AUTOREGRESSIVE ORDER SELECTION IN MISSING DATA PROBLEMS

[1]P.M.T. Broersen and [2]R.Bos

[1]Department of Multi Scale Physics, [2]Delft Center for Systems and Control
Delft University of Technology , P.O.Box 5046, 2600 GA Delft, The Netherlands,
email broersen@tnw.tudelft.nl

ABSTRACT

Maximum likelihood presents a useful solution for the estimation of the parameters of time series models when data are missing. The highest autoregressive (AR) model order that can be computed without numerical problems is limited and depends on the missing fraction. Order selection will be necessary to obtain a good AR model. The best criterion to select an AR order with an accurate spectral estimate is slightly different from the criterion for contiguous data. The penalty for the selection of additional parameters depends on the missing fraction. The resulting maximum likelihood algorithm can give very accurate spectra, sometimes even if less than 1% of the data remains.

1. INTRODUCTION

A robust and simple algorithm for spectral analysis with missing data is a useful tool for signal processing. Practical observations are often incomplete because sensor failure or outliers cause missing data. In meteorological observations, the weather conditions may disturb the equidistant sampling scheme. A general spectral estimator for missing data is the method of Lomb-Scargle [1]. This method computes Fourier coefficients as the least squares fit of sines and cosines to the available observations. The Lomb-Scargle spectrum is accurate in detecting strong spectral peaks in low level background noise, but the bias of the method prevents the description of spectral shapes or slopes in those parts of the spectrum with lower power [2].

Many missing data methods are derived from algorithms that have been developed for consecutive data. Firstly, they reconstruct the missing data, followed by the estimation of the spectral density from the reconstructed consecutive signal. Those methods can give rather accurate results for small missing fractions [2]. However, reconstruction methods are never as accurate as an AR (autoregressive) model estimated with a robust maximum likelihood algorithm [2]. Exact maximum likelihood using Kalman filtering has been described for missing data [3] and also an approximate method using predictions in a finite interval[2].

The layout of the paper is as follows. Time series models and a robust AR estimator are introduced. The paper deals only with randomly missing data. Estimation of parameters in a white noise and an AR(1) example gives an idea of the variance of parameters and the behavior of the likelihood. For very small remaining fractions, less than 1%, low order AR models are estimated. Some specific choices for the penalty factor in an order selection criterion are compared.

2. AR MODELS

Autoregressive or AR models can describe the characteristics of stationary stochastic processes [4]. The power spectral density function and the autocovariance function are determined completely by the parameters of the AR model. An AR(*p*) model can be written as:

$$x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p} = \varepsilon_n , \qquad (1)$$

where ε_n is a purely random white noise process with zero mean and variance σ_ε^2. Almost any *stationary* stochastic process can be described as an unique AR(∞) model, at least theoretically. In practice, finite orders are sufficient. Parameters can be estimated from given data x_n to determine a model that could have been used to generate those data from white noise as input. The autocovariance function and the power spectral density of the data x_n can be computed from the known or from the estimated model parameters. With consecutive data, the model order *p* can be selected automatically, based on reliable statistical criteria [5]. The power spectrum $h(\omega)$ of the AR(*p*) model (1) is given by [4]:

$$h(\omega) = \sigma_\varepsilon^2 \left| \frac{1}{A(e^{j\omega})} \right|^2 = \sigma_\varepsilon^2 \left| \frac{1}{1+a_1 e^{-j\omega} + \cdots + a_\mathrm{p} e^{-j\omega p}} \right|^2 \qquad (2)$$

This definition applies for true as well as for estimated parameters. Models are stationary if the roots of *A(z),* called poles, are inside the unit circle. Also the infinitely long autocorrelation function is determined by the *p* parameters of (1), rather than by the inverse Fourier transform of (2). The Yule-Walker relations [6] describe the complete autocorrelation function of an AR(*p*) process

$$\rho_n + a_1 \rho_{n-1} + \cdots + a_p \rho_{n-p} = 0 ,\ n \ge 0;\ \rho_{-n} = \rho_n \cdot \qquad (3)$$

Reflection coefficients k_i are used to recursively determine the AR parameters $A_m(z)$ of all model orders *m* between 1 and *p*, with the Levinson-Durbin formulas [6]:

$$\begin{aligned} &\hat{a}_1^1 = k_1 \\ &\hat{a}_i^m = \hat{a}_i^{m-1} + k_m \hat{a}_{m-i}^{m-1}, && 1 \le i < m \ \cdot \\ &\hat{a}_m^m = k_m && 1 \le m \le p \end{aligned} \qquad (4)$$

Those relations have the property that all poles are inside the unit circle if the reflection coefficients are all less than 1 in absolute value. It is this property that will be used 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 true and estimated spectra. Also a time domain expression for the model error ME exists as a normalized prediction error PE [5]:

$$ME = N\left(PE / \sigma_{\varepsilon}^{2} - 1\right). \tag{5}$$

The PE is defined as the expectation of the fit of an estimated model to consecutive new 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 contiguous data. It yields the number of parameters p' as the minimal expectation of the ME for unbiased estimated AR(p') models, with p' greater than or equal to the true process order p.

3. ROBUST ESTIMATORS

Two different maximum likelihood AR estimators can be used for missing data problems. Both give almost always the same results and both have the same problems with robustness and convergence. The exact estimator [3] requires a computation time that is proportional to the sum of the available and of the missing data. The approximate estimator's time is proportional to the available number of data only [2]. The speed of the exact algorithm is higher if less than 75% of the data is missing; the approximate can be much quicker for very sparse data. To improve the numerical robustness, the algorithms build the estimated $A_m(z)$ with reflection coefficients k_i, for all model orders, by using (4) with unconstrained optimization of $tan(\pi/2*k_i)$ for increasing orders m. This guarantees that the estimated k_i is always in the range $-1< k_i <1$. Hence, all AR models computed by non-linear numerical optimization routines are stationary. The usual Yule-Walker or Burg algorithms for AR estimation in contiguous data keep previous reflection coefficients constant in computing the new k_m [6] for increasing orders. For missing data, however, all k_i are optimized afresh and simultaneously in the missing data algorithms, for every candidate model order m.

Starting values for the non-linear optimization of the AR(m) model are the reflection coefficients of the AR(m-1) model, with an additional zero for the new k_m, with good properties [2]. Having obtained several AR models, it is essential to develop some criterion to choose the best among the candidates. Order selection for models estimated by likelihood maximization can be performed with a generalized information criterion GIC(p,α), defined as:

$$GIC(p,\alpha) = L(X;\underline{\hat{a}}_p) + \alpha p\,. \tag{6}$$

In selection criteria, the negative of the likelihood, denoted $L(X;\underline{\hat{a}}_p)$, is computed for the parameter vector $\underline{\hat{a}}_p$ for which the minimum is found for the available data X. The order p with the minimum among the GIC(p,α) is selected. The best value for the penalty α has to be investigated in the missing data case. The value $\alpha = 2$ is the famous AIC criterion [4]. In order selection for consecutive data, penalty $\alpha = 3$ has been proposed as preferable [5].

4. WHITE NOISE SIMULATIONS

The possibilities, the robustness and the accuracy of the missing data algorithm will be studied in simulations where consecutive data are generated first with an AR(p) process. Afterwards, those data are transformed into a missing data problem by randomly erasing observations.

The missing data maximum likelihood estimation algorithm can also be used for contiguous data. In that case, the highest possible candidate AR order is limited to 20, 30 or perhaps 40, depending on the specific data and on N. Numerical problems with the multidimensional optimization algorithms as well as the required computation time prevent the estimation of higher order models. The usual Burg algorithm can estimate models of order N-1 in contiguous data [5]. The limitation of the highest candidate model order is more inspired by the non-linear optimization that is required by the maximum likelihood algorithm than by the missing data character of the problem. Restriction of the highest AR order to 10 in simulations prevents numerical problems and gives a clear idea about the *average* quality of selected models as a function of the penalty factor α in (6). The convergence of the maximum likelihood algorithms is best for low order AR models and for few missing data and it becomes a problem in some simulation runs if higher orders must be calculated for a large missing fraction. Then, the *average* quality of the selected models will become poor, but that average is not always representative for the quality of most of the individual simulation runs.

Table 1 gives results for a simulation experiment with white noise; γ is the remaining data fraction. Similar simulations with contiguous data have lead to finite sample criteria for AR order selection [7]. Theoretically expected values for contiguous data would be for the three columns:

$$L \approx N\text{-}m, \quad ME \approx m, \quad N^{*}\mathrm{var}(k_m) \approx 1$$

The ≈ denotes the asymptotical approximation, which is almost equal to the finite sample result for m<5 if N is 1000. The behavior of the likelihood L is almost the same as it would be for contiguous data: the average of the likelihood diminishes with 1 for each additional order. The ME is quite different. The first values are approximately given by m/γ, which is 10m for the Table 1. However, the ME values become still much greater than this missing data value for m > 5, e.g. ME=450 for the AR(10) model. The table gives the variance of the reflection coefficients of the estimated AR(5) model. It is approximately $1/\gamma$ for all five orders. If the AR model order would be greater than 5, the variance of

Table 1. The average of 100 simulation runs of the likelihood, of the ME and of N times the variance of reflection coefficients, as a function of the model order m. N=1000 white noise observations are remaining, γ= 0.1

m	L	ME	N*var k_m
1	997.2	9.7	9.9
2	996.3	19.2	10.8
3	995.3	32.9	12.1
4	994.5	44.0	9.7
5	993.4	66.1	12.8

all reflection coefficients would become greater than $1/\gamma$ in the same simulations, e.g. about $2/\gamma$ for the AR(10) model. For higher orders, the ME and the variance of all estimated reflection coefficients increase sharply. This explains the unpredictable performance of estimated high order AR models. It is a good reason to restrict the highest candidate AR order for small values of the remaining data fraction.

The drop of 1 per parameter in L is the same as in contiguous data because that is the natural decrease of the likelihood L if superfluous parameters are estimated. If the remaining fraction is γ , if the remaining number of observations is N and if the data times are random, it can be shown that the expected number of data fragments with at least two contiguous measurements is γN [2]. Likewise, the average number of occurrences for every fixed lag is about γN. Therefore, the *effective* number of observations in a *white noise* experiment is about γN. That explains the global behavior of the ME and the parameter variance in Table 1 for low orders.

5. AR(1) SIMULATIONS

Table 2. The average of L, ME and the variance of the reflection coefficients for 1000 AR(1) observations with a_1=-0.9, γ= 0.1.

m	L	ME	N*var k_m
1	540.4	0.3	0.2
2	539.4	7.1	22.1
3	538.6	29.1	60.2
4	537.3	123.1	88.9
5	536.1	382.6	36.7

In contiguous data, the white noise behavior is found in the likelihood, the variance of reflection coefficients and in the ME of all unbiased models of true order and higher. Table 2 gives simulation results of an AR(1) process. The likelihood diminishes with about 1 for each order, like in white noise. The difference between the likelihood for order 0 and 1 is also about 1 in white noise, but that difference is 460 for Table 2, indicating that the likelihood decreases very much if the first parameter is estimated. The ME for the estimated AR(1) model is 0.3. This is much less than the ME of the AR(1) model estimated in white noise. It is even less than 1, the expected value if the AR(1) model is estimated from contiguous data [5]. The ME of the AR(2) and of the AR(3) model are also smaller than the white noise results. However, the ME of the AR(4) and AR(5) model are much greater in Table 2 than in white noise. The variance of the first reflection coefficient is much smaller than in white noise, the variance of the higher order reflection coefficients becomes much higher. This shows that only the behavior of the likelihood above the true order is independent of the true process characteristics above the true order, but the behavior of the model accuracy ME and the of the variance of reflection coefficients depends for all model orders on the true process characteristics.

An example showed that taking only every third contiguous observation will improve the model error for AR(1) processes with a pole close to the unit circle [2]. This regular pattern of missing data gives an improved model accuracy, for the case that the number of observations that is actually used for the estimation of the AR(1) parameter is the same. Table 3 shows that also randomly missing data give ME values smaller than 1, until very small fractions γ. The average reduction of the likelihood between the orders 0 and 1 is always (much) greater than 1. However, this significant average likelihood reduction does not give satisfactory AR(1) models if γN is much less than 10. In other words, as long as the expected number of data fragments with at least two contiguous measurements, γN, is greater than 10, AR(1) models can be estimated if the poles are close to the unit circle. For small γ, L_0-L_1 is more or less proportional to γ or γN in Table 3, in both examples.

Table 3: The average ME and reduction of the likelihood for N=5000 AR(1) observations with a_1=-0.8 and for N=1000 observations with a_1=-0.9, as a function of γ. $L_0 \approx N$.

γ	N=5000 a_1=-0.8		N=1000, a_1=-0.9	
	ME	$(L_0$-$L_1)/N$	ME	$(L_0$-$L_1)/N$
0.95	0.94	0.975	1.13	1.632
0.9	1.04	0.968	1.04	1.570
0.75	0.89	0.884	0.90	1.470
0.5	0.79	0.695	0.53	1.235
0.25	0.63	0.431	0.42	0.855
0.1	0.63	0.207	0.34	0.460
0.05	1.05	0.111	0.51	0.258
0.025	2.13	0.056	0.65	0.142
0.01	4.79	0.023	0.74	0.063
0.005	8.93	0.012	44.61	0.033
0.0025	25.61	0.006	1427.31	0.014
0.001	34.40	0.002	1957.90	0.005

Table 4. The average ME and reduction of the likelihood for AR(1) observations with a_1=-0.9, for γN = 10.

N	γ	ME	L_0-L_1
1000	0.01	0.74	63
500	0.02	0.93	56
250	0.04	0.41	54
100	0.1	0.42	45
50	0.2	0.68	32
25	0.4	2.24	20
10	1	15.45	3.5

Table 4 shows once more that the decrease of the likelihood is almost the same for different N, as long as γN remains the same and only if γ is small. The explanation is that that the expected number of fragments with at least two contiguous data is γN and the average number of occurrences for every other fixed lag is about γN. Only neighboring data can contribute significantly to the decrease of the likelihood, and that number of close data is given by γN.

6. AR(6) SIMULATIONS

Comparing order selection criteria (6) with different penalties α is not interesting for white noise: the model of order 0 will be the best and the criterion with the highest penalty will most often select that order 0. It is necessary that examples have both the possibility of overfit and of underfit, by selecting a too high or a too low order, respectively. To easily construct different interesting examples, the following procedure is used in simulations: the true parameters of a generating process of an arbitrary

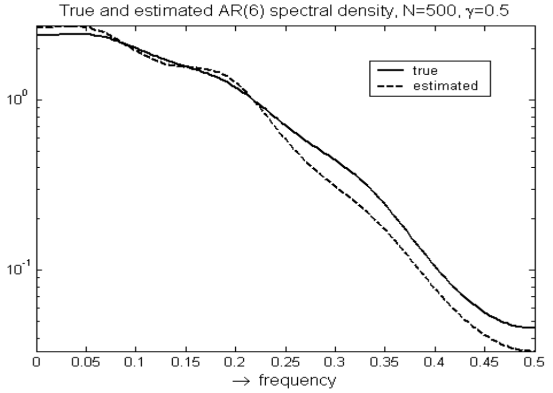

Fig.1. True and estimated AR(6) spectrum from 500 observations with 50 % missing , β= - 0.6. The ME is 7.6 in this realization.

Table 5. The average ME of 100 simulation runs of fixed order and of selected models for different penalties, as a function of the remaining fraction γ for N=500, β =-0.6.

→ γ	0.1	0.25	0.5	0.75	0.9	0.95
			Fixed order models			
AR(1)	134.2	128.3	122.9	120.6	119.0	118.8
AR(2)	59.5	49.6	44.6	41.1	40.1	40.2
AR(3)	***51.3***	30.4	20.7	16.8	16.0	15.5
AR(4)	126.7	***29.6***	***13.5***	9.1	8.2	8.0
AR(5)	358.3	55.7	14.2	***7.6***	***5.9***	6.0
AR(6)	790.0	113.6	19.7	8.4	6.1	***5.8***
AR(10)	15059	1884	101.4	10.6	11.8	10.1
			Selected model orders			
α=2	857.6	221.7	28.9	12.8	8.9	8.4
α=3	308.9	51.8	27.3	***12.4***	***8.8***	***8.4***
α=4	110.2	***51.2***	***26.9***	12.7	9.3	8.9
α=5	***95.7***	53.0	28.4	14.7	10.0	9.3
α=7	111.1	64.8	35.7	17.9	11.8	11.1
α=10	146.3	84.8	41.3	24.9	16.2	13.7

order p are built from reflection coefficients with (4), with $k_m=\beta^m$. In this way, all poles of the generating process have the same radius β. Figure 1 shows the true and the estimated AR(6) spectrum of a process with β = - 0.6. The accuracy of the spectrum is remarkable. The expected value of the ME for contiguous data would be 6. Randomly missing 50 % of the data has very little influence on the spectral accuracy in this individual simulation run.

For each missing data record, AR models of orders 1 to 10 are estimated. Order selection has not yet been successful for orders greater than 12 or 15. Simulation examples have been found in which the model with the lowest GIC(p,3) was poor if models with orders higher than 10 were candidates, and one of those orders was selected in a few simulation runs, destroying the average quality.

The best order for the estimated fixed order model depends on the true process parameters and on the missing fraction in Table 5. If the fraction of remaining data is lower, the optimal order tends to be lower.

The best value for the penalty depends on the compromise between underfit and overfit of the example processes and also on the highest candidate model order. In the given simulations, the penalty factor α was optimal for the values between 3 and 5. This has been found in many simulations. If overfit errors are not probable, e.g. because the maximum candidate order is close to the best order, a smaller penalty may seem appropriate. If underfit errors are not probable, like in the AR(1) examples of Table 3, the best penalty may become 10 or even higher. However, in examples where both the risks of underfit and of overfit are present, the range $3 < \alpha < 5$ gives always acceptable results, with a preference for a greater penalty factor α if the remaining fraction γ becomes smaller..

The average value of the ME in Table 5 for γ= 0.5 and the single example in Fig.1 are different. If more data are missing, the average ME becomes higher because of the very poor quality of a few simulation runs. In those runs, the minimization of the likelihood function did not converge properly. This will happen sometimes for small γ and for high candidate AR orders. For a single given missing data problem, it is always possible to improve the convergence by trying various starting values for the parameters.

CONCLUDING REMARKS

Both the exact and the approximating maximum likelihood algorithm can estimate accurate AR models in data with randomly missing observations. AR models are estimated for increasing model orders by using the previous model as starting values for the non-linear optimization, with an additional zero for the new order. The average accuracy is comparable to that of contiguous data if not more than 25 % of the data is missing. For higher missing fractions, the model quality depends on the true process. AR(1) models with a pole close to the unit circle can be estimated very accurately, even if more than 99 % of the data is missing.

The best value for the penalty factor in the order selection criterion depends on the missing fraction. The penalty 2 is too low, penalty 3 is a good value if less than 25 % is missing, 5 gives good results if less than 25 % remains and penalty 4 is a good compromise if about 50 % of the data is missing.

REFERENCES

[1] J. D. Scargle, Studies in astronomical time series analysis II. “Statistical aspects of spectral analysis of unevenly spaced data”. *The Astrophysics Journal*, no. 263, pp. 835-853, 1982.

[2] P. M. T. Broersen, S. de Waele and R. Bos, “Autoregressive spectral analysis when observations are missing”. *Automatica,* vol. 40, 2004.

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

[4] M.B. Priestley, *Spectral Analysis and Time Series*, London, Ac. Press, 1981.

[5] P.M.T. Broersen, “Automatic spectral analysis with time series models”. *IEEE Trans. on Instrumentation and Measurement*, vol. 51, no. 2, pp. 211-216, 2002.

[6] S.M. Kay and S.L. Marple, “Spectrum analysis-a modern perspective”. *Proc. IEEE*, vol. 69, pp. 1380-1419, 1981.

[7] P.M.T. Broersen, “Finite sample criteria for autoregressive order selection”. *IEEE Trans. on Signal Processing*, vol. 48, pp. 3550-3558, 2000.