Time-Series Analysis if Data Are Randomly Missing

Piet M. T. Broersen and Robert Bos

***Abstract*—Maximum-likelihood (ML) theory presents an elegant asymptotic solution for the estimation of the parameters of time-series models. Unfortunately, the performance of ML algorithms in finite samples is often disappointing, especially in missing-data problems. The likelihood function is symmetric with respect to the unit circle for the estimated zeros of time-series models. As a consequence, the unit circle is either a local maximum or a local minimum in the likelihood of moving-average (MA) models. This is a trap for nonlinear optimization algorithms that often converge to poor models, with estimated zeros precisely on the unit circle. With ML estimation, it is much easier to estimate a long autoregressive (AR) model with only poles. The parameters of that long AR model can then be used to estimate MA and autoregressive moving-average (ARMA) models for different model orders. The accuracy of the estimated AR, MA, and ARMA spectra is very good. The robustness is excellent as long as the AR order is less than 10 or 15. For still-higher AR orders until about 60, the possible convergence to a useful model will depend on the missing fraction and on the specific properties of the data at hand.**

***Index Terms*—Autocorrelation analysis, autoregressive moving-average (ARMA) model, incomplete data, missing observations, order selection, spectral analysis.**

I. Introduction

AN EFFICIENT, numerically stable, and simple algorithm for spectral analysis when data are missing would be a useful tool for signal processing in many areas of science and technology. Practical observations are often incomplete, because sensor failure or outliers cause missing data. Discarding all data records with spurious or missing observations may be expensive or impossible. Sometimes the only information that is available is a missing-data record. In meteorological, astronomical or satellite observations, the weather conditions may disturb the equidistant sampling scheme. In paleoclimatic data, the relation between the chronological time and the physical depth gives a time-base distortion, which is the cause that an observed time series has missing observations on an equidistant time grid [1].

An easily applicable spectral estimator for missing data is the method of Lomb [2]. This computes Fourier coefficients as the least squares fit of sines and cosines to the available remaining observations. The Lomb–Scargle spectrum is accurate in detecting strong spectral peaks, but this assumption biases the description of slopes and background shapes in the spectrum [3], [4].

A second group of methods relies on estimation algorithms that have been developed for uninterrupted equidistant data. First, they reconstruct the missing data with linear, cubic, or spline interpolation or with sample-and-hold or nearest neighbor resampling. That is followed by the estimation of the spectral density from the reconstructed uninterrupted signal. These methods can only give accurate or acceptable results for very small missing fractions [4].

A third type of estimators fits a time-series model directly to the available observations. An exact maximum-likelihood (ML) approach using Kalman filtering has been described for missing data [5]. Also, an approximate method has been developed for autoregressive (AR) models with good practical results [4]. The methods that have been tested for AR will be extended to moving-average (MA) and combined autoregressive moving-average (ARMA) estimation in this paper.

The layout of the paper is as follows. Time series are described as a parametrical description of the power-spectral density and of the autocorrelation function of measured data. The main question in this paper is whether it is possible to estimate reliable and accurate MA and ARMA models when data are missing. Some choices for a numerically stable and efficient ML algorithm are described for MA and ARMA models. However, the results of ML estimation for MA and ARMA are often disappointing in simulations. Therefore, a reduced-statistics (RS) algorithm is developed that uses the parameters of an intermediate estimated AR model to compute MA and ARMA models. This AR model is first estimated from the missing-data observations with an ML algorithm. Attention is given to the selection of appropriate MA- or ARMA-model orders.

II. Time-Series Models

ARMA models describe the characteristics of stationary stochastic processes [6]. The power spectrum and the autocovariance function are determined completely as functions of the estimated parameters of the ARMA model. An ARMA(p, q) model can be written as

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

where ε_n is a purely random white-noise process with zero mean and variance σ_ε^2 [6]. It is a pure AR model if $q = 0$, and

Manuscript received June 15, 2004; revised August 30, 2005.

P. M. T. Broersen is with the Department of Multi Scale Physics, Delft University of Technology, Delft, The Netherlands (e-mail: broersen@tnw.tudelft.nl).

R. Bos is with the Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands (e-mail: r.bos@dcsc.tudelft.nl).

Digital Object Identifier 10.1109/TIM.2005.861247

MA for $p = 0$. Almost any stationary stochastic process can be described as a unique AR(∞) or MA(∞) or ARMA(p, q) process. In practice, finite orders are sufficient. The power spectrum $h(\omega)$ of the ARMA(p, q) model (1) is given by [6]

$$h(\omega) = \sigma_\varepsilon^2 \left| \frac{B(\mathrm{e}^{j\omega})}{A(\mathrm{e}^{j\omega})} \right|^2 = \sigma_\varepsilon^2 \left| \frac{1 + b_1 \mathrm{e}^{-j\omega} + \cdots + b_q \mathrm{e}^{-j\omega q}}{1 + a_1 \mathrm{e}^{-j\omega} + \cdots + a_p \mathrm{e}^{-j\omega p}} \right|^2. \tag{2}$$

Also, the infinitely long true autocorrelation function is completely determined by the $p + q$ true parameters of (1). For MA(q) models, the finite autocorrelation length is given by [6]

$$\rho_n = \frac{\sum_{i=0}^{q-n} b_i b_{i+n}}{\sum_{i=0}^{n} b_i^2}, \qquad 0 < n \le q$$
$$\rho_n = 0, \qquad n > q; \quad \rho_{-n} = \rho_n \tag{3}$$

where ρ_n denotes the expectation $E(x_i x_{i+n})/\sigma_x^2$. The Yule–Walker relations [7] describe the complete autocorrelation function of an AR(p) process and also the autocorrelation for ARMA(p, q) for orders greater than the maximum of p and $q + 1$

$$\rho_n + a_1 \rho_{n-1} + \cdots + a_p \rho_{n-p} = 0$$
$$n \ge \max(p, q+1);\ \rho_{-n} = \rho_n. \tag{4}$$

For pure AR processes, (4) is valid for all $n > 0$. The complete ARMA autocovariance function for the first lags is easily derived [7]. Equations (2)–(4) can be used with the true process parameters as well as with the parameters of estimated models.

Reflection coefficients k_i are used to recursively determine the AR parameters $\hat{A}_m(z)$ of all model orders m between 1 and p, with the Levinson–Durbin formulas [7]

$$\hat{a}_1^1 = k_1$$
$$\hat{a}_i^m = \hat{a}_i^{m-1} + k_m \hat{a}_{m-i}^{m-1}, \qquad 1 \le i < m$$
$$\hat{a}_m^m = k_m, \qquad 1 \le m \le p \tag{5}$$

where $\hat{a}_i^m$ is the estimated parameter of order i in the model of order m. These relations have the property that all poles are inside the unit circle if and only if the reflection coefficients are all less than 1 in absolute value. This property is important for a robust algorithm for missing data.

The accuracy of estimated models is evaluated with the model error (ME). This is a relative measure in the frequency domain based on the integrated ratio of estimated and true spectra. Also, a time-domain expression for ME exists as a normalized prediction error (PE) [8]

$$\mathrm{ME} = N \left(\frac{\mathrm{PE}}{\sigma_\varepsilon^2} - 1 \right). \tag{6}$$

The PE is defined as the expectation of the mean-square error of the one-step-ahead prediction of an estimated model to new uninterrupted data of the same process. The multiplication with the number of observations N gives the ME an expectation that is independent of the sample size for unbiased models from uninterrupted data. It yields the number of parameters $p' + q'$ as the minimally obtainable expectation of the ME for unbiased estimated ARMA(p', q') models, with $p' \ge p$ and $q' \ge q$. When data are missing, the minimum value of the ME can sometimes be smaller, but it will generally be much greater than the number of estimated parameters [4].

III. MA and ARMA Estimators

A. Numerically Stable ML Estimator

Two different algorithms for the ML estimation of AR models have been described for missing-data problems [4], [5]. They can also be used for MA and ARMA models. The exact estimator [5] requires a computing time that is proportional to the sum of the number of available and missing data, which is equal to the total observation interval. The approximate estimator's time is proportional to only the available number of data [4]. The speed of the exact algorithm is higher if more than about 15% of the data remains; the approximate can be much quicker for very sparse data. To improve the numerical robustness, the algorithms build the estimated polynomials $A(\mathrm{e}^{j\omega})$ and $B(\mathrm{e}^{j\omega})$ in (2) with reflection coefficients k_i, like in (5). Then, the unconstrained optimization of $\tan(\pi/2 * k_i)$ gives estimated parameters for increasing orders p and q. This transform guarantees that the estimated k_i is always in the range $-1 < k_i < 1$. Hence, all AR, MA, and ARMA models computed by nonlinear numerical-optimization routines are stationary and invertible, with all the poles of $A(z)$ and zeros of $B(z)$ inside the unit circle [6].

B. Starting Values

The AR and the MA parts have separate starting values. As possible starting values for the nonlinear optimization of the ARMA(p, q) model, the reflection coefficients of the ARMA($p-1, q-1$) model have been considered. They are the reflections coefficients of the AR($p-1$) model and of the MA($q-1$) model, with additional zeros as the start for the new k_p and k_q. These starting values were successful in AR estimation [4]. However, this method fails completely for MA and ARMA. In estimating the ML parameter of an MA(1) model, that parameter will exactly be $+1$ if the correlation coefficient ρ_1 at lag 1 is greater than 0.5, and it would be -1 if $\rho_1 < -0.5$. This value 1 or -1 for k_1 also remains as a local minimum of the likelihood for higher order MA models. By adding an additional zero as the initial value for the second parameter in an MA(2) model, the MA(2) solution found by the nonlinear minimization always kept the first reflection coefficient at $+1$ or -1, if that was taken as the starting value for the optimization algorithm. In almost all simulation examples, it did not get out of this lower order local minimum. This behavior can be improved somewhat by taking 0.95 or 0.98 times the previous reflection coefficients as new starting values, which forces the initial search vector away from that local lower order minimum. However, this improved trick also often failed to converge to the global minimum. The same problem arises with the estimated zeros of an ARMA model.

In Monte Carlo simulations, it has been demonstrated that convergence to the global minimum really was a problem of choosing good starting values. Using the true parameters of the process as starting values for the ML estimator in simulations would often converge to a good model. However, knowledge of the truth is no option for practical data with unknown process parameters. No good recipe for practical starting values has been found in the literature or in our simulations.

ARMA models have a second problem in the ML method. If the true process has a pole and a zero that both have a rather small radius, they are almost canceling. The estimated model with one pole and one zero less would be more accurate than the estimated model of the true order. The likelihood will mostly converge then to a model with an almost-canceling pair of a pole and a zero, but their joint location has often a zero close to the unit circle in the complex plane, with a narrow spectral valley. For high-order ARMA models, there is an especially strong tendency to estimate such models. This happens if high-order models are estimated from measured observations where the true order is not known, which is usually the case in practice.

C. RS Estimator

A practical solution that has been found for the estimation of MA(q) or ARMA(p,q) models is the RS estimator [9]. This algorithm estimates MA and ARMA models from a limited number of estimated AR parameters, using the methods of Durbin [10], [11]. This RS MA(q) or ARMA(p,q) model is itself a good estimator for uninterrupted data [9]. However, if data are missing, it can also be used as a realizable starting point for the nonlinear ML optimization for that model. First, the best MA(q) or ARMA(p,q) model will be estimated and selected with the RS estimator in missing-data problems. Afterwards, a nonlinear ML optimization is carried out with that selected model as starting point. The model-quality ME of the RS model and the ME of the resulting ML model will be compared in this paper, to see how much the minimization of the likelihood function can improve the estimated model that was used as the recipe for the starting point.

D. Order Selection

Order selection for models estimated by likelihood minimization can be performed with a generalized information criterion (GIC) defined as

$$\text{GIC}(p+q,\alpha) = L\left(X;\underline{\hat{a}}_p,\underline{\hat{b}}_q,\hat{\sigma}_x^2\right) + \alpha(p+q). \qquad (7)$$

The model with the smallest GIC value is selected. In missing-data problems, N denotes the number of remaining observations. Only these N give a contribution to the likelihood; the fraction of missing data has no explicit influence in (7). In ML optimization, it is common to minimize the negative of the log-likelihood function, which is denoted L in (7). The best value for the penalty factor α for MA and ARMA estimation with missing data is investigated later in this paper. The famous Akaike's information criterion (AIC) is given by (7), with $\alpha = 2$ [6], [7]. In order selection for finite samples of uninterrupted data, $\alpha = 3$ gives better results [8]. For missing-data AR models, penalties α between 3 and 5 are good choices, depending on the missing fraction [12]. Penalty 3 was the best in simulations if less than 25% is missing, penalty 5 if more than 75% was missing, and penalty 4 for a missing fraction between 25% and 75%. The GIC criterion of (7) is used for the order selection of all estimation methods for missing data. If models are estimated with methods other than ML, e.g., with RS, the likelihood of that estimated model is computed afterwards for use in order selection.

IV. Fixed-Order Simulations

Several algorithms have been studied in simulations. If a total of N remaining observations is required, N/γ uninterrupted equidistant MA or ARMA observations are generated first, where γ denotes the remaining fraction and the missing fraction is given by $1-\gamma$. Those observations are transformed into a missing-data problem by randomly leaving out observations, keeping a fraction γ, which gives precisely N remaining observations. For each missing-data record, AR(p'), MA(q'), and ARMA$(r', r'-1)$ models have been estimated to show that it is possible to estimate all types of models from unknown data. The limitation to ARMA models with one order difference between AR and MA is computationally attractive. It is often used in spectral estimation for uninterrupted data if the order has to be selected [8], [9], and it has no serious influence on the accuracy of the selected models.

For the simulations with AR and MA processes, the true parameters of the generating process are built from reflection coefficients with (5), with $k_m = \beta^m$. In this way, all poles or zeros of the generating process have the same radius β. ARMA processes are generated with the same rule for the AR part; the MA part is found using $k_m = (-\beta)^m$, which gives the true zeros in the complex plane at the same radius β, but at different angles.

The convergence of the ML algorithms is rather good for low-order AR models and for few missing data [4] and it may become a problem if higher orders must be calculated for a large missing fraction. Furthermore, the convergence of the likelihood is generally much better for AR models than for the other model types.

Fig. 1 shows a true ARMA(4, 3) spectrum, together with three estimates in a single simulation run with 500 remaining observations and $\gamma = 0.75$. The best fitting model in this example is the AR(4) estimate selected with (7), with $q = 0$ and $\alpha = 3$. The ARMA(4, 3) RS estimate obtained from the AR(6) model has a spurious small wide peak around $f = 0.4$. Using this RS model as the starting value for the ML procedure gives the ARMA(4, 3) ML model with the neighboring peak and valley in Fig. 1. The estimated and selected AR(4) ML spectrum is quite accurate for this example; the ARMA(4, 3) ML estimate shows a narrow peak and a narrow valley.

It has been observed in many examples that ML solutions tend to diverge if more data are missing, giving spurious peaks and valleys in the estimated spectra. It happens for rather low orders in MA and ARMA, but high-order AR models can also have this behavior. The likelihood is used in (7) for order

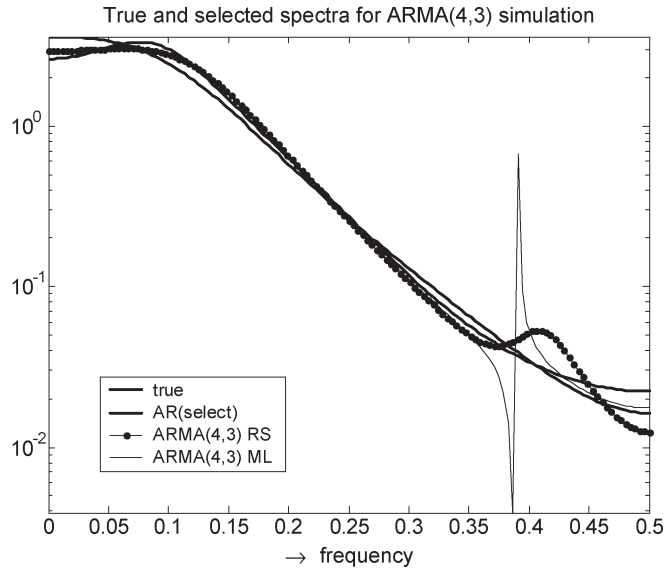


Fig. 1. True spectrum of the ARMA(4, 3) process and the estimated spectra of the selected AR(4), the ARMA(4, 3) RS estimated from the AR(6) model, and the ML estimate with starting values taken from the true process. All spectra are estimated from 500 remaining observations with 25% missing data, generated with $\beta = -0.5$. In this realization, the ME for AR(4) is 6.0, for ARMA(4, 3) RS is 11.4, and for the ARMA(4, 3) ML solution is 33.5.

selection. It is the smallest for the ARMA(4, 3) ML solution in the example of Fig. 1. Therefore, it would be selected with (7), but the ME quality (6) of that model is poor. Many other ARMA(4, 3) models and also the true ARMA(4, 3) process have a much better ME quality, although the value of the log likelihood is higher, especially for large missing fractions. In missing-data problems with a finite number of remaining observations, the relation between a small or minimum value of the log likelihood of ARMA models and a small ME is not as strong as it is when all data are available.

The occurrence of combined peaks and valleys is very common in higher order estimated ARMA models, if the ML optimization is used. It can also happen if no data are missing. If poles and zeros are distinct and statistically significant, ML estimation may be accurate. However, higher order ML models often give spurious peak–valley combinations that lead to a lower value of the likelihood but at the same time to a poor model quality. This is a serious problem in selecting the order of MA and ARMA models estimated with ML, because there is a strong tendency to select those questionable models.

V. Simulations With Order Selection

Without data missing, order-selection criteria are generally based on the log likelihood or on the decrease of the residual variance for higher model orders. If there are data missing, the value of the likelihood function is the only available basis for order selection with (7). Even for RS models that do not use the likelihood for estimation but a long AR model, the log likelihood has to be computed for use in the order-selection criterion (7).

In missing-data records, convergence and order selection have never been a problem with AR models for orders lower than 10 or 15. Sometimes, very-high-order AR models until order 60 could be estimated accidentally in a single simulation run, at considerable computational costs. It has been found that the GIC criterion of (7) sometimes selected a very high AR order in those situations. The selected model would then always have a number of strong peaks in the estimated spectrum for data where the true spectrum and lower order estimated AR models were smooth. For reasons of convergence, automatic order selection, and computation time, the maximum AR candidate order is taken to be 15 in the simulations.

The AR model order is selected as the order K with the minimum GIC (K, α). It has been shown that the best compromise penalty α for AR depends on the remaining fraction γ. This gives penalty $\alpha = 3$ for $\gamma > 0.75$, $\alpha = 5$ for $\gamma < 0.25$, and $\alpha = 4$ in the range between those limits [12]. The intermediate AR order to determine MA or ARMA models for uninterrupted data is chosen with a sliding window [8]. That order is then much higher than the selected order K of the best AR model. For AR models estimated from uninterrupted data, too-high model orders are only slightly less accurate than the model with the best order. However, it turns out that high-order AR models from missing-data problems can have very much greater values for the ME, and are poor representations for the character of the data. Therefore, the choice for the intermediate order has to be adapted for missing data to make sure that no spurious details are present in that AR model. The intermediate AR order to estimate MA or ARMA models is taken as the highest order P for which the ME in comparison with the selected AR(K) model is less than $2(P - K)$

$$P = \arg\max_{\hat{P}} \left[\text{ME}\left(\text{AR}(\hat{P}), \text{AR}(K)\right) \leq 2(\hat{P} - K) \right]. \quad (8)$$

The orders K and higher are candidates for P. This ensures that the spectra of the AR(K) and the AR(P) models are similar, because a difference of 2 in the ME for each extra parameter is rather small. This is a practical compromise between the high preferred orders for uninterrupted data [8] and the fact that high-order estimated AR models can become very inaccurate for small values γ of the remaining fraction.

Table I gives the average ME of models for simulations with an ARMA(2, 1) process, all with $N = 500$ remaining observations and six different values for the remaining fraction γ. Table II gives the ME of models estimated for MA(4) data with 250 remaining observations. The ME of two different AR models, two ML models, and five RS models estimated from a long AR model are compared in each table. All algorithms can also be applied to uninterrupted data; the accuracy for $\gamma = 1$ is given in the final columns of the tables for comparison.

1) The first two rows of the tables give the average ME in 100 runs of two ML AR models: the AR(K) model selected with (7) and the highest order AR(15) model that has been computed.

TABLE I
AVERAGE ME OF 100 MONTE CARLO SIMULATION RUNS, AS A FUNCTION OF THE REMAINING FRACTION γ. MODELS ARE ESTIMATED FROM $N = 500$ OBSERVATIONS OF AN ARMA(2, 1) PROCESS WITH $\beta = 0.4$

$\rightarrow \gamma$	.25	.5	.75	.9	.95	1
AR						
Selected AR model	29.2	13.6	7.6	6.4	6.1	5.6
AR(15)	72612	555.0	30.8	17.5	17.2	15.0
ARMA with ML						
ML ARMA(2,1) start true	1587	260.9	5.3	3.8	3.3	3.3
ML start sel $\alpha = 3$	1407	310.2	21.5	3.9	4.6	3.6
ARMA from AR(*P*)						
ARMA(2,1)	27.6	10.5	5.1	4.2	3.4	3.4
Selected with $\alpha = 2$	28.3	11.2	5.9	5.0	4.9	4.8
Selected with $\alpha = 3$	28.1	10.6	5.4	4.2	4.2	3.6
Selected with $\alpha = 4$	29.9	10.5	5.3	4.2	3.8	3.4
ARMA(2,1) from AR(15)	918.4	159.0	9.3	4.1	3.4	3.4

TABLE II
AVERAGE ME OF 100 MONTE CARLO SIMULATION RUNS, AS A FUNCTION OF THE REMAINING FRACTION γ. MODELS ARE ESTIMATED FOR $N = 250$ OBSERVATIONS OF A TRUE MA(4) PROCESS WITH $\beta = -0.8$

$\rightarrow \gamma$	.25	.5	.75	.9	.95	1
AR						
Selected AR model	72.0	42.5	25.6	18.0	15.3	13.1
AR(15)	133020	1774	56.0	22.0	18.4	15.6
MA with ML						
ML MA(4), start true	7676	4973	227.2	10.4	6.8	4.8
ML start sel $\alpha = 3$	34204	4822	507.9	28.0	8.2	5.4
MA from AR(*P*)						
MA(4)	63.7	30.0	14.8	7.2	5.7	4.5
Selected with $\alpha = 2$	66.3	35.6	17.9	9.6	8.2	5.8
Selected with $\alpha = 3$	67.2	35.0	16.8	8.6	6.6	4.9
Selected with $\alpha = 4$	68.6	34.6	16.7	8.0	6.1	4.8
MA(4) from AR(15)	4616	1161	32.0	6.3	5.1	4.4

2) Then, the ME of two different ML models is given. ARMA models are estimated in Table I for the ARMA(2, 1) data and MA models in Table II for MA(4) data. The first row gives the true-order model and is computed with the true process parameters as starting values. This is only possible in simulations where the truth is known and it is considered as the best possible starting value for the ML minimization. If it did not converge from these starting values, it never converged in our simulations. Also, a realizable ML solution is given, with the parameters of the RS model that was selected with penalty $\alpha = 3$ as the starting values and with the same order as that selected RS model. In this way, it can be verified whether postprocessing with an ML algorithm can improve the quality of the model estimated previously with the RS algorithm.
3) Then, the quality of five RS models follows. Four are estimated from the same AR(P) model, with an individually ordered P in every simulation run, selected with (8). The first and the fifth of these rows have the true model order, the others are selected with $\alpha = 2, 3,$ and 4, respectively. To illustrate the advantage of the AR(P) model as an intermediate AR model in comparison with a fixed high-order AR model, the ME of the true-order model computed from AR(15) is also given in the final row.

The results presented in the two tables are just representative illustrations of what has been found in numerous other simulations with different generating processes. Only conclusions that are supported by similar results in all other simulations are drawn. Therefore, the conclusions to be presented are more general than only for the examples in Tables I and II.

All AR models are estimated with the ML method of Jones [5]. The first two rows show that selecting the AR order is always better than taking order 15. The quality of the AR(15) model deteriorates strongly if more than 25% is missing. Generally, lower AR orders are selected from the same number of remaining observations if the missing fraction becomes greater.

The quality of the true-order ARMA(2, 1) and MA(4) models obtained from the intermediate AR(P) models is about the same as from AR(15) models if 10% or less of the data is missing. It is also close to the accuracy without missing observations. For more data missing, say $\gamma < 0.9$, the quality of the RS models obtained from the AR(P) model is the best. Therefore, selecting the intermediate order P with (8) gives good results in the simulations for all missing fractions.

The comparison of the three values for the penalty factor α in the RS estimates obtained from AR(P) definitely shows that penalty 2 is not the best. Hardly any preference for 3 or 4 can be given in those examples. However, it is known that penalty 4 can lead to an undesirable underfit in simulations with small true-parameter values with uninterrupted data. Based on this experience, the penalty 3 is advised. In contrast with AR order selection [12], there is no reason to make the penalty dependent on the missing fraction for MA and ARMA order selection.

Two different models have been estimated with ML. The difference between using the realizable starting values obtained from the RS estimate with the AR(P) model and selected with $\alpha = 3$ and using the true parameters as starting values for the ML minimization is small, if not too many data are missing. This is a strong indication that it is not possible to improve the quality of the ML algorithm by using still other starting values. The optimization converges mostly to the true minimum of the likelihood function.

However, the quality of the model belonging to the minimum of the likelihood is sometimes disappointing, especially if many data are missing. The poor quality of the ML solution for ARMA models is also illustrated in Fig. 1, where the true-order model with the smallest likelihood has a spurious peak–valley combination for $f = 0.39$. An extensive search did not deliver an example where the average ME quality of the ML model was much better than the average ME of the RS model obtained from the intermediate AR(P) model. That agrees more or less with the conclusion that has been obtained for uninterrupted data. There, it has been concluded that the RS estimators for MA and ARMA estimation were as accurate as the best of the other existing algorithms [9].

In the examples, the RS estimates obtained from the AR(P) model and selected with $\alpha = 3$ or 4 provide the best estimated ARMA and MA models. There is no reason to optimize the

likelihood function to estimate MA or ARMA models. The result of postprocessing that selected the model by using it as a starting value for an ML optimization does not improve the quality of the model. A smaller value of the likelihood does not give a better value of the model quality measured with the ME. This follows from the comparison of the second ML row with the third RS row in the tables.

The Cramér–Rao lower bound for the ME is given by the number of estimated parameters if all observations are available. For small missing fractions, several estimated models are close to that benchmark value, which is 3 and 4 in Tables I and II, respectively.

The quality of the selected RS model is good in comparison with the true-order RS model. Hence, if good models are among the estimated candidates, they will be selected with the described algorithm. Selection of the MA or ARMA order gives about the same quality as estimating the true-order model. Not knowing the true order is not a problem because the order can be selected with a very small loss of accuracy.

Taking a model of too-high order can have an enormously detrimental influence on the accuracy. This is clear from the ME of the AR(15) models for small γ in the second row of both tables. This is not the case in the RS MA and ARMA models of different orders, because these are all derived from the same intermediate AR(P) model, with P defined in (8).

The average quality of the selected MA or ARMA models in the presented examples is always better than the quality of the selected AR model. It was significantly better for small missing fractions. The difference becomes small if more data are missing. Therefore, MA and ARMA models are a welcome addition as candidate models for measured unknown data. Using only AR models may lead to less accuracy.

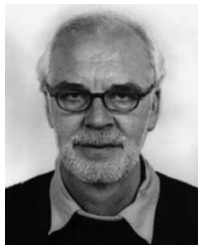
AR models can be estimated for very small values of γ, as long as γN is great enough [4]. This number γN can be seen as an effective number of observations because it gives approximately the number of pairs of two contiguous observations and of pairs separated by one missing observation and of pairs separated by any specific larger gap. The RS algorithm can always be used to investigate whether a better fit to the data can be found with MA or ARMA models.

VI. Concluding Remarks

AR models of increasing orders can be estimated with the ML approach. Starting values for AR are found by using the reflection coefficients of the previous model with an additional zero for the nonlinear optimization for the higher order. The AR models can also be used as input for an RS algorithm for MA and ARMA estimation. That turns out to be better for the accuracy of MA and ARMA models than the nonlinear optimization of the log-likelihood function. The minimum of the log-likelihood function is often found at MA or ARMA models with spurious spectral details. For each model type, examples have been given where that specific type gives the most accurate estimated spectrum. Therefore, it is advisable in missing-data problems to estimate AR models with the ML estimator and to estimate MA and ARMA models with the RS estimator. MA and ARMA model orders can be selected automatically with an order-selection criterion based on the afterwards-calculated likelihood of those candidate models, with penalty of 3.

References

[1] J. R. Petit *et al.*, "Climate and atmospheric history of the past 420.00 years from the Vostok ice core, Antarctica," *Nature*, vol. 399, no. 6735, pp. 429–436, Jun. 1999.
[2] J. D. Scargle, "Studies in astronomical time series analysis II. Statistical aspects of spectral analysis of unevenly spaced data," *Astrophys. J.*, vol. 263, no. 2, pp. 835–853, Dec. 1982.
[3] R. Bos, S. de Waele, and P. M. T. Broersen, "Autoregressive spectral estimation by application of the Burg algorithm to irregularly sampled data," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 6, pp. 1289–1294, Dec. 2002.
[4] 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.
[5] R. H. Jones, "Maximum likelihood fitting of ARMA models to time series with missing observations," *Technometrics*, vol. 22, no. 3, pp. 389–395, 1980.
[6] M. B. Priestley, *Spectral Analysis and Time Series*. London, U.K.: Academic, 1981.
[7] S. M. Kay and S. L. Marple, "Spectrum analysis—A modern perspective," *Proc. IEEE*, vol. 69, no. 11, pp. 1380–1419, Nov. 1981.
[8] P. M. T. Broersen, "Automatic spectral analysis with time series models," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 2, pp. 211–216, Apr. 2002.
[9] P. M. T. Broersen and S. de Waele, "Automatic identification of time series models from long autoregressive models," *IEEE Trans. Instrum. Meas.*, vol. 54, no. 5, pp. 1862–1868, Oct. 2005.
[10] J. Durbin, "Efficient estimation of parameters in moving average models," *Biometrika*, vol. 46, no. 3/4, pp. 306–316, 1959.
[11] ——, "The fitting of time series models," *Rev. Inst. Int. Stat.*, vol. 28, no. 3, pp. 233–243, 1960.
[12] P. M. T. Broersen and R. Bos, "Order selection for autoregressive spectral estimation with randomly missing data," in *Proc. IEEE Benelux Signal Processing Symp. (SPS)*, Hilvarenbeek, The Netherlands, Apr. 15–16, 2004, pp. 33–36.



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

He is currently with the Department of Multi Scale Physics of the Delft University of Technology. His main research interest is in automatic identification on statistical grounds by letting measured data speak for themselves. He developed a practical solution for the spectral and the autocorrelation analysis of stochastic data by the automatic selection of a suitable order and type for a time-series model of the data.

Robert Bos was born in Papendrecht, The Netherlands, in 1977. He received the M.Sc. degree in applied physics from the Delft University of Technology, Delft, The Netherlands, in 2001. He is currently pursuing the Ph.D. degree at the Delft University of Technology.

He currently works at the Delft Center for Systems and Control, Delft University of Technology. His research is aimed at the development and the application of state-estimation techniques for high-order-first-principle models of complex processes in the process industry.