions.
FIGURE 1. Hard Rejection Function

An intuitively appealing feature of both psi-functions is that

\[ \psi\left(\frac{u_t}{s_t}\right) = 0 \text{ when } |u_t| > b \cdot s_t, \] and so \( \hat{x}_t = \hat{x}_{t-1} \) when \( |u_t| \) is sufficiently large. This rejection feature seems quite natural and so we use \( \psi_{HA} \) in our algorithm. One reason for not using the hard rejection function \( \psi_{HR} \), obtained as a limiting case of \( \psi_{HA} \) with \( a = b \), is that \( \psi \) should be continuous, see Hampel (1974).\(^1\)

In the next three sections we will give detailed rationale for the proposed robust filtering algorithm including a maximum likelihood-type justification. As explained in Section 5, our algorithm will be called the approximate conditional mean (ACM) filter. Those more interested in empirical results could, at the first reading of the paper, go directly to Section 10 containing two examples. Preceding Section 10 there are three sections dealing with model identification, initial parameter guess values, and forecasting. These three topics form the basis for ongoing research.

---

\(^1\) For the case where the measurement error \( v_t \) has a contaminated normal distribution defined in equation (4.6), Martin (1979b) provides additional arguments for using Hampel's psi-function \( \psi_{HA} \) (see also our discussion at the end of Section 5). Although Martin (1980b) suggested choosing \( a \) and \( b \) in the vicinity of \( a = 2.6 \) and \( b = 3.6 \), these values remain to be better established. It is now felt that the value \( b = 3.6 \) is perhaps too small relative to \( a = 2.6 \). The reason is that for these values the derivative \( \psi' \) appearing in (5.15) takes on rather an undesirable large value. In fact, an appeal to the principle of qualitative robustness suggests using a more smooth \( \psi \) function than \( \psi_{HA} \), namely a \( \psi \) function with continuous derivatives. These issues remain to be explored more carefully, perhaps by Monte Carlo.
4. Robust Maximum Likelihood-type Estimates

Let the parameter vector $\pi' = (\theta', \sigma', \sigma_e^2)$ represent the parameters of the AO-model. Its log-likelihood can be written as

$$\log h(\gamma^n | \pi) = \sum_{t=1}^n \log h(y_t | \gamma^{t-1}, \pi)$$ (4.1)

where $\gamma^t = (y_1, y_2, \ldots, y_t)'$ is the vector of observations up to and including observation $y_t$. The density $h(y_t | \gamma^{t-1}, \pi)$ is the conditional density of the observation $y_t$ given $\gamma^{t-1}$, and $h(y_1 | \gamma^0, \pi)$ denotes the unconditional density $h(y_1 | \pi)$. The vector $\gamma^0$ stands for the initial observation $(y_0, y_{-1}, \ldots, y_{1-p})$.

Since $x_t$ and $v_t$ are by assumption independent, we can write

$$h(y_t | \gamma^{t-1}, \pi) = f_x(y_t - \xi | \gamma^{t-1}, \pi) dF_v(\xi)$$ (4.2)

where $F_v$ is the distribution function of the measurement error $v_t$ and $f_x$ is the conditional prediction density function of $x_t$ given $\gamma^{t-1}$. In engineering literature $f_x$ is sometimes called the "signal" or "state" prediction density.

Let

$$a_t | t-1 = E(x_t | \gamma^{t-1}) = E(y_t | \gamma^{t-1})$$ (4.3a)

and

$$m_t = E [(x_t - \tilde{x}_t | t-1)^2 | \gamma^{t-1}]$$ (4.3b)

denote the conditional prediction mean and conditional prediction variance associated with $f_x$. We make the simplifying assumption that $f_x$ may be well-approximated by setting

$$f_x(x_t | \gamma^{t-1}, \pi) = \frac{1}{\sqrt{m_t}} f \left( \frac{x_t - \tilde{x}_t | t-1}{\sqrt{m_t}} \right)$$ (4.4)

for some fixed density $f$ which is independent of the parameters $\pi$. For $t = 1$, the expectations are taken to be unconditional.

\footnote{From now on we use centered stationary data, see (2.1). When $\mu$ is unknown it may be estimated robustly and the estimate $\hat{\mu}$ can be used to form centered observations. When $\mu$ is consistent, estimators based on the centered data typically behave asymptotically as if $\mu$ were known and exactly centered observations were used.}
Now, using (4.4) we can rewrite (4.2) as

$$h(y_t | y_{t-1}, \pi) = g_t(u_t)$$  \hspace{1cm} (4.5)

where $u_t = y_t - \hat{x}_{t|t-1}$ and the subscript $t$ on the function $g_t$ indicates the dependence of $g_t$ on time.

In practice we very rarely know the noise distribution $F_v$ in the tails with high accuracy. For the contaminated normal noise distribution\(^1\)

$$F_v = (1 - \gamma)N(0, \sigma_o^2) + \gamma N(0, \sigma^2)$$ \hspace{1cm} (4.6)

with $\sigma_o^2 < \sigma^2$ and small $\gamma > 0$, Martin (1979b) gave some detailed motivation for approximating $g_t$ by setting

$$g_t(u_t) = \frac{1}{s_t} g\left(\frac{u_t}{s_t}\right)$$ \hspace{1cm} (4.7)

where the density $g$ is obtained by convolution

$$g = f * F_v.$$ \hspace{1cm} (4.8)

Although the functional forms assumed in (4.4) and (4.7) are generally not valid in situations with non-Gaussian $v_t$, we believe that the use of these forms when $F_v$ is nearly normal involves an approximation error that is small enough to be relatively inconsequential.

Note that the scale measure $s_t$ represents the scale of the $y$-prediction residuals $u_t = y_t - \hat{x}_{t|t-1}$. Since the $x$-prediction residuals $x_t - \hat{x}_{t|t-1}$ have as scale measures the quantity $\sqrt{m_t}$, and since $y_t = x_t + v_t$, it is reasonable to let

$$s_t = \sqrt{m_t + \sigma_o^2}$$ \hspace{1cm} (4.9)

when $F_v$ at most deviates from a nominal $N(0, \sigma_o^2)$ distribution because of its heavy tail, e.g., as in (4.6). Of course when the errors $v_t$ are zero most of the time,

\(^1\)See Huber (1977a, p. 2) for a similar suggestion to model the error law of "good data" with a contaminated normal distribution.
say \( P(y_t = 0) = 1 - \gamma \) with \( \gamma \) small, obtained for example by setting \( \sigma^2_0 = 0 \) in (4.6), then we have

\[
s_t = \sqrt{m_t}.
\]  

(4.10)

Using (4.5) and (4.7) we can rewrite (4.1) as

\[
\log h(y^n | \pi) = - \sum_{t=1}^{n} \log s_t + \sum_{t=1}^{n} \log g \left( \frac{u_t}{s_t} \right).
\]  

(4.11)

Now, it seems natural by analogy with Huber's (1964, 1977a) M-estimates (maximum likelihood type estimates) to replace \( -\log g \) with a properly chosen symmetric function \( \rho \). Thus, we propose to define the parameter estimates as the values that minimize the following robustified objective function:

\[
R(\pi) = \sum_{t=1}^{n} \log s_t(\pi) + \sum_{t=1}^{n} \rho \left( \frac{u_t(\pi)}{s_t(\pi)} \right).
\]  

(4.12)

The parameter vector \( \pi \) is included in (4.12) to indicate explicitly the dependency of \( s_t(\pi) \) and \( u_t(\pi) \) on the parameter vector \( \pi' = (\phi', \theta', \sigma^2_\varepsilon) \).

If \( \rho(t) = -\log g(t) \) and the density \( g \) is normal the corresponding function \( \psi \) is the identity function and minimization of \( R(\pi) \) is identical to Gaussian maximum likelihood estimation.

The motivation for the particular choice of the function \( \rho \) is the same as that for Huber M-estimates for location and regression (see e.g., Huber (1977a), Hampel (1974)): \( \rho \) should have bounded and continuous derivative \( \psi(r) = \rho'(r) \). Boundedness insures that no one single observation \( y_t \) can have an unbounded influence on the parameter estimates \( \hat{\pi} \), and continuity insures that small changes in \( y_t \) (e.g., due to rounding) do not produce large changes in \( \hat{\pi} \). In fact we shall use Hampel's \( \psi_{HA} \) given by (3.3) for the example presented in Section 10.
5. State Variable Representations and Robust Filters

5.1 State Variable Representation of the ARMA Model

To determine the parameter estimates which minimize the objective function \( R(\theta) \) defined by (4.12), we need to express \( \hat{x}_t \) and \( s_t \) as functions of \( \phi, \theta, \sigma_e, \) and \( y_t \). This is readily done by casting the AR model (2.1), (2.2) in the following vector state variable representation

\[
\begin{align*}
\alpha_t & = \phi \alpha_{t-1} + \epsilon_t \\
y_t & = z' \alpha_t + \nu_t
\end{align*}
\]

where \( z = (1, 0, \ldots, 0)' \) and the first element of \( \alpha_t \) is \( x_t \). Equation (5.1) is a first order vector autoregressive model, and (5.2) represents the measurement equation associated with (5.1). In (5.1), it is assumed that \( \epsilon_t \) is white noise with \( E(\epsilon_t) = 0 \) and \( \text{Var}(\epsilon_t) = \sigma^2 \). Notice that (5.2) for our model equals \( y_t = x_t + \nu_t \).

Further details are given in Appendix A where it is shown that the matrix \( \phi \) may be chosen to have the following form

\[
\phi = \begin{bmatrix}
\phi_1 \\
\phi_2 \\
\vdots \\
\phi_k
\end{bmatrix} = \begin{bmatrix}
I(k-1) \\
0
\end{bmatrix}
\]

where \( I(k-1) \) is a \((k - 1) \times (k - 1)\) identity matrix, and \( 0 \) is a \((k - 1) \) column vector of zeros. The dimensionality\(^2\) of the square \( \phi \) matrix is \( k = \max \{ p, q + 1 \} \)

\(^1\)For the sake of brevity, this section presents a particular state variable representation for stationary ARMA models. In Appendix A the model is extended to ARIMA models which include differencing.

\(^2\)The dimensionality of the state vector is in general defined as

\[ k = \max(p + sP + d + sD, q + sQ + 1) \]

where \( p \) and \( q \) are the orders of the autoregressive and moving average parameters, and \( P \) and \( Q \) the orders of the seasonal autoregressive and seasonal moving average parameters, respectively, \( s \) is the seasonal periodicity, and \( d \) and \( D \) are the order of the consecutive and seasonal differencing. Because it is quite customary to us \( s \) for the seasonal periodicity, we adhered to this notation, although it could possibly lead to some confusion with the scale of the \( y \)-prediction residuals \( s_t \).
If \( q \geq p \), the first column of \( \Phi \) contains the autoregressive parameters \( \phi_1, \phi_2, \ldots, \phi_k \), but with \( \phi_i = 0 \) for \( i > p \). Corresponding to this choice of \( \Phi \), the vector \( \mathbf{r} \) in (5.1) is a \( k \times 1 \) column vector defined as

\[
(1, -\theta_1, -\theta_2, \ldots, -\theta_{k-1})'
\]

with \( \theta_i = 0 \) for \( i > q \) in case \( p > q \).

5.2 Conditional Mean and Approximate Conditional Mean Filters

The problem of estimating the state vector \( \mathbf{x}_t \) of a stochastic dynamic system of the type (5.1), (5.2) from noisy observations \( y_1, \ldots, y_t \), often called the filtering problem, is of central importance in certain engineering problems (see, e.g., Jazwinski (1970)). It is well known that the conditional mean

\[
a_t = E(\mathbf{x}_t | \mathbf{y}^t)
\]

is the optimal estimator, or filter, of the state vector \( \mathbf{x}_t \) based on a mean-squared-error criterion.

Kalman (1960a, 1960b) introduced a recursive technique for updating the estimator \( a_t \) and the filtering error covariance matrix

\[
P_t = E[(\mathbf{x}_t - a_t)(\mathbf{x}_t - a_t)' | \mathbf{y}^t]
\]

as new observations become available. One of the major advantages of this technique (other than computational efficiency) is that the predictions of the future values of the state vector \( \mathbf{x}_t \), together with their associated covariance matrices \( P_t \) (conditional on the parameter values), may be calculated using only the most recent observation \( y_t \) and the lagged filtered value \( a_{t-1} \). This means, in particular, that forecast values may be obtained by simply extending the recursions; see also Section 9.

When \( \epsilon_t \) and \( \nu_t \) are independently and normally distributed, \( \epsilon_t \sim N(0, \sigma_\epsilon^2) \), and \( \nu_t \sim N(0, \sigma_\nu^2) \), the straightforward computation of \( a_t = E(\mathbf{x}_t | \mathbf{y}^t) \), by any one of a variety of techniques yields the Kalman filter recursions:
\[ a_t = a_{t|t-1} + m_t(y_t - z' a_{t|t-1})/(m_t + \sigma_y^2) \]  \hspace{1cm} (5.6)

and

\[ P_t = M_t + \frac{1}{m_t + \sigma_y^2} m_t m'_t, \]  \hspace{1cm} (5.7)

where

\[ a_{t|t-1} = E(x_t|y^{t-1}) \]  \hspace{1cm} (5.8)

is the optimal one-step-ahead predictor of \( a_t \). The first component \( a_{t|t-1} = \hat{x}_t|t-1 \) \( E(x_t|y^{t-1}) \) of this predictor vector is the best one-step-ahead predictor of both \( x_t \) and \( y_t \), see (4.3a). The state prediction-error covariance matrix \( M_t \) is defined as

\[ M_t = E[(a_t - a_{t|t-1})(a_t - a_{t|t-1})'|y^{t-1}]. \]  \hspace{1cm} (5.9)

The vector \( m_t \) appearing in (5.6) and (5.7) is the first column of \( M_t \), and

\[ m_t = E((x_t - \hat{x}_t|t-1)^2|y^{t-1}) \]  \hspace{1cm} (cf. (4.3b)) is the first element of \( m_t \), or the 1-1 element of \( M_t \). It is a special consequence of the totally Gaussian model that the \( M_t \) and \( P_t \) do not depend upon \( y_t \). The values for \( a_{t|t-1} \) and \( m_t \) are obtained from the values \( a_{t-1} \) and \( P_{t-1} \) by applying the prediction equations:

\[ a_{t|t-1} = \Phi a_{t-1} \]  \hspace{1cm} (5.10)

\[ M_t = \Phi P_{t-1} \Phi' + Q \]  \hspace{1cm} (5.11)

where

\[ Q = \sigma_y^2 R T R'. \]  \hspace{1cm} (5.12)

To start the recursions (5.6), (5.7), (5.10), and (5.11), initial values of \( a_0 \) and \( P_0 \) are needed. The initial values are discussed in Section 7.

It may be noted that for the perfectly observed Gaussian ARMA model \( (\nu_t \equiv 0) \) Harvey and Phillips (1976, 1979) also used a state variable representation and demonstrated various theoretical and computational advantages of using the Kalman filter for obtaining maximum likelihood parameter estimates. Of
course, for the AO model (2.1), (2.2) when $v_t$ is non-zero and non-Gaussian
their approach no longer works. One reason is that the Kalman filter is no
longer the conditional mean, another is that the Kalman filter lacks robustness
towards additive outliers.

Masreliez (1975) developed a non-Gaussian approximate conditional mean
filter which leads to intuitively appealing robust filter recursions. In
obtaining his result, Masreliez used the assumption that the state prediction
density is Gaussian

$$f(\alpha_t | y^{t-1}) = N(a_{t-1}, M_t). \tag{5.13}$$

This assumption means, in particular, that the prediction density of the first
component of $\alpha_t$, i.e., the conditional density of the underlying ARMA process $x_t$
given $y^{t-1}$, is normal, $x_t \sim N(a_{t-1}, m_t)$. This assumption is stronger than
assumption (4.4).

Masreliez' recursions for the conditional mean $a_t = E(\alpha_t | y^t)$ are

$$a_t = a_{t-1} + m_t \psi_t(y_t) \tag{5.14}$$

and

$$P_t = M_t - \psi'_t(y_t) M_t m_t' \tag{5.15}$$

The prediction equations remain the same as (5.10) and (5.11). The function
$\psi_t(y_t)$ equals the negative of the score function (see, e.g., Rao, 1973, p. 366ff),
of the observation prediction density $h(y_t | y^{t-1}, \pi)$, i.e.,

$$\psi_t(y_t) = -(\partial / \partial y_t) \log h(y_t | y^{t-1}, \pi). \tag{5.16}$$

The function $\psi'_t(y_t)$ is defined as

$$\psi'_t(y_t) = \frac{\partial}{\partial y_t} \psi_t(y_t). \tag{5.17}$$

When the measurement errors $v_t$ are independently and normally distrib-
uted, the score function $\psi_t$ is a linear function, and (5.14) and (5.15) are
identical to the classical Kalman recursions (5.6) and (5.7).
5.3 Robust Filter Related to the Minimization Problem

When the measurement errors \( v_t \) are non-Gaussian, use of the assumptions made in Section 4 yield

\[
\psi_t(y_t) = \frac{1}{s_t} \psi\left(\frac{u_t}{s_t}\right) \tag{5.18}
\]

and

\[
\psi'_t(y_t) = \frac{1}{s_t^2} \psi'\left(\frac{u_t}{s_t}\right) \tag{5.19}
\]

where \( u_t = y_t - \hat{x}_{t|t-1}, \hat{x}_{t|t-1} = a_t|t-1, \psi(z) = -\frac{d}{dz} \log g(z), \) and \( \psi'(z) = \frac{d}{dz} \psi(z), \)

where the density \( g(*) \) is given by (4.8).

Now, once again using the rationale stated in Section 4, we may replace \(-\log g(*)\) with a properly chosen robustified function \( \rho(*) \). Then with \( \psi = \rho \), say \( \psi = \psi_{HR} \) or \( \psi = \psi_{UA} \) specified in Section 3, the Masreliez' filter becomes:

\[
a_t = a_t|t-1 + \frac{1}{M_t} \frac{1}{s_t} \psi\left(\frac{u_t}{s_t}\right) \tag{5.20}
\]

\[
P_t = M_t - \frac{1}{M_t} M_t \frac{1}{s_t^2} \psi'\left(\frac{u_t}{s_t}\right). \tag{5.21}
\]

We shall henceforth refer to the above recursions, along with the prediction equations (5.10), (5.11) as an approximate conditional mean (ACM) type filter or robust filter (cf. Martin (1979b), (1980b)). The reader should now compare the first component of (5.20) with the basic filtering equation given in (3.1).

A considerably simplified version of the filter was used by Thompson (1977) and Kleiner, et al. (1979). However careful examination of (5.20) and (5.21) reveal two attractive features: (i) the scale \( s_t \) adapts to the data in a natural way; (ii) the vector recursion (5.20) allows a good data point which

---

1Although we continue using the same symbols \( a_t, P_t \), etc., the reader should be aware that from now on these symbols refer to values obtained from the Masreliez recursions (5.20) and (5.21).
follows one outlier to help correct the entire prediction vector so that the
next prediction will be improved. Neither of these features were retained
in the simplified versions referred to above.

In the case of the Hampel function $\psi_{HA}$ and when the time series con-
tains only a small fraction of outliers, e.g., $0 < \gamma < .1$, we find that the
filter leaves most of the data unaltered, while a small fraction of the data
containing large outliers are replaced with one-sided predictions. Under these
conditions the ACM-type filter operates as a one-sided outlier interpolator.

The robust filter components $a_t|_{t-1}$ and $s_t$ are expressible as functions
of the parameter vector $\pi$ and the observations $y_t$. Hence, we can find the robust
parameter estimates by direct minimization of the robustified objective function
$R(\pi)$. Although it is our intention at some later time to implement the full
minimization estimates, we describe in the next section somewhat simpler esti-
mates.
6. Iterative Robust Estimation Procedure

In this section we will make some further assumptions which will simplify the robustified objective function $R(\pi)$ (4.12). We will then be able to use the ACM-type filter presented in Section 5 in conjunction with the usual nonlinear unconditional least squares ARIMA model fitting algorithm to minimize $R(\pi)$. The function $R(\pi)$ is used here only to determine the parameters $\hat{\phi}$ and $\hat{\theta}$ directly and then $\sigma^2$ is estimated robustly based on the implied residuals.

We will again assume, as in Section 4, that the measurement error $v_t$ is zero most of the time, i.e., $P(v_t = 0) = 1 - \gamma$ with small $\gamma > 0$, and, hence according to (4.10), $s_t = \sqrt{m_t}$.

Differentiating $R(\pi) = R(\hat{\phi}, \hat{\theta}, \sigma^2)$ in (4.12) with respect to parameters $(\phi, \theta)$ we obtain the following equation for the full maximum-likelihood-type estimates:

$$
\frac{1}{\sqrt{m_t}} \sum_{t=1}^{n} \frac{1}{\sqrt{m_t}} \psi \left( \frac{u_t}{\sqrt{m_t}} \right) \frac{\partial \hat{\tau}_t}{\partial \phi} + \frac{1}{2} m_t \psi \left( \frac{1}{\sqrt{m_t}} \frac{\partial m_t}{\partial \phi, \theta} \right) \frac{1}{\sqrt{m_t}} \frac{u_t}{\sqrt{m_t}} = 0
$$

where as before $u_t = y_t - \hat{R}_t | t-1$ and $\psi(z) = \rho'(z)$.

We now assume that the first term in (6.1) dominates the second one. This will be the case, for example, if $\frac{\partial m_t}{\partial \phi, \theta} = 0$ most of the time, see Martin (1980b) for a similar approximation. Under this assumption equation (6.1) becomes

$$
\frac{1}{\sqrt{m_t}} \sum_{t=1}^{n} \psi \left( \frac{u_t}{\sqrt{m_t}} \right) \frac{\partial \hat{\tau}_t}{\partial \phi} = 0
$$

(6.2)

Notice that equation (6.2) has the same form as Huber's robust nonlinear regression equation (see Huber 1977a, pp. 36-37, 1981, p. 179 ff.) but with a time-varying scale.

There exist several iterative procedures for solving the equations of the form (6.2), see, for example, Huber (1977a), Holland and Welsch (1977). The iterative procedure we use is a modification of the so-called H-algorithm
due to Huber (1977a, p. 39). The H-algorithm is the standard iterative non-linear least squares algorithm, but with a certain smoothing transformation to the residuals. The convergence of this algorithm in the linear regression situation was proved by Huber (1981).

Our algorithm is comprised of the following steps (see Martín (1980b)).

1) Compute initial robust estimates of $\hat{\phi}$, $\hat{\theta}$, and $\hat{\sigma}_e$; one possible method of doing this is discussed below in Section 7.

2) Run a one-sided ACM-outlier interpolator which computes the "cleaned" observations, i.e., the filtered values $x_t^{(j)}$, the predicted values $\hat{x}_{t|t-1}^{(j)}$, and scale $s_t^{(j)} = \sqrt{m_t^{(j)}}$ based on the current parameter estimate $\hat{n}(j) = (\hat{\phi}(j), \hat{\theta}(j), \hat{\sigma}_e^2(j))$.

At this point we would like to reiterate that as mentioned earlier much of our robust filtering applies equally well to the case where $(\hat{\phi}, \hat{\theta})$ corresponds to a nonstationary ARIMA model. As demonstrated in Appendix A, if the data is nonstationary but can be made stationary by appropriate differencing, we can rewrite the ARIMA model in an expanded state variable form, and then apply the one-sided ACM-outlier interpolator to this expanded state variable form. This has as major advantage that we interpolate outliers in the raw data and hence avoid the proliferation of outliers due to differencing the raw data.

3) Use $a_t^{(j)}$ and $\hat{n}(j)$ as input to the standard ARIMA nonlinear least squares estimation procedure which computes $(\hat{\phi}^{(j+1)}, \hat{\theta}^{(j+1)})$. Note again that if differencing is required, we now difference the filtered data $x_t^{(j)}$ before estimating $\hat{\phi}$ and $\hat{\theta}$. Using the fact that

$$\frac{1}{s_t} \left( \frac{u_t}{s_t} \right) = \frac{1}{m_t} (x_t - \hat{x}_{t|t-1})$$
which follows immediately from (5.20), we can represent the next iteration as follows:

\[
\begin{bmatrix}
\hat{\theta}^{(j+1)} \\
\hat{\phi}^{(j+1)} \\
\end{bmatrix}
= \begin{bmatrix}
\hat{\theta}^{(j)} \\
\hat{\phi}^{(j)} \\
\end{bmatrix}
+ (C'C)^{-1}C'd^{(j)}
\]

(6.3)

where \(d^{(j)} = \{d_t^{(j)}\}\) is an \(n\)-component vector with \(d_t^{(j)} = \frac{x_t^{(j)} - \hat{x}_t^{(j)}}{m_t^{(j)}}, \quad t=1,...,n\).

and \(C\) is an \(n\times(p + q)\) matrix with elements

\[
C_{tk} = \begin{cases}
\frac{\partial R_t}{\partial \phi_k} |_{t-1}, \quad k = 1,...,p, \\
\frac{\partial R_t}{\partial \phi_{k-p}} |_{t-1}, \quad k = p+1,...,p+q
\end{cases}, \quad t = 1,...,n
\]

Naturally, the matrix \(C\) is reevaluated at each iteration.

4) Next, calculate the standard deviation of the new least squares residuals \(\hat{r}_t^{(j+1)}\) as a new scale estimator \(\sigma_{c}^{(j+1)}\).

5) Iterate back to 2) until the estimates do not change by more than a prespecified tolerance.

---

\(^1\)Currently we do not stop the standard ARIMA estimation procedure after one iteration, but fully iterate using the current filtered data \(x_t^{(j)}\).
7. Starting Values

For stationary processes the recursions (5.10), (5.11), (5.20), (5.21) are initialized by setting \( a_{1|0} = 0 \) and solving the so-called filtering Liapunov equations for the initial state variance matrix, i.e., solve the equation

\[ P_o = \Phi P_o \Phi' + Q. \]  

(7.1)

for \( P_o \), the covariance matrix for the stationary ARMA process \( x_t \), see Gardner et al. (1980).

For a nonstationary ARIMA model the above approach no longer makes sense. Similar to suggestions made in the literature, in the nonstationary case we start the recursion at time \( t = 0 \), with an initial covariance matrix

\[ P_o = \lambda I \]  

(7.2)

where \( \lambda \) is a sufficiently large number\(^1\). After \( k \) recursions, with \( k \) the dimension of the state vector, the covariance matrix will be approximately equal to the one obtained according to (7.1).

Next, we need initial guess values for the ARIMA parameter values \( \hat{\phi}, \hat{\theta} \) and \( \sigma_e^2 \). As is well known, an ARMA process can always be rewritten as a infinite order autoregressive process. If we now approximate an ARMA process with a finite order AR model we could then estimate the parameters of the finite AR process robustly using the so-called autoregression GM-estimation procedure (Martin, 1980a). Next, given these AR estimates, we could use Durbin (1959) formulae (see also Fuller (1976, p. 353 and p. 359)), to obtain the initial estimates of the ARMA model parameters \( \hat{\phi} \) and \( \hat{\theta} \). At present we use either guess values obtained from the acf or estimates obtained after first estimating the

\(^1\)In our empirical work we currently used a value for \( \lambda = 1000 \). At present we are evaluating values for \( \lambda \) determined by the variance of the data.
model using the standard ARIMA nonlinear estimation. We have found that the final results are not affected by a particular choice of the guess values, only the number of iterations is.

A robust starting value for the scale of the error in equation (5.1) of the state vector model could be obtained using a robust scale estimate, e.g.,

$$\hat{\sigma}_e = \text{median}|y_t - \text{median}(y_t)| / 0.6745$$

(7.3)

where the choice of 0.6745 insures that $\hat{\sigma}_e$ is asymptotically unbiased when the data is $N(0, \sigma^2)$. Other robust scale estimators which are more efficient under certain conditions have been proposed in the literature (see e.g., Kafadar (1979) and Bell (1980)) and deserve to be implemented.
8. **Model Identification**

When there are no outliers, Box and Jenkins (1976) suggest a procedure to identify or select an appropriate subclass of the general ARIMA models based on the sample autocorrelation function (acf) and partial autocorrelation function (pacf) of the stationary series. However, the use of the standard acf and pacf formulae when the data contain outliers can be very misleading. These estimates lack robustness just as do the usual product moment correlation estimates for independent bivariate samples. One way around the difficulty may be to adapt robust correlation and covariance methods for i.i.d. multivariate samples (see Devlin et al. (1975), Maronna (1976), Huber (1977b), Marazzi (1980), Rieder (1980)) to the time series setting.

Currently we are handling the problem in another way based on an iterative identification procedure. We begin by using the usual Box-Jenkins identification procedure based on the initial unfiltered data to specify an initial model. Next, the cleaned data obtained by the robust filtering is used as the basis of a new model identification pass. If the same model is identified as when using the raw data, and if the diagnostic checks on the estimation results (e.g., check on over- and underspecification, residual analysis) do not suggest model misspecification, we are done. Otherwise we carry out the robust estimation on the new model, and apply the same diagnostic checks. Hopefully, a small number of iterations will suffice. For the examples presented in Section 10, and several other unreported examples, this approach seemed to work adequately.
9. **Forecasting**

One method of computing the forecasts based on an ARIMA(p,q) model is to use the so-called random-shock form of the process, i.e., rewriting the model in the form of an MA(\infty) process. When using this form the mean squared error (MSE) of a particular forecast may be calculated linearly, given estimates of \( \hat{\phi} \) and \( \hat{\theta} \), see Box and Jenkins (1976, Ch. 4). However, since we are using a state variable representation, it is not necessary to explicitly recast the model in the form of an MA(\infty) process, but the forecasting is carried out by projecting the recursion formulae forward in such a way that they yield the appropriate predictions and their MSE's.

Consider the prediction of \( x_{n+1} \), given all \( n \) observations \( y_1, \ldots, y_n \). We will use the notation \( \hat{x}_{n+h|n} \), to denote the forecast of \( x_{n+h} \) made at time \( n \) with \( h \) the forecast horizon. Recall that the first component of \( a_n \) is \( x_n \).

Therefore we will forecast \( a_n \) and then select the first component as forecast of \( x_n \). Based on the recursive equations (5.6) and (5.7) with \( a_{t-1} \) and \( M_t \) defined in (5.10) and (5.11), we have

\[
\hat{a}_{n+1|n} = \hat{\phi} a_n
\]

and the forecast covariance matrix of \( \hat{a}_{n+1|n} - a_{n+1} \) is given by

\[
P_{n+1|n} = M_{n+1} = \hat{\phi} P_n \hat{\phi}' + Q.
\]

Again, as long as a nonstationary model can be modeled with an ARIMA(p,d,q) model, there are no formal difficulties in representing such a model in a state variable form, see Appendix A. Therefore simple extension of equations (9.1) and (9.2) allow us to generate forecasts for an ARIMA(p,d,q) model.

In general, forecasts can be made for any horizon as

\[
\hat{a}_{n+h|n} = \hat{\phi} a_{n+h-1|n} = \hat{\phi}^h a_n.
\]

The important feature of these forecasts is that they are driven by the robustly filtered data, not the original data which may contain outliers.
The associated forecast covariance matrix for the vector forecast $\hat{a}_{n+h|n}$ is

$$P_{n+h|n} = \Phi P_{n+h-1|n} \Phi' + Q$$

$$= \phi^h P_n \phi^h + \sum_{\ell=1}^{h} \phi^{\ell-1} Q \phi^{\ell}$$

with $\phi^h = \phi^{h-1} \phi^0 = I$.

These forecast formulae can be regarded as the Kalman recursion equation with the updating equations bypassed, i.e., (5.6) becomes $\hat{a}_{n+1} = \hat{a}_{n+1|n}$ and (5.7) becomes $P_{n+1} = P_{n+1'}$.

The observation prediction for $y_t$, the first component of $\alpha_t$, can now be obtained using the measurement equation (2.2):

$$y_{n+h|n} = z' a_{n+h|n} = a_{n+h|n} = \hat{x}_{n+h|n}$$

with $a_{n+h|n}$, the first component of $\hat{a}_{n+h|n}$, and the forecast variance $y_{n+h|n}$ becomes

$$\text{var} \ y_{n+h|n} = z' P_{n+h|n} z = \sigma^2 f_{n+h|n}$$

with $f_{n+h|n}$ the l-1 element of $P_{n+h|n}$, divided by $\sigma^2$. 


10. Examples

10.1 Series RESEX

Series RESEX is a monthly series of Bell Canada inward movement of residential telephone extensions in a fixed geographic area from January 1966 to May 1973, a total of 89 data points. Figure 3 contains a time series plot of the data and clearly shows two extremely large values in November and December 1972 due to a November "bargain month", i.e., free installation of residence extensions, and a spillover effect in December because all November's orders could not be filled in the same month.

Brubacher (1974) identified the stationary series as an ARIMA $(2,0,0) \times (0,1,0)_{12}$ model, i.e., the RESEX data is represented by an AR(2) model after seasonal differencing. Estimating the parameters of this model produces the results presented in Table 1. As can be seen, the ACM-type robust filter results are quite different from the least squares estimates without correction for outliers. This is particularly the case for the $\phi_2$ parameter which even has the opposite sign although its SE is large relative to its point estimate. In the last column of Table 1, we also report results based on a method proposed by Brubacher. This method is briefly described in Appendix B. The basic difference is that Brubacher's method requires a priori designation of the outlier points. As can be seen from this Table, Brubacher's results are very similar to the ACM-type robust filter estimates. This should not come as a complete surprise because it is trivial to correctly specify the outliers in the RESEX series. Figure 4 is a plot of the data and the filtered data.

---

1 We thank J. Schmitz for modifying the basic Kalman filter subroutines. The Kalman filter is based on Gardner, et al. (1980). The programs are available as part of the TROLL system for IEM machines and the TS package for DEC10, DEC20, and VAX machines.

2 The data are listed in Appendix C.
FIGURE 3. RESEX
January 1966 – May 1973
(thousands)
FIGURE 4. RSEX and RSEX Robustly Filtered

January 1966 to May 1973

□ RSEX

◇ RSEX Filtered
(thousands)
Table 1
RESEX: January 1966 - May 1973
Estimates a)

<table>
<thead>
<tr>
<th></th>
<th>No Correction for Outliers b)</th>
<th>ACM - type Robust Filter</th>
<th>Brubacher (1974)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_1$</td>
<td>.537 (.113)</td>
<td>.522 (.109)</td>
<td>.51 (.11)</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>-.106 (.115)</td>
<td>.350 (.109)</td>
<td>.36 (.11)</td>
</tr>
<tr>
<td>RSE*10^3 c)</td>
<td>6.5</td>
<td>1.3</td>
<td>1.2</td>
</tr>
</tbody>
</table>

a) Large sample standard errors between parentheses. The ACM-type robust filter standard errors are obtained from the unconditional least squares ARIMA algorithm applied to the filtered data in the last iteration.
b) Box-Jenkins unconditional backforecasted least squares estimates.
c) RSE: residual standard error.

10.2 Series UNFTV 1

The next series we will analyze is the value of Unfilled Orders, Radio and TV (UNFTV). This is a monthly series for the period January 1958 to October 1980, a total of 274 data points. Figure 5 contains a plot of the data. There is an annual trade show in the Spring which is usually reflected in this series by an increase around trade show time. Throughout the year there are promotional deals and model close-out deals which affect unfilled orders, new orders, and sales.

The identification of the model, using the autocorrelation and partial autocorrelation function in the standard Box-Jenkins approach based on the raw data, indicated that the data should be differenced consecutively and

The data are listed in Appendix C.
FIGURE 5. Value of Unfilled Orders, Radio and TV (UNFTV)
January 1958 - October 1980
(millions of dollars)
seasonally and that we should include a first-order moving average term and a first-order seasonal moving average term. Thus the initial model for UNFTV is ARIMA(0,1,1)×(0,1,1)_{12}.

Table 2 contains the estimates for this series. The point estimates and large sample standard errors obtained using the standard ARIMA estimation procedure are presented in the first column. The observant user might well express doubt about these results since the value for September 1978 appears to be a clear outlier. We tried to mimic what certain users might do with such an outlier. The results given in the next column are based on the standard ARIMA estimation results after modifying the September 1978 value from $973 million to $600.544 million. This latter value is not an interpolation from previous and next value, but is obtained after robustly filtering the data. The results in Table 2 show very little change in the estimates due to modifying the September 1978 value. Finally, in the last column we present the ACM-type robust filter results. Notice that both coefficient estimates, and in particular the MA(1) coefficient, are now substantially different. It must therefore be the case that, besides the September 1978 point, there are other outliers in the data. In Figure 6a, we have represented the difference between UNFTV series and the UNFTV filtered series. As can be seen from this figure, there appear to be quite a number of additional outliers. In Figure 6b we plotted UNFTV and UNFTV filtered from January 1976 on.

These results indicate rather vividly that in time series data it may be quite difficult to specify by visual inspection where there are outliers in the data. This is just a reflection of the fact that an outlier is not necessarily a globally extreme point, as in the case of the September 1978 value; it may only be a locally extreme point that is cloaked by the general structure of the process. Basically outliers are determined by the prediction residuals.
Table 2
UNFTV: January 1958 - October 1980
Estimates a)

<table>
<thead>
<tr>
<th></th>
<th>No Correction for Outliers b)</th>
<th>Corrected b), c) Sept. 1978</th>
<th>ACM-type Robust Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_1$</td>
<td>.407 (.067)</td>
<td>.389 (.059)</td>
<td>.083 (.067)</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>.816 (.056)</td>
<td>.818 (.057)</td>
<td>.880 (.036)</td>
</tr>
<tr>
<td>RSE*10^6 d)</td>
<td>55.9</td>
<td>50.7</td>
<td>38.2</td>
</tr>
</tbody>
</table>

a) Large sample standard errors between parentheses. The ACM-type robust filter standard errors are obtained from the unconditional least squares ARIMA algorithm applied to the filtered data in the last iteration.
b) Box-Jenkins unconditional backforecasted least squares estimates.
c) The value for September 1978 was changed from $973 million to $600.544 million.
d) RSE: residual standard error.
FIGURE 6a. UNFTV - UNFTV Robustly Filtered
January 1958 - October 1980
(millions of dollars)
FIGURE 6b. UNFTV and UNFTV Robustly Filtered
January 1976 - October 1980
□ UNFTV
FIGURE 7. Scatter Diagram UNFTV Least Squares Residual

Y axis: One Period Lagged Residuals
X axis: Residuals

(Notice, the Y axis and X axis have different scales)
FIGURE 8. Scatter Diagram UNFTV Robustly Filtered Residuals

Y axis: One Period Lagged Residuals
X axis: Residuals
In Figures 7 and 8 we have presented the scatter diagrams of the least squares residuals, based on the results in column 1 of Table 2, and the robustly filtered residuals, based on the results in the last column of Table 2. For white noise residuals such a scatter diagram should be a circle centered at zero. A comparison of these two figures clearly indicates that the least squares residuals contain several outliers, whereas this is less clearly the case for the robustly filtered residuals.
11. Conclusions and Topics for Further Research

In this paper we have shown that the ACM-type robust filter can effectively handle a variety of outlier situations. In addition the proposed method is easy to implement in its current form. All that is needed is a robust filtering program together with a standard ARIMA nonlinear optimizer. The robust filtering program could be obtained by simple modifications to a good (i.e., numerical stable) Kalman filtering program (see Bierman, 1977).

One of the additional appealing features of the approach is that the program can easily be modified to handle missing data. For the missing data case we choose \( \psi(u_t) = 0 \) if the corresponding observation is missing. In effect we can treat a missing observation as an observation with a very large prediction residual and set the corresponding filtered value \( \hat{x}_t \) equal to the predicted value \( \hat{x}_{t-1} \). In fact this amounts to making the \( \psi \) function depend upon time, \( \psi = \psi_t \), with \( \psi_t \) the identify function if \( t \) corresponds to a time at which data is observed, and with \( \psi_t \) the zero function otherwise. This corresponds to maximizing the likelihood function for missing data problems, see Jones (1980), Dunsmuir and Robinson (1981).

However, there remains an important distinction between missing data problems and outlier problems in that one does not usually know in advance where the outliers occur.

A number of topics are open for research:

1. Model Identification/Model Selection. As mentioned in Section 8 it could be useful to calculate robust autocorrelation and robust partial autocorrelation functions. One such method which yields a consistent correlogram at the Gaussian has recently been suggested by Yohai (1981).
The issues in calculating robust autocorrelations for time series are, first, that we want to enforce Toeplitz structure on the correlation or covariance matrix estimate, thereby reducing the number of parameters to be estimated; and second, that we must enforce positive definiteness. Therefore the current proposal to produce robust correlation and covariance estimates for i.i.d. multivariate data, see Devlin et al. (1975), Maronna (1976), Huber (1977b), Marazzi (1980), Rieder (1980) cannot be used as such and in particular would have to be adapted to produce a Toeplitz structure.

As a complementary or alternative approach we could adapt Akaike's (1974b) model selection criterion to obtain a robust model selection criterion. Consider the case of an AR(p) model. Akaike suggests selecting the order p based on the minimization of the AIC information criterion:

$$\text{AIC}(p) = -\log h(y|\hat{\pi}, p) + (p + 2)$$  \hspace{1cm} (11.1)

where $\hat{\pi}$ is the MLE of $\pi$ with $\pi = (\phi, \sigma^2)$. This criterion has been the subject of much research, see e.g., Bhansali and Downham (1977), Shibata (1976, 1980), Geweke and Meese (1981). Ozaki (1977) used this method to determine the optimal orders in an ARMA(p,q) model.

One natural and obvious possibility is to apply the AIC criterion on the robust loss function as defined in (4.12), so that a Robust Akaike Information Criterion can be defined as

$$\text{RAIC}(k) = R(\hat{\pi}, k) + (k + 2)$$

with $k = p + q$, and $R(\hat{\pi}, k)$ defined according to (4.12).

A disturbing property of the AIC approach is that it tends to produce overelaborate models. This has been demonstrated theoretically and empirically by Shibata (1976) in the case where the choice is between alternative pure autoregressive processes. One might suspect that, when mixed autoregressive-moving average alternatives are contemplated, the situation would be even worse. Recently strongly consistent procedures for pure autoregressive models have been introduced, see Akaike (1977), Rissanen (1978), Schwarz (1978), Hannan
and Quinn (1979). The Hannan and Quinn procedure is, in particular, interesting because in large samples it will underestimate the order less than those other strongly consistent procedures.

2. Another consideration is that of extending the current robust filter from a one-sided filter to a two-sided filter. The ACM-type filter uses only the data up to time \( t \) in evaluating if the observation at time \( t + 1 \) is an outlier. It would certainly be desirable to use all the data. One possibility for using all the data has been described in Martin (1979b). We propose to study the use of such a two-sided filter in future work.

3. We have not addressed theoretical considerations such as consistency. A highly desirable procedure should be both robust and consistent at the nominal Gaussian model. Although our approach is appealing and robust, at the nominal Gaussian model it is in general not consistent. A study of the linear part of the robust filter, see Mallows (1980), indicates that the asymptotic bias will be small at the Gaussian model, see Martin (1981). In some recent work by Yohai and Bustos (1981), consistent estimates that are robust have been proposed for some classes of ARMA models. What is therefore needed is a Monte Carlo study to evaluate the bias and variance in small samples.
Appendix A

State Variable Representation

As mentioned in Section 5, the ARMA(p,q) model
\[ \phi(B)x_t = \theta(B)\varepsilon_t \]  \hspace{1cm} (A.1)
with measurement equation
\[ y_t = x_t + \nu_t \]  \hspace{1cm} (A.2)

can also be represented in vector state variable form\(^1\) as

\[ \alpha_t = \phi \alpha_{t-1} + \Gamma \varepsilon_t \]  \hspace{1cm} (A.3)

\[ y_t = z'\alpha_t + \nu_t \]  \hspace{1cm} (A.4)

where vector \( z = (1, 0, \ldots, 0)' \) and \( \alpha_t \) is the vector state variable, with
first component \( \alpha_{1t} = x_t \). Other quantities in (A.3) and (A.4) are defined
below.

We consider here only one possible form of the state variable representa-
tion. This state variable form has also been used by Ledolter (1976, 1979),
used a somewhat different form. See also Priestley (1980) for the use of the
Kalman filter for nonlinear time series models.

For expositional purposes, first assume that \( p > q \), and let
\[ \alpha_{1t} = \phi_1 \alpha_{1,t-1} + \alpha_{2,t-1} + \varepsilon_t \]
with
\[ \alpha_{2,t-1} = \phi_2 \alpha_{1,t-2} + \phi_3 \alpha_{1,t-3} + \ldots + \phi_p \alpha_{1,t-p} - \theta_1 \varepsilon_{t-1} \]
\[ - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \]

\(^1\)Although the vector state variable forms (A.3) and (A.4) have
similarities with the linear regression model with time-varying coefficients,
these two models are fundamentally different. For an expositional paper on
random and changing coefficients models, see Chow (1980).
Next define

\[ \alpha_{t-2}^3 = \phi_{3t-1}^1 + \cdots + \phi_{t-1}^p \alpha_{t-p}^1 - \theta_{2t-2}^1 - \cdots - \phi_{t-q}^q \epsilon_t^q. \]

Then continuing in this fashion we obtain

\[ \alpha_{t-1}^1 = \phi_{1t-1}^1 + \alpha_{t-1}^2, \]
\[ \alpha_{t}^2 = \phi_{2t}^1 + \alpha_{t-1}^3, \]
\[ \alpha_{t}^3 = \phi_{3t}^2 + \alpha_{t-1}^4, \]
\[ \vdots \]
\[ \alpha_{p-1}^{p-2} = \phi_{p-1t}^p + \alpha_{t-1}^{p+1}, \]
\[ \alpha_{p}^p = \phi_{pt}^1 - \theta_{p-1t}^p. \]

with \( \theta_i = 0 \) for \( i > q \).

The corresponding state transition matrix is

\[
\Phi = \begin{bmatrix}
\phi_1 & & & \\
& \ddots & & \\
& & I_{(k-1)} & \\
& & & \ddots \\
& & \cdots & \\
& & & \ddots \\
\phi_k & & & \ddots \\
& & & & & 0
\end{bmatrix}
\]

with \( k = p \), \( I \) is the \((k-1)\times(k-1)\) identity matrix, \( 0 \) is the \((k-1)\) column vector of zeros.

If \( q \geq p \), then the above procedure leads to a state equation of dimension \( q + 1 \) and the first column of \( \Phi \) contains \( \phi_1, \phi_2, \ldots, \phi_{q+1} \), with \( \phi_i = 0 \).
for \( i > p \). Therefore, in general, the dimension of the state transition matrix (A.6) is \( k = \max(p, q + 1) \). The vector \( \mathbf{r} \) in (A.3) is a \( k \times 1 \) column vector defined as \((1, -\theta_1, -\theta_2, \ldots, -\theta_{k-1})\) with \( \theta_i = 0 \) for \( i > q \) if \( p > q \). Based on this representation the first coordinates of \( \alpha_t \), \( \alpha_{t-1}, \ldots, \alpha_{t-p+1} \) contain \( x_t \), \( x_{t-1} \), \ldots, \( x_{t-p+1} \), respectively.

The same form can also be used to represent a general ARIMA\((p,d,q)\) model

\[
\phi(B)(1-B)^d x_t = \theta(B)\epsilon_t \tag{A.7}
\]

or

\[
\phi(B)x_t = \theta(B)\epsilon_t \tag{A.8}
\]

with

\[
\phi(B) \equiv \phi^*(B)(1-B)^d. \tag{A.9}
\]

The order of \( \phi(B) \) is now \( p + d \). For example, the parameters of the polynomial \( \phi(B) \) in an ARIMA \((1,1,q)\) model can be shown to be equal to

\[
\phi_1 = 1 + \phi^*_1 \]
\[
\phi_2 = -\phi^*_1.
\]

As a result, the above derivations carry over after replacing \( p \) with \( p + d \) and therefore the dimensionality of \( k \) is now \( k = \max(p + d, q + 1) \).

Similarly, a multiplicative seasonal model

\[
\phi^*(B) \phi(B^s)(1-B)^d (1-B^s)^D x_t = \theta^*(B)\theta(B^s)\epsilon_t \tag{A.10}
\]

can be represented as

\[
\phi(B)x_t = \theta(B)\epsilon_t \tag{A.11}
\]

with

\[
\phi(B) = \phi^*(B)\phi(B^s)(1-B)^d (1-B^s)^D \]
\[
\theta(B) = \theta^*(B)\theta(B^s)\]
where

\[ \Phi(B^s) \] is the seasonal autoregressive polynomial

\[ \Theta(B^s) \] is the seasonal moving average polynomial

\[ (1-B^s)^D \] is the seasonal difference operator

\[ s \] is the seasonal span.

The dimensionality of the state vector representation is now determined by the orders of the \( \Phi(B) \) and \( \Theta(B) \) polynomials in (A.11).
Appendix B

Brubacher Estimation Method

Brubacher (1974), see also Brubacher and Wilson (1976), proposes to estimate the ARMA parameters iteratively by taking into account the important fact that the least squares estimation of the interpolates is a linear estimation problem conditional upon the parameter values.

For a particular ARMA model the method becomes

1. Select initial values for the model parameters.

2. Use linear estimation to obtain the least squares estimates for the outliers, the interpolates, conditional upon the parameter values of step 1.

3. Conditional upon the interpolates, obtain (nonlinear) least squares estimates for the ARMA parameters.

4. Go to step 1 until convergence of the model parameters and interpolates is attained.

The fundamental difference between the Brubacher method and our robust filter method is that with the Brubacher method the user must identify a priori which data points are to be treated as outliers. The robust filter method needs no such a priori decision.
**RESEX: Residence Telephone Extensions Inward Movement (Bell Canada)**

<table>
<thead>
<tr>
<th></th>
<th>JAN</th>
<th>FEB</th>
<th>MAR</th>
<th>APR</th>
<th>MAY</th>
<th>JUN</th>
<th>JUL</th>
<th>AUG</th>
<th>SEP</th>
<th>OCT</th>
<th>NOV</th>
<th>DEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1967</td>
<td>10545.</td>
<td>10120.</td>
<td>11877.</td>
<td>14752.</td>
<td>16932.</td>
<td>14123.</td>
<td>14777.</td>
<td>14943.</td>
<td>16573.</td>
<td>15548.</td>
<td>15838.</td>
<td>14159.</td>
</tr>
</tbody>
</table>
## UNFTV: Value of Unfilled Orders, Radio and TV

(millions of dollars)

US Bureau of the Census

<table>
<thead>
<tr>
<th>JAN</th>
<th>FEB</th>
<th>MAR</th>
<th>APR</th>
<th>MAY</th>
<th>JUN</th>
<th>JUL</th>
<th>AUG</th>
<th>SEP</th>
<th>OCT</th>
<th>NOV</th>
<th>DEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>208</td>
<td>219</td>
<td>224</td>
<td>224</td>
<td>233</td>
<td>242</td>
<td>256</td>
<td>245</td>
<td>245</td>
<td>230</td>
<td>224</td>
<td>227</td>
</tr>
<tr>
<td>221</td>
<td>217</td>
<td>244</td>
<td>247</td>
<td>246</td>
<td>287</td>
<td>299</td>
<td>296</td>
<td>306</td>
<td>297</td>
<td>277</td>
<td>269</td>
</tr>
<tr>
<td>261</td>
<td>263</td>
<td>239</td>
<td>237</td>
<td>254</td>
<td>291</td>
<td>345</td>
<td>351</td>
<td>343</td>
<td>324</td>
<td>288</td>
<td>301</td>
</tr>
<tr>
<td>293</td>
<td>266</td>
<td>271</td>
<td>314</td>
<td>319</td>
<td>331</td>
<td>381</td>
<td>366</td>
<td>420</td>
<td>394</td>
<td>389</td>
<td>373</td>
</tr>
<tr>
<td>359</td>
<td>388</td>
<td>391</td>
<td>380</td>
<td>401</td>
<td>371</td>
<td>412</td>
<td>370</td>
<td>342</td>
<td>304</td>
<td>268</td>
<td>272</td>
</tr>
<tr>
<td>283</td>
<td>288</td>
<td>296</td>
<td>332</td>
<td>364</td>
<td>371</td>
<td>385</td>
<td>427</td>
<td>387</td>
<td>378</td>
<td>375</td>
<td>436</td>
</tr>
<tr>
<td>402</td>
<td>453</td>
<td>481</td>
<td>481</td>
<td>524</td>
<td>438</td>
<td>507</td>
<td>511</td>
<td>491</td>
<td>474</td>
<td>446</td>
<td>432</td>
</tr>
<tr>
<td>436</td>
<td>423</td>
<td>415</td>
<td>468</td>
<td>491</td>
<td>524</td>
<td>545</td>
<td>597</td>
<td>597</td>
<td>585</td>
<td>547</td>
<td>552</td>
</tr>
<tr>
<td>637</td>
<td>625</td>
<td>600</td>
<td>607</td>
<td>657</td>
<td>727</td>
<td>774</td>
<td>758</td>
<td>724</td>
<td>738</td>
<td>657</td>
<td>609</td>
</tr>
<tr>
<td>618</td>
<td>580</td>
<td>494</td>
<td>504</td>
<td>494</td>
<td>495</td>
<td>560</td>
<td>577</td>
<td>528</td>
<td>535</td>
<td>438</td>
<td>431</td>
</tr>
<tr>
<td>389</td>
<td>450</td>
<td>419</td>
<td>376</td>
<td>387</td>
<td>587</td>
<td>606</td>
<td>525</td>
<td>546</td>
<td>488</td>
<td>445</td>
<td>480</td>
</tr>
<tr>
<td>480</td>
<td>530</td>
<td>514</td>
<td>497</td>
<td>486</td>
<td>536</td>
<td>570</td>
<td>508</td>
<td>533</td>
<td>543</td>
<td>430</td>
<td>397</td>
</tr>
<tr>
<td>439</td>
<td>438</td>
<td>449</td>
<td>441</td>
<td>476</td>
<td>533</td>
<td>526</td>
<td>533</td>
<td>526</td>
<td>491</td>
<td>433</td>
<td>463</td>
</tr>
<tr>
<td>437</td>
<td>453</td>
<td>456</td>
<td>439</td>
<td>465</td>
<td>461</td>
<td>468</td>
<td>486</td>
<td>493</td>
<td>495</td>
<td>485</td>
<td>460</td>
</tr>
<tr>
<td>463</td>
<td>478</td>
<td>503</td>
<td>529</td>
<td>574</td>
<td>671</td>
<td>637</td>
<td>653</td>
<td>743</td>
<td>737</td>
<td>650</td>
<td>671</td>
</tr>
<tr>
<td>567</td>
<td>664</td>
<td>649</td>
<td>602</td>
<td>591</td>
<td>681</td>
<td>729</td>
<td>691</td>
<td>740</td>
<td>724</td>
<td>698</td>
<td>606</td>
</tr>
<tr>
<td>615</td>
<td>569</td>
<td>551</td>
<td>561</td>
<td>626</td>
<td>608</td>
<td>657</td>
<td>691</td>
<td>629</td>
<td>599</td>
<td>505</td>
<td>471</td>
</tr>
<tr>
<td>408</td>
<td>410</td>
<td>369</td>
<td>374</td>
<td>357</td>
<td>376</td>
<td>438</td>
<td>462</td>
<td>460</td>
<td>454</td>
<td>468</td>
<td>453</td>
</tr>
<tr>
<td>443</td>
<td>439</td>
<td>443</td>
<td>384</td>
<td>406</td>
<td>455</td>
<td>540</td>
<td>587</td>
<td>452</td>
<td>558</td>
<td>495</td>
<td>525</td>
</tr>
<tr>
<td>370</td>
<td>508</td>
<td>318</td>
<td>348</td>
<td>381</td>
<td>434</td>
<td>454</td>
<td>708</td>
<td>554</td>
<td>598</td>
<td>513</td>
<td>491</td>
</tr>
<tr>
<td>413</td>
<td>425</td>
<td>370</td>
<td>467</td>
<td>485</td>
<td>559</td>
<td>549</td>
<td>695</td>
<td>973</td>
<td>682</td>
<td>513</td>
<td>542</td>
</tr>
<tr>
<td>469</td>
<td>645</td>
<td>634</td>
<td>537</td>
<td>317</td>
<td>543</td>
<td>478</td>
<td>618</td>
<td>671</td>
<td>433</td>
<td>357</td>
<td>339</td>
</tr>
<tr>
<td>320</td>
<td>453</td>
<td>360</td>
<td>292</td>
<td>316</td>
<td>424</td>
<td>409</td>
<td>520</td>
<td>492</td>
<td>447</td>
<td>447</td>
<td>447</td>
</tr>
</tbody>
</table>
REFERENCES


