ARMAsel for Detection and Correction of Outliers in Univariate Stochastic Data

Piet M. T. Broersen

***Abstract*—For stationary random data, an automatic estimation algorithm can now select a time series model with a spectral accuracy close to the Cramér–Rao lower bound. The parameters of that selected time series model accurately represent the spectral density and the autocovariance function of the data. That is all the possible information for Gaussian data, as well as the most important information for arbitrarily distributed data. A single model type and order is selected from many candidate time series models by looking for the smallest prediction error. The single selected model precisely includes only the statistically significant details that are present in the data. The residuals of the automatically selected time series model reveal the location of outliers or other irregularities that may not be visible in the measured signal. The program requires no user interaction and can be incorporated into automatic measurement instruments and protocols.**

***Index Terms*—Autocorrelation, autocovariance, autoregressive model, autoregressive moving average (ARMA) model, feature extraction, moving average model, order selection, parametric model, spectral estimation.**

I. Introduction

THE autocorrelation function and the power spectral density contain all the possible statistical information of stationary stochastic data with a joint normal distribution with zero mean. For other distributions, the first and second moments are also very valuable information. Until about 2002, modified periodograms and lagged-product autocorrelation estimates were the standard analysis tools for random data [1]. Modification of the periodogram with a window was necessary to obtain a useful spectral estimate. The best choice of window depended on the data and on the preferences of the data analyst [1] and could not be automatically made.

Time series models can also be used for spectral analysis [1]. Apart from the dedicated univariate time series modeling, the treatment as a linear system, with white noise as the input signal, has also been used for a time series with known model order and type [2]. The computational demands for time series modeling were very heavy in the past for the routine automatic application to measurement data with unknown model type and model order. Only a limited number of models, with order and type specified by the data analyst, could be mutually computed and compared with order selection criteria. The automatic estimation and selection of a single time series model for a given set of random observations enables many new applications to real-life data [3]. It will be investigated whether the automatically selected time series model, with all the spectral details that are statistically significant in the given data, is a good starting point to detect outliers or changes.

Irregular and undesirable disturbances in measured signals can be distinguished in innovational or additive outliers and permanent or transient level changes [4]. AutoRegressive Moving Average (ARMA) time series models can be used to detect outliers, and an example with an AR(1) process shows how all types of disturbances can be incorporated into a univariate model [4]. The detection of outliers can be extended to multivariate time series data [5]. The interaction between the size and the dynamic structure of the model gives special properties to the multivariate time series case [5]. Outliers in multivariate regression problems often require robust estimation methods to be detected because simple least squares estimation becomes biased by the outliers. Recently, many new methods have been developed [6]. An active research field in data mining is the detection of outliers. Distance-based outliers can be detected, but in some examples, the perception can be subjective to what an outlier is [7]. Moreover, outliers should have sufficient distance from inliers for reliable detection. Otherwise, outliers tend to precisely bias the parameters of time series models to reduce the residuals of those models. Furthermore, the residual variance estimate is very large due to outsiders. Simple rules of calling points further than three standard deviations (SDs) away from the mean will be the reason that very few outliers may be detected in dynamic processes [8].

Before 1980, AR models could be estimated for fixed specified orders, with the Burg method [9] and with the Yule–Walker method [10]. However, the AR model class is not generic. The traditional order selection criterion known as Akaike's information criterion [11] often selected very high model orders, particularly in finite and small sample records. MA and combined ARMA models could also be estimated [1] for specified fixed orders. However, different estimation methods give rather different results for the same data [12]. No preference for one single estimation method could be given [13]. Improved order selection criteria for finite samples in AR models, numerically efficient estimation algorithms without iterations for MA and ARMA models, and faster computers altogether give the possibility of a reliable automatic analysis of data with the ARMAsel algorithm [3], [14].

Modified periodograms have been the only practical tools for the routine analysis for spectra of random data for a long time, together with lagged-product estimates for autocovariance functions [1]. The SD of the raw periodogram is approximately equal to its expectation and does not become smaller for longer

Manuscript received July 15, 2006; revised September 18, 2007.

The author is with the Department of Multiscale Physics, Delft University of Technology, 2628 Delft, The Netherlands (e-mail: P.M.T.Broersen@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIM.2007.911576

data sets. Furthermore, the nonzero length of estimated autocovariance functions is always precisely equal to the length of the data record, independent of the true correlation length. Spectral or lag windows are therefore required to improve accuracy, unless the data are periodical. The final results depend on the choice of the shape and width of the window, which has to be subjectively chosen by the data analyst. Therefore, these methods are not suitable for an automatic analysis of random data with unknown characteristics. Moreover, it has been proved that lagged products are not efficient for autocovariance estimation [15]. This means that time series models can extract more accurate information for the autocorrelation function from the same amount of data.

Automatic time series analysis with selection of the model order and type has recently become a new perspective with the ARMAsel program [14] that selects between AR, MA, and ARMA candidate models. Time series models can be subdivided into three types: 1) AR; 2) MA; and 3) combined ARMA models [1]. Finite or low-order AR, MA, or ARMA models are accurate enough in practice because the true parameters of higher orders rapidly decrease for most processes. Using only the class of AR models is not generic and can be very inaccurate for processes with spectral zeros close to the unit circle.

From a single selected model of the ARMAsel program [14], the autocovariance function and the spectral density are computed. The accuracy of this parametric spectrum is typically better than the best of all possible modified periodogram estimates. The data select a model on statistical grounds, the accuracy of which generally approaches the Cramér–Rao lower bound in simulations with sufficient data [3]. According to the invariance property of the maximum likelihood theory, functions that are derived from maximum likelihood parameter estimates are also maximum likelihood estimates for the functions, under mild conditions. This explains the high quality of the time series estimates for the spectrum and the autocorrelation function. An automatic selection algorithm for the model type and order selects a single AR, MA, or ARMA model, which is denoted as the ARMAsel model. Subspace methods constitute another framework for the automatic identification of stochastic observations based on state-space modeling [16], [17]. Asymptotic properties of subspace estimators have been derived [18]. Unfortunately, it has been concluded that automatic procedures based on recommendations from asymptotic theory lead to poor accuracy in finite samples [19]. So far, other automatic methods suffer from poor performance on some types of data and are less suitable [19].

This paper treats an example with turbulence data obtained with direct numerical simulations [20], [21]. The Navier–Stokes equations are solved on a shared-memory supercomputer SGI Origin 3800. A study of the flow around a cylinder gives information about heat and mass transfer to a clothed human limb in extreme outdoor conditions. It very much required computing time, and it is difficult to repeat the computations for an independent verification of the random data. The spectral density of the turbulence signal has been analyzed before with time series analysis [3]. This showed that the selected time series estimates had interesting properties. Irregularities or outliers in the data will be detected and repaired with programs that are available in the ARMAsel toolbox [14]. It is demonstrated how one single bad observation distorts the spectrum of the flow data in more than 50% of the total frequency domain.

II. Time Series Models

Time series models have three different linear types: 1) AR; 2) MA; and 3) combined ARMA models. An ARMA(p,q) model can be written as [1]

$$x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p} = \varepsilon_n + b_1 \varepsilon_{n-1} + \cdots + b_q \varepsilon_{n-q} \tag{1}$$

where ε_n is a purely random process of independent identically distributed stochastic variables with zero mean and variance σ_ε^2. If applied to estimation, then (1) is called a model. If the formula is used to generate new data or to describe the true characteristics of the data, then it is generally called a process. The process or model is purely AR for $q = 0$ and purely MA for $p = 0$. A shift operator is defined such that $z^{-1}x_n$ is equal to x_{n-1}. The roots of the AR parameter polynomial $A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p}$ are denoted as the poles of the ARMA(p,q) model. The roots of $B(z)$, which are defined as $B(z) = 1 + b_1 z^{-1} + \cdots + b_q z^{-q}$, are the zeros. Models are called stationary if all poles are within the unit circle and invertible if all zeros are within it. A shorthand notation for an ARMA model is

$$A(z)x_n = B(z)\varepsilon_n. \tag{2}$$

Assume that the data represent a stationary stochastic process. The power spectrum $h(\omega)$ of the ARMA(p,q) process is completely determined by the parameters in (1), together with the innovation variance σ_ε^2, and is given by

$$h(\omega) = \frac{\sigma_\varepsilon^2}{2\pi}\frac{\left|B(e^{j\omega})\right|^2}{\left|A(e^{j\omega})\right|^2} = \frac{\sigma_\varepsilon^2}{2\pi}\frac{\left|1 + \sum_{i-1}^{q} b_i e^{-j\omega i}\right|^2}{\left|1 + \sum_{i-1}^{p} a_i e^{-j\omega i}\right|^2}. \tag{3}$$

The transform of an AR(p) model into a positive semidefinite autocovariance function $r_{\mathrm{AR}}(k)$ is made with the Yule–Walker equations [12]. The complete autocovariance function can be written in a compact style as a function of the parameters of (1) as [3]

$$r(k) = \sum_{m=-q}^{q}\left[r_{\mathrm{AR}}(k+m)\sum_{i=0}^{q} b_i b_{i+|m|}\right] \qquad \forall k. \tag{4}$$

The ARMA autocovariance function can be written as a convolution of the separate autocovariances of the AR and MA parts. The autocorrelation $\rho(k)$ is found by dividing $r(k)$ by $r(0)$. The method of (4) in deriving the autocovariance function from the parameters is available in the ARMA Spectral Analysis (ARMASA) Toolbox [14] that also contains the ARMAsel algorithm. This Matlab program estimates the parameters of (1) for a large number of candidate model orders and types and automatically selects the best model for the data at hand.

The prediction error (PE), as a measure for accuracy, allows a mutual comparison of AR, MA, and ARMA models. It is defined as the squared error of prediction when applying the estimated model to another independent realization of the same true process. It is estimated by using the residual variance with a correction term. In simulations, knowledge of the true process parameters is available. This can be used in the model error (ME) [3]. This is a normalized and scaled version of the PE. With $A(z)$ and $B(z)$ being the true ARMA(p, q) process polynomials, $\hat{A}(z)$ and $\hat{B}(z)$ denoting the estimated ARMA(p', q') model, and N denoting the number of observations, the ME is defined as

$$\text{ME} = \text{ME}\left[\frac{\hat{B}(z)}{\hat{A}(z)}, \frac{B(z)}{A(z)}\right] = N\left(\frac{\text{PE}}{\sigma_{\varepsilon}^2} - 1\right). \tag{5}$$

By taking $q' = 0$, the estimated ARMA(p', q') model becomes AR(p'), and $p' = 0$ gives a pure MA(q') model. The minimal expectation of the ME for unbiased efficiently estimated models is the Cramér–Rao lower bound, and it is independent of the sample size N. The expectation of the ME is asymptotically equal to the number of estimated parameters in unbiased models, with at least all truly nonzero parameters included. An efficient expression for the computation of the ME as an objective accuracy measure in the time domain requires only the knowledge of the process and model polynomials [3].

A quality measure for measured data is the ratio between the variance of the data and the variance of the residuals of the selected model. The residuals are the unexplained part in (1), and they will be dependent on outliers. The power gain P_g is defined as the quotient of the variances of output and excitation of the time series process in (1) or as the quotient of the signal variance σ_x^2 and the residual variance $\sigma_{\hat{\varepsilon}}^2$ for estimated models. It is given by

$$P_g = \frac{\sigma_x^2}{\sigma_{\hat{\varepsilon}}^2}. \tag{6}$$

Outliers will cause a smaller value of P_g because the variance of the residuals increases if outliers are present.

III. Turbulence Data

This paper treats an example of turbulence data obtained by solving the Navier–Stokes equations on a supercomputer [20], [21]. It required a total of 15 500 CPU hours, which is still 16 days when it runs on 40 parallel CPUs. Therefore, verification of the internal consistency of the generated signal during the temporal storage of data for the shared-memory operation is a problem. The time has been normalized with T_{St}, the dimensionless Strouhal time, which, in turn, is given by $1/f_{St}$. The Strouhal number f_{St} is a measure for the vortex-shedding frequency in turbulent flow [21].

Fig. 1 shows $N = 3810$ observations of the turbulence signal. The original signal had 76 200 observations, but the spectral density was more than 80 dB lower in the last 95% of the frequency range [22]. This irrelevant part of the frequency range was removed by downsampling with a factor of 20 by simply deleting 19 of every 20 observations. Decimating the data with standard antialiasing filters was not appropriate here [22]. The steep spectral slopes of the filter would be within the frequency range of the final signal. This steep slope requires very high-order time series models for the filtered signal, which creates spurious details in the rest of the frequency range after filtering [22]. Downsampling folds back the high-frequency contents of the spectrum to the remaining frequency range. It can only be an acceptable solution if the power of the signal that is above half of the new resampling frequency is (almost) negligible. Antialiasing filters leave the first part of the spectrum undisturbed but give a very strong distortion at the frequencies that are filtered out. The distortion is already obvious for frequencies below the cutoff frequency of these filters. The performance of antialiasing filters in combination with time series analysis can be improved. It has been verified that the smallest distortion of the spectrum is found by taking the cutoff frequency of the antialiasing filter at 95% of the remaining frequency range. To demonstrate this, a white noise signal has been decimated with antialiasing filters. Time series modeling only selects a flat white noise spectrum with ARMAsel [14] for the decimated white noise signal if the cutoff frequency was chosen at 0.95 times half the resampling frequency. Other choices of the cutoff frequency gave a colored spectrum for the model selected from the wideband decimated white noise. It should be realized that all other properties of the original signal are strongly disturbed by the filtering; only the spectral density until the cutoff frequency is left the same. As an example, outliers that are obvious or recognizable in the original data are not visible or detectable in decimated filtered data because the high frequencies are filtered out.

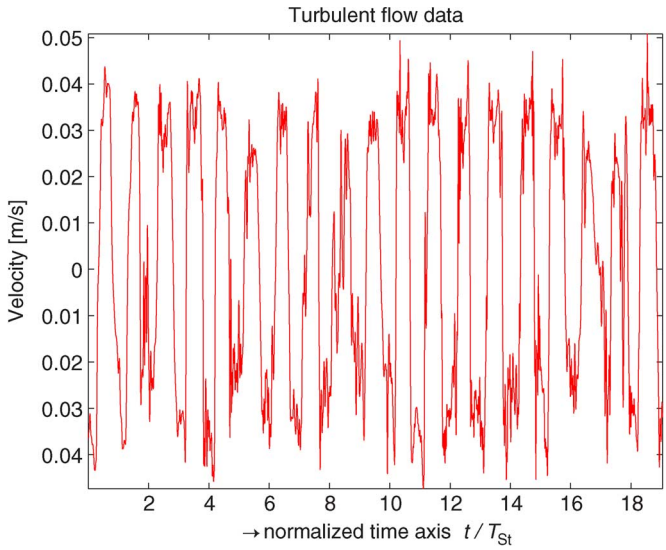


Fig. 1. Turbulent flow data obtained with computational fluid dynamics. The original data have been downsampled with a factor of 20 without damage. Antialiasing filters would severely influence the spectrum; simple downsampling does much less. The timescale has been normalized with T_{St}, which is the Strouhal time for the vortex-shedding period.

The signal in Fig. 1 shows some irregular periodicity because the zero crossings are not equidistant. Moreover, it is obvious that the signal is not normally distributed. A normal distribution would have its maximal density around the mean value, and the

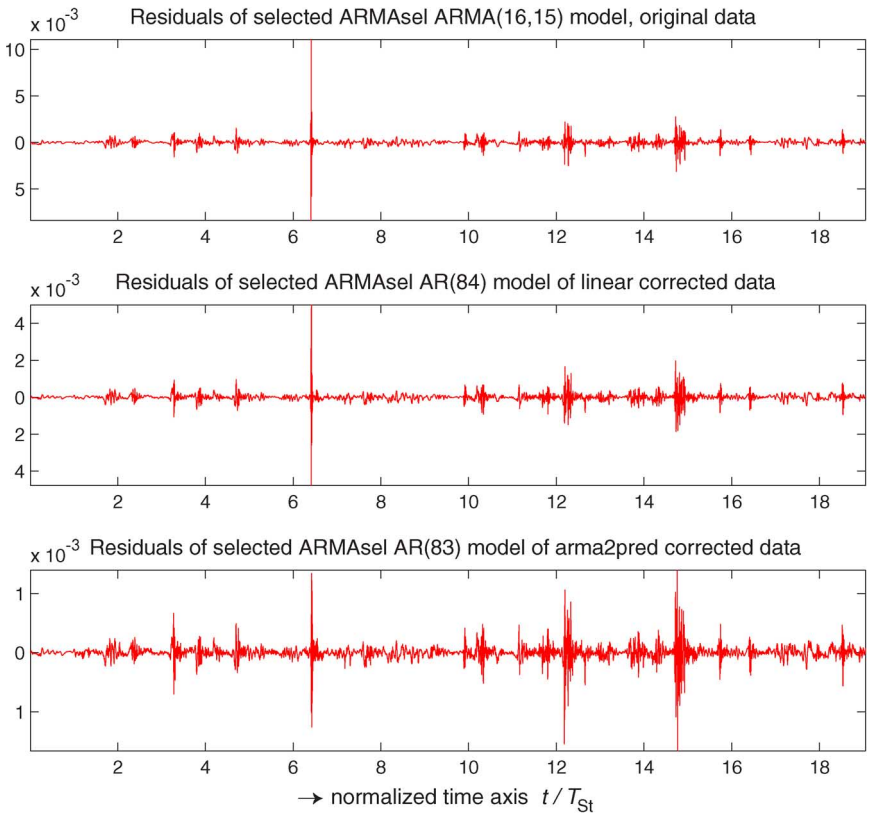


Fig. 2. Residuals of the selected time series model of the original downsampled data and of the data where one erroneous observation at $t/T_{St} = 6.4$ has been replaced by its linear interpolation and finally replaced by the best time series prediction based on the observations around that time.

density would gradually decrease for greater deviations from the mean. The signal in Fig. 1 has only a few observations around the mean and more at $+0.03$ and -0.03 m/s. The autocorrelation function and the power spectral density do not contain all the relevant statistical information of the given data. Higher order moments will also contain valuable information for this signal. Only the second-order moments will be analyzed here with ARMAsel [14]. This paper shows the possibilities of the automatic time series analysis to practical data, even if they are not normally distributed.

IV. Residuals of ARMAsel

The ARMAsel program has been applied to find the best time series model for the downsampled data. The automatic ARMAsel program selected the AR(16, 15) model for these data. The power spectral density of the selected model will be shown later. The residuals of the ARMA(16, 15) model for the measured data are shown in the upper plot of Fig. 2. It is obvious that a very irregular behavior is found at $t/T_{St} = 6.4$, which is the observation x_n for $n = 1280$. This single observation is an outlier, which is probably caused by one of the interruptions of the turbulence computations where the shared memory of the supercomputer requires data swapping. The residual plot shows the capacity of the automatic ARMAsel program to detect outliers. Studying the residuals of the selected model definitely traces this outlier that is much greater than all other irregularities in the residual signal.

The best solution in dealing with outliers is to completely remove them and to treat the remaining signal as a missing-data problem [23]. This requires no assumptions at all. A powerful algorithm based on the ARMAsel program for stationary stochastic data is available for missing-data problems, including automatic order selection [3]. This algorithm still gives accurate spectra close to the Cramér–Rao lower bound if less than 10% of the data is missing [23]. However, it very much requires computing time, and it is sometimes difficult to estimate models with more than about 20 AR parameters. If the missing fraction is very small, like one outlier on 3810 observations, it may be accurate enough to use some sort of interpolation.

Based on experience with missing data, two different methods have been investigated to replace the outlier by a corrected observation. The first method is linear interpolation, which has a disappointing accuracy in missing-data problems [23]. The outlier is replaced by the average of the preceding and the following observations. ARMAsel was applied to the new signal after the correction of the outlier. It selected the AR(84) model for the corrected data. This model was used to again compute the residuals after the linear interpolation of the outlier. This reduces the residual error variance with a factor of 3. It makes the largest corrected error more than a factor of 2 smaller, as shown in the middle plot of Fig. 2. However, it is obvious that the residual at $t/T_{St} = 6.4$ remains, by far, the largest. It should be possible to find a better correction.

The experience with missing data [23] learns that a better replacement is found with an algorithm that is based on the

iterative expectation–maximum (EM) principle. The E-step calculates the conditional expectation of the missing data with a model; the M-step computes maximum likelihood estimates from the combined observed and reconstructed data. This can be iterated, and it can be interpreted as a method to minimize the PE of the missing data [2]. The single outlier turns out to have a significant influence on the estimated time series parameters. All estimation methods minimize, in some sense, the sums of squared residuals. Therefore, the largest residual has a significant influence if it is very much larger than the others. The erroneous observation at $t/T_{St} = 6.4$ is observation x_n for $n = 1280$. It can be approximated by the prediction that can be made with the ARMA(16,15) time series model that has been estimated for the given data, with the outlier, from the observations from $n = 1$ until 1279. The prediction is made with the ARMA2pred algorithm from the ARMASA Toolbox [14]. The program ARMA2pred is based on a state-space representation of the time series model in Kalman filtering notation, which was originally developed for the calculation of the likelihood function for missing-data problems [24]. Likewise, a backward prediction is made from all observations with $n > 1280$. Finally, the average of these two predictions is taken to replace the outlier at $n = 1280$. It is the average because the estimated accuracy of the two predictions is equal. If two or more consecutive observations are outliers, the same two-sided method can be used with weighting factors based on the covariance matrix of the predictions, which is also computed in the ARMA2pred algorithm. It is the E-step of an EM algorithm.

ARMAsel selected the AR(83) model for the data with the ARMA2pred correction at $n = 1280$. The residuals of this AR(83) model are shown in the lower plot of Fig. 2. The difference between the analyzed signals is only the replacement of the observation at $n = 1280$ by its two-sided predicted value. After that, the magnitude of all residuals is reduced to about 40% of the upper part of residuals, with the outlier included. Therefore, replacing the single outlier gives a much better time series model for the rest of the data, with much smaller residuals over the whole signal. The SD of the measured signal is 0.0270, the SD of the residuals in Fig. 2 is 0.00039, 0.00023, and 0.00013, respectively. The values of the P_g measure of (6) of the data with an outlier, linear interpolation, and ARMA2pred correction are 5.1 10^3, 1.5 10^4, and 4.8 10^4, respectively. Therefore, replacing one outlier by its ARMA2pred prediction reduces the residual variance with a factor of about 10.

The automatic canonical correlation analysis (CCAsel) subspace procedure [19] has been applied to the same data. Unfortunately, it will fail to converge if all 3810 observations were used. It has been verified that the lack of convergence of the automatic subspace algorithm also frequently occurred with other data sets. However, using only the first 3000 observations of the turbulence signal gives a result of the automatic subspace algorithm CCAsel, with the ARMA(7, 7) model selected. The SD of the residuals was 0.00024, and P_g was 1.4 10^4 then—close to the performance of linear interpolation and much worse than that of ARMAsel. The candidates of this automatic subspace algorithm are only ARMA(p, p) models. This is a problem because, here, the best model was AR(83), which has been selected with ARMAsel for the turbulence data. The subspace ARMA(p, p) model would require 166 parameters for an exact unbiased description of the AR(83) model, with 83 superfluous MA parameters. The minimal ME value of (5) would become $2p$ for unbiased subspace models of a true AR(p) process, which is a factor of 2 above the Cramér–Rao lower bound for unbiased models. Moreover, order selection for subspace methods can become difficult if the true process is a high-order pure AR or MA process because too many insignificant parameters disturb the performance of order selection criteria. Therefore, the order of the selected ARMA(7, 7) model is really very low here. Convergence problems, limited ARMA(p, p) candidates for order selection, and the existence of example processes for which the subspace method gives poor results [19] have never been found for the ARMAsel algorithm. At the moment, this leads to a definite preference for ARMAsel in automatic signal processing and spectral analysis. The same conclusion has been obtained with a comparison of automatic model selection methods [19].

The residual at $n = 1280$ could still be made smaller by iterating the EM estimation with the ARMAsel algorithm. Then, the new AR(83) model for the corrected data would be used for an improved prediction and correction. It can be a good solution if more outliers are present or if the outcome of the first iteration is not satisfactory. This can happen if a very large outlier is present in the original data. However, this EM iteration has not been used here because the new residual at $n = 1280$ is already of the same level as the largest residuals for the given data at other times in Fig. 2. One iteration step already fulfilled all sensible requirements for the correction of the outlier. A second iteration may be necessary if the outlier is much greater.

The rather large residuals at $t/T_{St} = 12.18$ and $t = 14.765$ are not related to visible irregularities in the densely sampled signal. These residuals are of the same magnitude as the corrected residual at $t/T_{St} = 6.4$ in the lower plot of Fig. 2, and outliers of this size cannot be detected. Nevertheless, if they are synchronized with swapping of the memory of the supercomputer, they could also be indications of inaccuracies in the data generation.

Although all computations have been made with the downsampled signal, the full densely sampled signal is also available here. This can be used to visualize the signal around $t/T_{St} = 6.4$, which agrees with observation number $n = 1280$. Fig. 3 gives the segment with the outlier in the original densely sampled signal before downsampling. This demonstrates, without any doubt, that the observation at $n = 1280$ was an outlier. The fortuitous availability of a more densely sampled signal makes sure that there really is an outlier in those practical data. The difference between a linear interpolation of the observations at $n = 1279$ and 1281 and the ARMA2pred correction is only small, as can be concluded from Fig. 3. This demonstrates the sensitivity of the residuals to very small deviations. The computed SD of the ARMA2pred prediction at $n = 1280$ was about 0.0003.

It is evident in Fig. 3 that the original observation at $t/T_{St} = 6.4$ is not a very clear outlier after downsampling. In fact, the differences with the neighboring observations are not much greater than the usual differences between neighbors

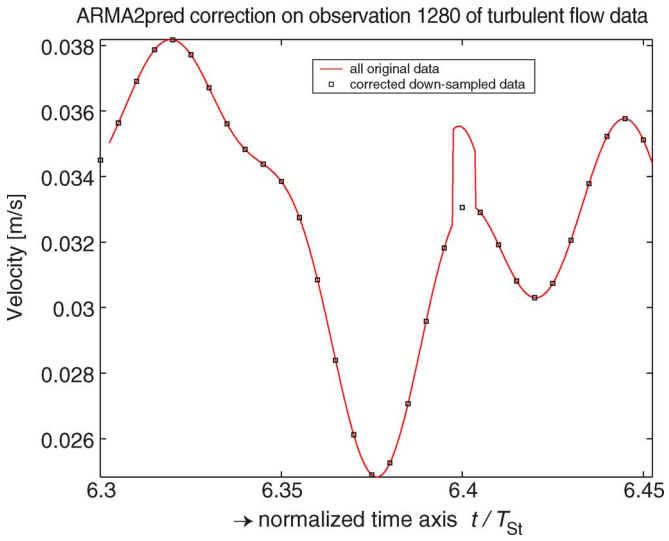


Fig. 3. Original turbulence observations before downsampling and the down-sampled data corrected with the ARMA2pred prediction. The signal properties around $n = 1280$ or $t/T_{St} = 6.4$ are definitely different from the rest of the data. The correction is only slightly different from a linear interpolation between the two neighbors.

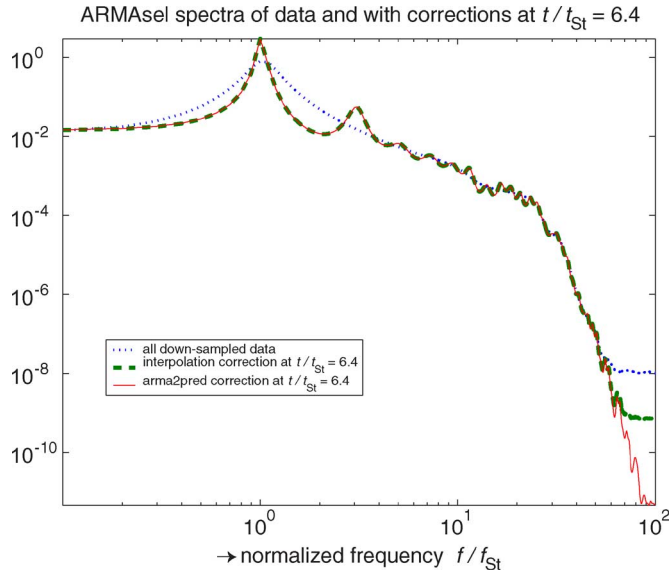


Fig. 4. Selected ARMA(16, 15) spectrum of the original data, the AR(84) spectrum of the data with linear interpolation, and AR(83) of the ARMA2pred correction. The outlier causes a strong spectral distortion in the normalized frequency range between 50 and 100.

in Fig. 3. Moreover, the peak at $t/T_{St} = 10.3$ in Fig. 1 and many other peaks in the downsampled signal are at least as sharp as the outlier at $t/T_{St} = 6.4$, without giving the large residuals. In other words, the outlier cannot be detected in the downsampled data by a visual inspection, without detecting many false alarms. This demonstrates that an inspection of the residuals of a selected time series model is a powerful method to locate irregularities in data that are generated by a supercomputer with a shared-memory.

V. Accuracy of ARMAsel

The spectra of the selected models for the data with an outlier, which were corrected with linear interpolation and with ARMA2pred correction, are given in Fig. 4. The correction has a strong influence on the spectrum for frequencies in the normalized range from 50 to 100, which is more than half of the linear frequency range. The difference at the end is more than a factor of 1000. The correction with ARMA2pred reveals more high-order spectral details than linear interpolation here, with a large dynamic spectral range in this example.

Three different signals have been analyzed with the ARMAsel program, with the correction at $n = 1280$ being sole difference. The estimated model accuracies of all estimated time series candidate models for two of these signals are given in Figs. 5 and 6. The global shapes of both figures are similar. The most important difference between these two figures is the vertical scale. The second figure has about ten times smaller values for the accuracy on the vertical axis. These accuracies are squared PEs. This means that the SD of the errors is about three times smaller after the correction. This approximately agrees with the differences in amplitude of the residuals, which had already been found between the upper and lower plots in Fig. 2.

ARMA(16, 15) is selected in Fig. 5. After replacing the single outlier x_n at $n = 1280$ by its time series prediction,

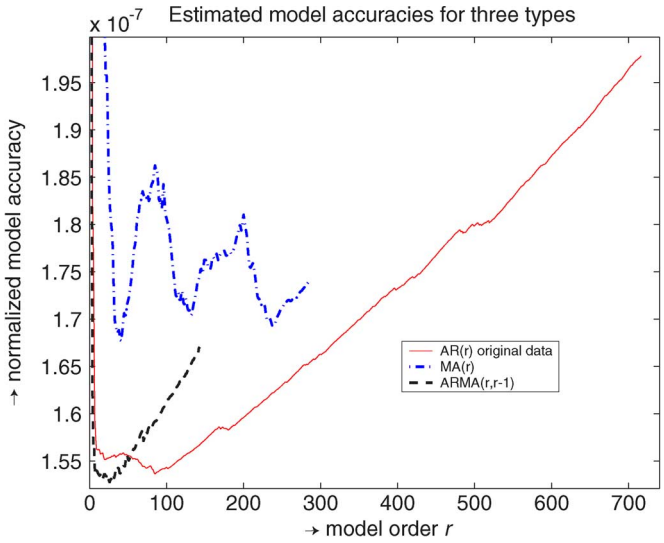


Fig. 5. Estimated model accuracies of all the models estimated with the ARMAsel program for $N = 3810$ downsampled observations of the turbulence data with outlier. ARMA(16, 15) is selected with the normalized model accuracy 1.53 10^{-7}.

AR(83) is selected in Fig. 6, but the ARMA(28, 27) model would have been a good candidate, with almost the same spectrum. MA models are not attractive for these data. It is remarkable that a small modification of a single observation can have such a strong influence on the estimated model accuracies if the spectrum extends over ten decades of magnitude. The explanation is that a constant is added to the spectrum by the outlier, which causes the flat distortion of the spectrum in Fig. 4 at the high frequencies with very low power.

In stationary ARMA(p, q) data, the expected model accuracy first becomes better for increasing model orders of the three

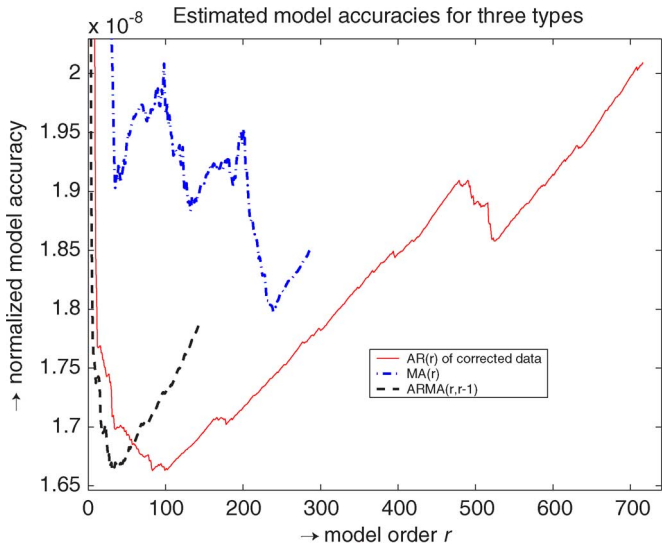


Fig. 6. Estimated model accuracies of all the models estimated with the ARMAsel program for $N = 3810$ downsampled observations of the turbulence data, with the ARMA2pred correction for the outlier at $n = 1280$. AR(83) is selected with the normalized model accuracy 1.67 10^{-8}.

types until minima are found in Fig. 5 or 6. Afterward, the accuracy will become worse for all types. In theory, the ME will linearly increase with the order for overcomplete models with too many parameters included [3]. This is seen for the AR and ARMA models in Fig. 5. The slope of ARMA is twice the slope of AR because the ARMA$(r, r-1)$ model has two extra parameters for one order higher. This behavior is also found in simulation experiments where it is known that there are no deviations from the stationary character of the signal. A remarkable local dip in the AR models at order 517 in Fig. 6 can be attributed to the large residuals at $t/T_{St} = 12.18$ and 14.765 in Fig. 2. These are the observations $n = 2436$ and 2953, with the same difference of 517 as the AR order with the dip. Obviously, something irregular happened at that distance, which becomes important for AR models of orders greater than 517. The continuation of the AR model accuracy in Fig. 6 for orders above 550 has the same slope of the line for orders between 100 and 500, but it is shifted downward. It becomes much more significant in Fig. 6 than in Fig. 5 after correcting the outlier at $n = 1280$. Apart from the automatic selection of the model with the best spectrum for the given data, the ARMAsel program [14] gives a lot of additional information in Fig. 5 or 6 that can be used to trace peculiarities in the data.

It is remarkable that the estimated model accuracy of all models with more than ten estimated parameters in Fig. 6 is about a factor of 10 smaller than the accuracy of the models in Fig. 5. The single outlier has a very significant influence on the spectra of all the estimated models of all types and orders.

VI. Adding Deliberate Errors

The signal with the correction of the time series prediction can be used as a prototype signal. One or more outliers can be added to this corrected signal to verify the quality of the outlier detection. The same distorted spectrum of the downsampled signal in Fig. 4, with the additional outlier, is also found by adding an error of 0.002 to any arbitrary observation of the corrected data. This error value happens to be the amplitude of the outlier in Fig. 3. The distorted spectrum is given by the original spectrum (without an outlier) plus a constant. This constant is about 10^{-8} in Fig. 4, with a P_g value of about 5000. Adding an error of amplitude 0.02 at some place gives a constant spectral level at 10^{-6}, with a P_g value of about 620, distorting the spectra above the normalized frequency 40. An error of 0.2 gives a constant of 10^{-4}, with a P_g value of about 30, distorting the spectra above the normalized frequency of 25. Therefore, the spectral distortion due to an outlier is characterized by the addition of a constant to the spectrum. The influence is only strong in the frequency range where the original spectrum is of the same magnitude as the constant. Spectral details in the rest of the frequency range are almost completely undisturbed by adding this small constant error. ARMAsel selected ARMA(16, 15) for those data with one deliberate additional error, like for the original data with an outlier. This model gives a spectrum that is very similar to the spectrum of all original data, where ARMA(16, 15) was also selected. Without an outlier, the AR(83) model is selected, which shows more details in the high-frequency range and also gives significant details around the normalized frequency of 3.

It is also possible to add more outliers and deliberate errors at the same time. Adding four errors of amplitude 0.01 has about the same influence as adding one error of size 0.02. Both situations have the same error energy. The constant level in the spectrum with those four errors was around 10^{-6}, with a P_g value of about 620. ARMAsel would have selected the ARMA(14, 13) model if four outliers were added. The four errors could be corrected with the ARMA2pred predictions for each individual outlier. Therefore, it has been verified that the correction method is also applicable to multiple outliers.

VII. Conclusion

The ARMAsel algorithm estimates the parameters of time series models, automatically selects a single model, and provides additional information about the model accuracies. The plots of the residuals of the automatically selected model and of the accuracies of all the estimated models contain a lot of detailed information about the data. Outliers that are invisible in the data themselves can be visually detected in the plot of the residuals, for signals with a high power gain. ARMAsel opens new perspectives for a refined analysis of random data.

Linear interpolation between neighbors is much less efficient for outlier correction than prediction with an estimated time series model. This paper has shown that the application of time series analysis to data that are not normally distributed can reveal different interesting details that cannot be easily found with other methods. It gives a useful tool to analyze the influence of interruptions in dynamic numerical simulation computations with supercomputers with data swapping.

Only one iteration of the prediction is sufficient for the correction of the outlier. If a better accuracy would be desired, repeated iterations can be applied.

REFERENCES

[1] M. B. Priestley, *Spectral Analysis and Time Series*. London, U.K.: Academic, 1981.

[2] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1999.

[3] P. M. T. Broersen, *Automatic Autocorrelation and Spectral Analysis*. London, U.K.: Springer-Verlag, 2006.

[4] K. Vaage, "Detection of outliers and level shifts in time series: An evaluation of two alternative procedures," *J. Forecast.*, vol. 19, no. 1, pp. 23–37, 2000.

[5] R. S. Tsay, D. Pena, and A. E. Pankratz, "Outliers in multivariate time series," *Biometrika*, vol. 87, no. 4, pp. 789–804, Dec. 2000.

[6] S. Frosch Møller, J. von Frese, and R. Bro, "Robust methods for multivariate data analysis," *J. Chemom.*, vol. 19, no. 10, pp. 549–563, 2005.

[7] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-based detection and prediction of outliers," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 2, pp. 145–160, Feb. 2006.

[8] R. K. Pearson, "Outliers in process modeling and identification," *IEEE Trans. Control Syst. Technol.*, vol. 10, no. 1, pp. 55–63, Jan. 2002.

[9] J. P. Burg, "Maximum entropy spectral analysis," in *Proc. 37th Meeting Soc. Exploration Geophysicists*, 1967. 6 pp.

[10] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[11] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[12] S. M. Kay and S. L. Marple, "Spectrum analysis—A modern perspective," *Proc. IEEE*, vol. 69, no. 11, pp. 1380–1419, Nov. 1981.

[13] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NJ: Prentice-Hall, 2005.

[14] P. M. T. Broersen, *Matlab ARMASA Toolbox*. under signal processing, spectral analysis. [Online]. Available: www.mathworks.com/matlabcentral/fileexchange

[15] B. Porat, *Digital Processing of Random Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1994. p. 144.

[16] M. Viberg, "Subspace-based methods for the identification of linear time-invariant systems," *Automatica*, vol. 31, no. 12, pp. 1835–1851, Dec. 1995.

[17] P. van Overschee and B. de Moor, "A unifying theorem for three subspace system identification algorithms," *Automatica*, vol. 31, no. 12, pp. 1853–1864, Dec. 1995.

[18] D. Bauer, "Asymptotic properties of subspace estimators," *Automatica*, vol. 41, no. 3, pp. 359–376, Mar. 2005.

[19] D. Bauer and S. de Waele, "A finite sample comparison of automatic model selection methods," in *Proc. 13th IFAC Symp. Syst. Identification*, 2003, pp. 1790–1795.

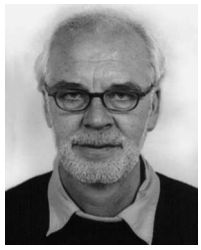
[20] M. P. Sobera, C. R. Kleijn, H. E. A. van den Akker, and P. Brasser, "Convective heat and mass transfer to a cylinder sheathed by a porous layer," *AIChE J.*, vol. 49, no. 12, pp. 3018–3028, 2003.

[21] M. P. Sobera, C. R. Kleijn, and H. E. A. van den Akker, "Subcritical flow past a circular cylinder surrounded by a porous layer," *Phys. Fluids*, vol. 18, no. 3, pp. 1–4, 2006. Art. 038106.

[22] P. M. T. Broersen, "ARMAsel for identification of univariate measurement data," in *Proc. IEEE IMTC Conf.*, Sorrento, Italy, 2006, pp. 107–112.

[23] P. M. T. Broersen, S. de Waele, and R. Bos, "Autoregressive spectral analysis when observations are missing," *Automatica*, vol. 40, no. 9, pp. 1495–1504, Sep. 2004.

[24] R. H. Jones, "Maximum likelihood fitting of ARMA models to time series with missing observations," *Technometrics*, vol. 22, no. 3, pp. 389–395, Aug. 1980.



Piet M. T. Broersen was born in Zijdewind, The Netherlands, in 1944. He received the M.Sc. degree in applied physics and the Ph.D. degree from Delft University of Technology, Delft, The Netherlands, in 1968 and 1976, respectively.

He is currently with the Department of Multiscale Physics, Delft University of Technology. He developed statistical measures to let measured data speak for themselves as a practical solution for the spectral and the autocorrelation analysis of stochastic data. His main research interest is in the automatic and unambiguous identification of the character of stationary random data.