EURASIP Journal on Applied Signal Processing 2002:1, 30–37

Multiple Optima in Identification of ARX Models Subject to Missing Data

Ragnar Wallin

S3-Process Control, Royal Institute of Technology, SE-100 44 Stockholm, Sweden
Email: ragnarw@kth.se

Alf J. Isaksson

S3-Process Control, Royal Institute of Technology, SE-100 44 Stockholm, Sweden
Email: alf@s3.kth.se

Received 31 July 2001

Special system identification algorithms are required if there are significant amounts of data missing. Some such algorithms have been developed previously and typically result in iterative procedures for the parameter estimation. Since missing data can be viewed as irregular sampling (decimation) of the signals, it is obvious that there is a risk for aliasing. In system identification aliasing manifests itself as potential multiple global optima of the identification loss function. The aim of this paper is to investigate under what circumstances this may in fact occur. The focus of the paper is on periodic missing data patterns. It is shown that it is, in fact, not the fraction of missing data that is important, but rather what time lags of the input and output correlation and cross-correlation functions that can be estimated. For ARX models with all input data observed we verify that there is indeed only one global optimum.

Keywords and phrases: parameter estimation, irregular sampling, linear systems.

1. INTRODUCTION

Surprisingly often data sets used for system identification are incomplete. Some observations are missing, either according to a periodical pattern or at random. Examples of randomly missing data include sensor failures, outliers and temporary plant shutdown. Periodically missing data appear, for instance, in time sharing of sensors, radar scans, and multirate sampling. As identification experiments are expensive and time consuming, methods that can cope with missing data are attractive. They make it possible to use all data sets that are available.

The missing data problem has been studied extensively in statistics, but less so in engineering literature. A survey of the research in statistics is given in the book by Little and Rubin [1]. Estimation of ARMA models is studied in [2, 3, 4], estimation of AR models in [5, 6, 7, 8]. In the engineering literature we find [9, 10, 11, 12, 13, 14, 15, 16].

The specific problem studied in this paper is the existence of multiple global optima of the system identification procedure. That multiple optima can occur is obvious realizing that missing data can be viewed as sampling (or decimation), albeit often irregular, of the signals. Hence, the sampling theorem indicates that aliasing may occur. It is, however, not entirely obvious for what combinations of missing data pattern and model order there may be more than one system that optimally predicts the observed data.

The problem with multiple optima is important as we do not want to estimate a model that predicts an incorrect spectral behaviour. This paper presents some results for linear time invariant systems with and without input and gives some examples for autoregressive (AR) and autoregressive models with an exogenous input (ARX).

2. MOTIVATING EXAMPLES

The estimate of AR model parameters is unique when all data is observed and the model order is chosen correctly [17]. This section presents two examples showing that there may indeed be more than one solution when data is missing.

Example 1. Consider a first-order AR model. It is described by the difference equation

$$y_k + a y_{k-1} = e_k, \tag{1}$$

where e is white Gaussian noise of variance λ. Now assume that only every second data point is observed. We will describe

such a periodic pattern with the notation {10}, where 1 means that the data point is observed and 0 means it is missing. A difference equation in observed data only is

$$y_{k+1} - a^2 y_{k-1} = e_{k+1} - a e_k. \tag{2}$$

This is in fact also a first-order AR model. As we only can estimate a^2, it is obviously impossible to tell if the data is coming from the model with parameter $-a$ or a.

Example 2. The second-order AR model

$$y_k - 1.07 y_{k-1} + 0.49 y_{k-2} = e_k \tag{3}$$

has poles in $0.54 \pm i0.45$. Figure 1 is a contour plot of the likelihood function versus locations of the pole in the upper half plane. Four hundred data points are observed. We clearly see two maxima for poles located in $0.55 + i0.45$ and $-0.55 + i0.45$.

The results in Examples 1 and 2 actually follow directly from the sampling theorem, since the pattern {10} is simply decimation. Hence the Nyquist frequency is reduced a factor two. In the z-domain this corresponds to complex poles with an angle in $[-\pi/2, \pi/2]$ to the positive real axis. Of much more interest are irregular patterns like, for example, {1010000}. When can we expect multiple optima of the likelihood function in such cases?

3. MAXIMUM LIKELIHOOD ESTIMATION

The optimal way of estimating model parameters (with or without missing data) is the maximum likelihood method. The aim is to find model parameters that maximize the probability that the data come from this model.

To derive the likelihood function we need the probability density function for the observed data. For linear models the probability density function can be found by putting the model into state-space form and using the Kalman predictor, as is shown in [2, 3]. For Gaussian noise the prediction errors produced by the Kalman predictor are mutually independent and Gaussian. The likelihood function is then given by

$$f_\theta(\zeta) = \prod_{k \in O} \frac{1}{\sqrt{2\pi \det S_k}} \times \exp\left\{ -\frac{1}{2} (\zeta_k - \hat{\zeta}_k)^T S_k^{-1} (\zeta_k - \hat{\zeta}_k) \right\}. \tag{4}$$

The set O denotes all the time instants where output data, ζ_k, are observed and S_k is the covariance of the prediction errors.

One way of computing the likelihood function when both outputs and inputs are missing (if we do not want to treat the missing inputs as parameters) is to introduce an input model as in Section 3.1.

3.1. Linear stochastic models

Most linear finite state stochastic models can be written on innovation form

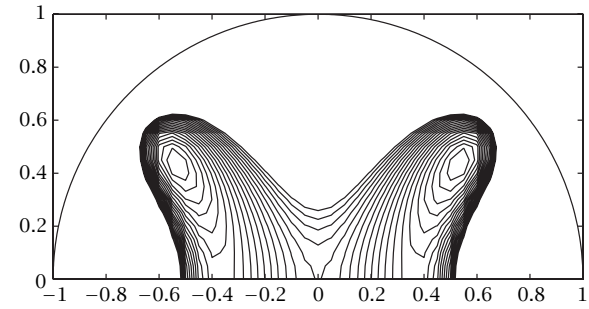

Figure 1: Contour plot of the likelihood function versus the pole location in the upper half plane.

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Fn_k, \\ y_k &= Cx_k + n_k, \end{aligned} \tag{5}$$

where at time k, y_k is the output, u_k is the input and n_k is white noise. Some examples of frequently used models where this is possible include OE, ARX, ARMAX, and BJ. If we extend the model with a time series input model, for example, AR or ARMA, it is still possible to write the model on innovation form with different matrices A, C, and F. Then, we get a multivariate time series where the output, z_k, and the noise vector, n_k, are

$$z_k = \begin{bmatrix} y_k \\ u_k \end{bmatrix}, \qquad n_k = \begin{bmatrix} v_k \\ y_k \end{bmatrix}. \tag{6}$$

The model is an innovation form, without an input matrix B

$$x_{k+1} = Ax_k + Fn_k, \qquad z_k = Cx_k + n_k. \tag{7}$$

The noise vector has variance

$$\Lambda = \begin{cases} \lambda & \text{for system (5)}, \\ \begin{bmatrix} \lambda & 0 \\ 0 & \sigma \end{bmatrix} & \text{for system (7)}. \end{cases} \tag{8}$$

If only some outputs are measured, we can introduce the matrix D_k that picks out the outputs that actually are observed at time k. For the system (5), D_k is

$$D_k = \begin{cases} 1, & \text{if } y_k \text{ is observed}, \\ \text{empty}, & \text{otherwise}, \end{cases} \tag{9}$$

and for the system (7)

$$D_k = \begin{cases} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, & \text{if } y_k \text{ and } u_k \text{ are observed}, \\ \begin{bmatrix} 1 & 0 \end{bmatrix}, & \text{if only } y_k \text{ is observed}, \\ \begin{bmatrix} 0 & 1 \end{bmatrix}, & \text{if only } u_k \text{ is observed}, \\ \text{empty}, & \text{otherwise}. \end{cases} \tag{10}$$

An empty matrix has dimension zero. Whenever such a matrix appears in an equation it may be omitted. Also, if a matrix is multiplied by an empty matrix the resulting one is empty. This results in the time-varying state-space model

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Fn_k, \\ \zeta_k &= D_k C x_k + D_k n_k = \bar{C}_k x_k + D_k n_k. \end{aligned} \tag{11}$$

Of course, B is an empty matrix for the system (7). One possible state-space form for an ARX model with an AR input model is given in [14]. This is what is used in the examples at the end of this paper.

3.2. The Kalman predictor

The Kalman predictor for the system (11) is given by (cf. [18, pages 429–430])

$$\begin{aligned} \hat{x}_{k+1} &= (A - K_k \bar{C}_k)\hat{x}_k + Bu_k + K_k \zeta_k, \\ \hat{\zeta}_k &= \bar{C}_k \hat{x}_k, \end{aligned} \tag{12}$$

where K_k is given by

$$\begin{aligned} K_k &= (AP_k \bar{C}_k^T + R_{12_k})(\bar{C}_k P_k \bar{C}_k^T + R_{2_k})^{-1}, \\ P_{k+1} &= AP_k A^T + R_1 - K_k(\bar{C}_k P_k A^T + R_{12_k}^T). \end{aligned} \tag{13}$$

The matrices R_1, R_{12_k}, and R_{2_k} are

$$R_1 = F\Lambda F^T, \qquad R_{12_k} = F\Lambda D_k^T, \qquad R_{2_k} = D_k \Lambda D_k^T. \tag{14}$$

Often we have no information about the initial state of the Kalman filter. It is then usually chosen to be zero. The output of the Kalman predictor $\hat{\zeta}_k$ may, in that case, be expressed as (see [19, pages 392–393])

$$\hat{\zeta}_k = \begin{cases} 0, & k = 0, \\ \displaystyle\sum_{j=0}^{k-1} \bar{C}_k \Phi_{k,j+1}(K_j \zeta_j + Bu_j), & k \geq 1, \end{cases} \tag{15}$$

$$\Phi_{k,j} = \begin{cases} \displaystyle\prod_{i=j}^{k-1} (A - K_i \bar{C}_i), & k \geq j+1, \\ I, & k = j. \end{cases} \tag{16}$$

As K_j is an empty matrix when the output is not observed, it follows from (15) that $\hat{\zeta}_k$ is a function of observed data only.

Periodic Kalman predictor

If data is observed in an M-periodic pattern and the realization is minimal, the Kalman predictor converges to a periodic steady state predictor with period M [20]. Each position, p, in the observation pattern corresponds to a constant P_p and a constant K_p. The steady state Kalman predictor consequently has the following properties:

$$\begin{aligned} P_{p+iM} &= P_p, \\ K_{p+iM} &= K_p, \\ \Phi_{k+iM,j+iM} &= \Phi_{k,j}, \end{aligned} \tag{17}$$

where i is an integer and $p \in [0, M-1]$. Methods to compute the periodic steady state Kalman predictor are given in [20]. The case that all data is observed is a special case with $M = 1$.

4. SUFFICIENT STATISTIC

The main vehicle to the analysis of why two systems may be equally likely to have produced the data observed will be the notion of sufficient statistic. A statistic is said to be sufficient if it contains all the information in y that is useful for estimating θ. The probability density function of the prediction errors of the Kalman predictor is multivariate Gaussian as is seen in (4). This is an exponential family distribution as it can be written in the form

$$f_\theta(y) = a(y)c(\theta) \exp\left[\sum_{i=1}^{k} \pi_i(\theta) t_i(y) \right], \tag{18}$$

where $t = (t_1, t_2, \ldots, t_k)$ is the sufficient statistic [21]. In the exponential family, any statistic that is sufficient is also minimal.

The significance of the sufficient statistic concept is that the maximum likelihood estimate can be computed equally well from the sufficient statistic as from data directly. We are now ready to present the main result of this paper.

Result 3. *For systems* (5) *and* (7)*, where data is observed in a periodical pattern, one sufficient statistic are all lags of the sample correlation and sample cross-correlation functions of the input and output, that can be obtained from data.*

Derivation. If we have measurements from $k = 1$ to $k = N$ and data is observed in a periodical fashion, the number of data points in position p of the M-periodic observation pattern is

$$N_p = \mathrm{INT}\left(\frac{N-p}{M}\right), \tag{19}$$

where $\mathrm{INT}(\cdot)$ is the integer part of $(\cdot)$. Equation (4) may now be written as

$$\begin{aligned} f_\theta(\zeta) = \prod_{p \in \mathcal{P}} & \frac{1}{(2\pi \det S_p)^{N_p/2}} \\ & \times \exp\left\{ -\frac{1}{2} \sum_{p \in \mathcal{P}} \sum_{i=0}^{N_p - 1} (\zeta_{p+iM} - \hat{\zeta}_{p+iM})^T \right. \\ & \left. \times S_p^{-1}(\zeta_{p+iM} - \hat{\zeta}_{p+iM}) \right\}, \end{aligned} \tag{20}$$

where the set $\mathcal{P}$ denotes the observed positions in the M-periodic pattern. To find a sufficient statistic we only have to rearrange the terms of the exponent of equation (20).

If we, instead of the true Kalman predictor, use the periodic Kalman predictor we do not get the true likelihood function but the difference is very small (caused by a transient). The terms $\hat{\zeta}_k$ are linear functions of older data according to equation (15). The coefficients in the sum in the exponent of

equation (20) are periodic as the predictor is periodic. The sum can thus be divided into smaller sums with a constant coefficient in front of each sum. Doing so we see that one sufficient statistic is sums of the type

$$\text{coefficient} \sum \zeta_{k+\tau}\zeta_k^T, \tag{21}$$

and for the system (5) we also get sums of the type

$$\text{coefficient} \sum u_{k+\tau}u_k, \qquad \text{coefficient} \sum \zeta_{k+\tau}u_k. \tag{22}$$

This is sample correlations and sample cross-correlations of the input and output signals.

Remarks. (1) Two cases where the sufficient statistic actually is not all lags of sample correlation and cross-correlation functions is AR and ARX models when all data is observed and AR(1) models where every Mth sample of the output is observed. The reason for this is that the transition matrix, $\Phi_{k,k+M}$, is nilpotent in those cases.

(2) Two models are equally likely if the data they produce fits the correlation and cross-correlation functions equally well.

(3) In [22] it is shown that, irrespective of if data is missing in a periodic pattern or at random, a signal can be reconstructed if the sampling frequency is high enough. Define the fraction of the observed data, γ, as

$$\gamma = \lim_{N\to\infty} \frac{N_{\text{obs}}}{N}. \tag{23}$$

If we observe a fraction γ of the data, the Nyquist frequency decreases a factor γ. Maximum likelihood estimation of the model parameters can (at least asymptotically) be viewed as sampling of the correlation functions. It is consequently the fraction, γ, of observed lags of the correlation functions that decide if there can be any aliasing effects. Hence, if the identification is restricted to searching for systems with poles in $[-\gamma\pi, \gamma\pi]$ to the positive real axis, the absence of alias systems is guaranteed. We could apply Marvasti's nonuniform sampling theorem to the missing correlation pattern. However, this seems to be a bit too conservative. Some combinations of model orders and missing correlation patterns result in a unique maximum of the likelihood function even though Marvasti's theorem says that there is a possibility that there may be more than one correlation function that fit the observed lags equally well. The reason is that the flexibility of the correlation function is limited by the model order.

(4) It is the sampling of correlations and not the sampling of data that is important. This gives a more generous bound on when alias effects can occur. Take, for example, a signal with the missing data pattern $\{11010\}$, we are observing three out of five data points. The missing correlation pattern is, however, $\langle 11111\rangle$. So, even though we only have 60 percent of the data we can calculate all sample correlations.

(5) Randomly missing data, if we assume that every sample has a positive likelihood to be observed, will not cause a problem. Asymptotically we can consistently estimate all lags of the correlation functions.

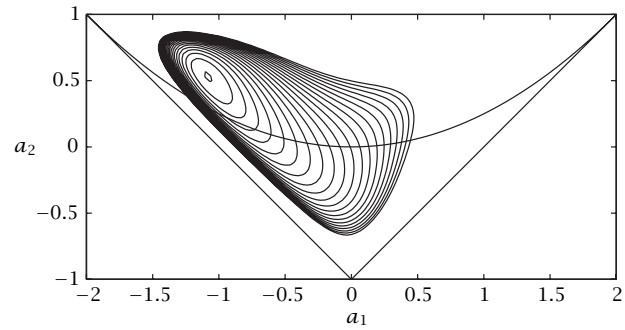

FIGURE 2: Contour plot of the log-likelihood values, as a function of a_1 and a_2, with observations according to the pattern $\{110\}$. Levels from -546 to -164 with the increment 20.

5. EXAMPLES

In the next examples the second-order AR model

$$y_k - 1.0725y_{k-1} + 0.49y_{k-2} = e_k, \tag{24}$$

with noise variance 0.3662 is used. The noise variance is chosen to give $r_y(0) = 1$. In all the examples four hundred data points are observed. In the plots of the maximum likelihood function λ is scaled to give an $r_y(0) = 1$.

If we use the Yule-Walker equations to express the autocorrelations in the model parameters, we get

$$\begin{aligned} r_y(0) &= \frac{(1+a_2)\lambda}{(1+a_2)(1-a_2^2) - a_1^2(1-a_2)}, \\ r_y(1) &= \frac{-a_1\lambda}{(1+a_2)(1-a_2^2) - a_1^2(1-a_2)}, \\ r_y(k) &= -a_1 r_y(k-1) - a_2 r_y(k-2) \quad \text{for } k \geq 2. \end{aligned} \tag{25}$$

5.1. *Observation pattern* $\{110\}$

We observe two thirds of the data points, but we can estimate all lags of the autocorrelation function. As the parameter estimation problem can be looked upon as sampling of the autocorrelation function the maximum likelihood function has only one unique maximum. This is verified in Figure 2. The triangle in the figure is the stability triangle. All points above the curve in the triangle correspond to complex conjugated poles and all points below correspond to two real poles.

5.2. *Observation pattern* $\{101000\}$

The observation pattern $\{101000\}$ illustrates that there can be more than one global maximum of the likelihood function. We have the missing autocorrelation pattern $\langle 101010\rangle$. It is easy to see from (25) that the autocorrelations for even time lags are the same irrespective of if $a_1 < 0$ or if $a_1 > 0$. We get two AR models that match the sample autocorrelations just as well.

A plot of the log-likelihood function is shown in Figure 3. The estimated parameters and the values of the log-likelihood function are shown in Table 1. There are two global maxima

Table 1: Parameter estimates and log-likelihood when data is observed according to the pattern {101000}.

a_1	a_2	λ	Log-likelihood
−1.0129	0.4258	0.3880	−360.03
1.0129	0.4258	0.3880	−360.03

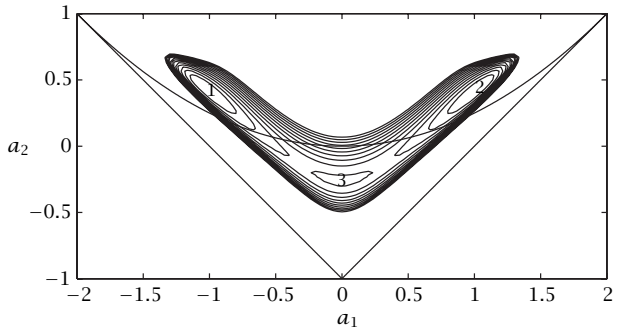

Figure 3: Contour plot of the log-likelihood values, as a function of a_1 and a_2, with observations according to the pattern {101000}. Levels from −364 to −398 with the increment 4.

Table 2: Parameter estimates and log-likelihood when data is observed at every third time sample.

a_1	a_2	λ	Log-likelihood
−1.0774	0.4890	0.3449	−374.25
−0.0521	0.4879	0.7233	−374.17
1.0761	0.4491	0.3406	−374.53

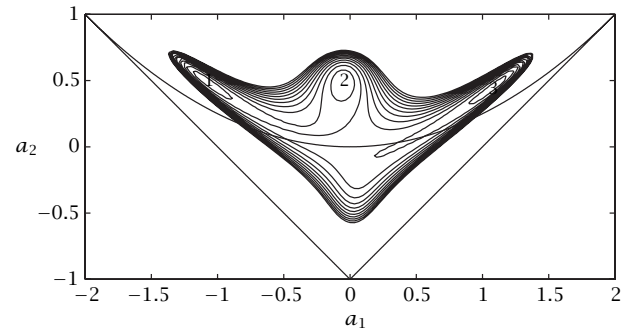

Figure 4: Contour plot of the log-likelihood values, as a function of a_1 and a_2, with observations according to the pattern {100}. Levels from −410 to −377 with the increment 3.

and one local. The local maximum has a value of −368.86. Local maxima are undesired as most numerical algorithms are only guaranteed to converge to a local, not a global, maximum.

5.3. Observation pattern {100}

The third example illustrates that the wrong model can be more probable than the real one if the number of data points is low. We have the observation pattern {100}. There are no multiple solutions of (25) evaluated at every third time lag but there are three models that match the data almost equally well (Table 2 and Figure 4). That it is indeed the autocorrelations that make the models almost equally probable is illustrated in Figure 5.

5.4. Observation pattern {1010000}

Another example where the local maximum has almost the same value as the global one is the observation pattern {1010000}. The missing autocorrelation pattern is ⟨1010010⟩. The estimated parameters and the log-likelihood values are shown in Table 3 and a plot of the log-likelihood function is given in Figure 6.

5.5. Examples for models with input

In this section, we consider the ARX model

$$y_k = -a y_{k-1} + b u_{k-1} + v_k \tag{26}$$

with input AR model

$$u_k = -c u_{k-1} + \gamma_k. \tag{27}$$

The variance of the noise v_k is λ and the variance of the noise γ_k is σ. A state-space representation of the form (5) for the system is

$$\begin{aligned} x_{k+1} &= \begin{bmatrix} -a & b \\ 0 & -c \end{bmatrix} x_k + \begin{bmatrix} v_k \\ \gamma_k \end{bmatrix}, \\ \begin{bmatrix} y_k \\ u_k \end{bmatrix} &= \begin{bmatrix} -a & b \\ 0 & -c \end{bmatrix} x_k + \begin{bmatrix} v_k \\ \gamma_k \end{bmatrix}. \end{aligned} \tag{28}$$

5.6. Missing output pattern {10} and missing input pattern {01}

Assume the observation pattern {10} for the outputs and {01} for the inputs. At each time instant either an input or an output is observed. As a result only even lags for the autocorrelation of output and input can be obtained from data. For the cross-correlation instead only odd lags can be obtained. The correlation matrices for z_k are

$$\begin{aligned} r_z(0) &= \begin{bmatrix} \dfrac{\lambda(1-c^2)(1-ac)+\sigma b^2(1+ac)}{(1-c^2)(1-ac)(1-a^2)} & -\dfrac{\sigma bc}{(1-c^2)(1-ac)} \\ -\dfrac{\sigma bc}{(1-c^2)(1-ac)} & \dfrac{\sigma}{1-c^2} \end{bmatrix}, \\ r_z(\tau) &= \begin{bmatrix} r_y(\tau) & r_{yu}(\tau) \\ r_{yu}(-\tau) & r_u(\tau) \end{bmatrix} \\ &= A^\tau r_z(0) \\ &= \begin{bmatrix} -a & b \\ 0 & -c \end{bmatrix}^\tau r_z(0), \quad \tau \geq 0. \end{aligned} \tag{29}$$

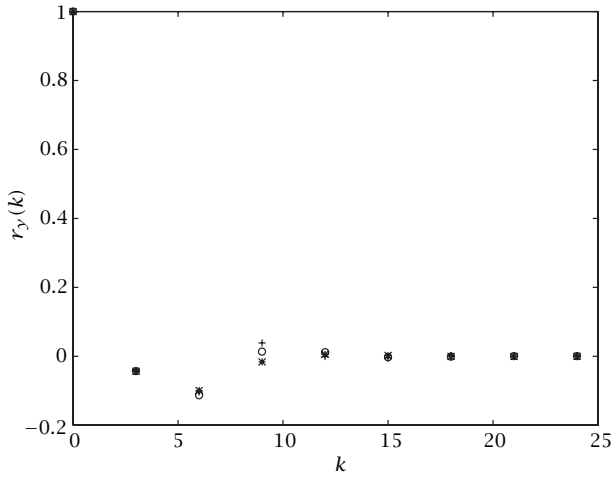

FIGURE 5: The autocorrelations that can be estimated for the three models with every third data point observed.

TABLE 3: Parameter estimates and log-likelihood when data is observed according to the pattern {1010000}.

a_1	a_2	λ	Log-likelihood
−1.0387	0.4435	0.3918	−380.98
0.4000	−0.1414	0.7784	−382.44

If we examine $r_z(\tau)$ for the entries that can be estimated we see that a and c always appear with an even exponent or in a product of two odd exponents between them. Hence, it is impossible to distinguish the pair $(-a, -c)$ from the true values (a, c). As an example the parameters

$$a = -0.5, \qquad b = 1, \qquad c = -0.7 \tag{30}$$

are chosen. In Figure 7, the log-likelihood function of one data realization (with 200 time instants) is plotted as a function of a and c (while b, λ, and σ assume their true values). The two global maxima are clearly visible.

5.7. *Missing output pattern* {10} *and missing input pattern* {10}

Here we consider the system (26) and (27) again. The output and input are observed according to the missing data pattern {10} only even lags of the autocorrelation functions and cross-correlation functions from input and output can be obtained. For those lags the parameters a, b, and c appear in combinations that makes it impossible to distinguish the triplet $(-a, -b, -c)$ from the true values (a, b, c).

5.8. *Missing output pattern* {10} *and no missing inputs*

Again we look at the system (26) and (27). The output is observed according to the missing data pattern {10} but all inputs are observed. In this case we will only get one maximum of the likelihood function as we can estimate all lags

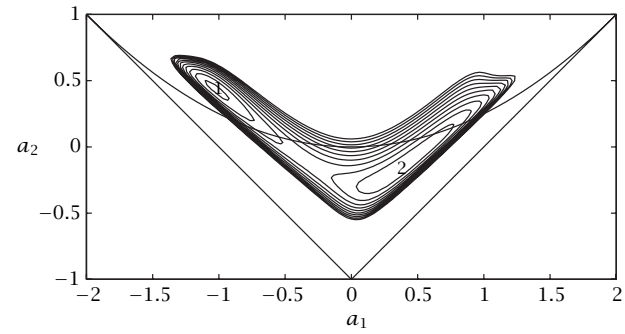

FIGURE 6: Contour plot of the log-likelihood values, as a function of a_1 and a_2, with observations according to the pattern [1010000]. Levels from −412 to −382 with the increment 3.

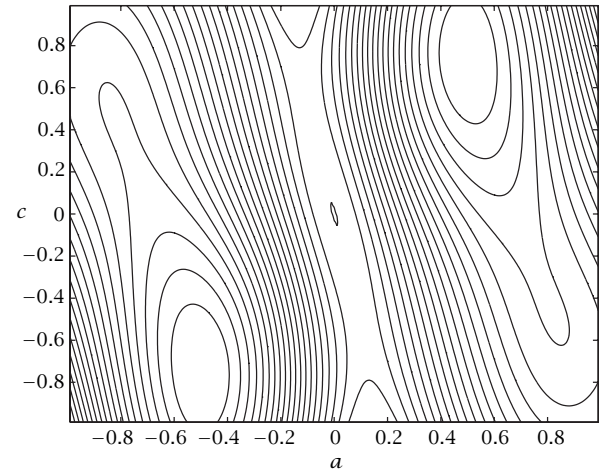

FIGURE 7: Contour plot of the log-likelihood values, as a function of a and c.

of the correlation function of the input and all lags of the cross-correlation function between the input and output and solve a and b from the following equations:

$$\begin{aligned} r_{yu}(1) &= -a r_{yu}(0) + b r_u(0), \\ r_{yu}(2) &= -a r_{yu}(1) + b r_u(1). \end{aligned} \tag{31}$$

This is possible as

$$\det \begin{bmatrix} -r_{yu}(0) & r_u(0) \\ -r_{yu}(1) & r_u(1) \end{bmatrix} \neq 0 \quad \text{when } c^2 \neq 1. \tag{32}$$

The parameter c has to be less than one in magnitude if the input process is to be stationary and computing correlations should be relevant. A contour plot of the log-likelihood function of one realization of the system (with 200 time instants) is shown in Figure 8.

6. ARX MODELS WITH ONLY MISSING OUTPUTS

As we saw in Section 5.8 we only had one global optimum. Is this always the case when all input data is observed? When

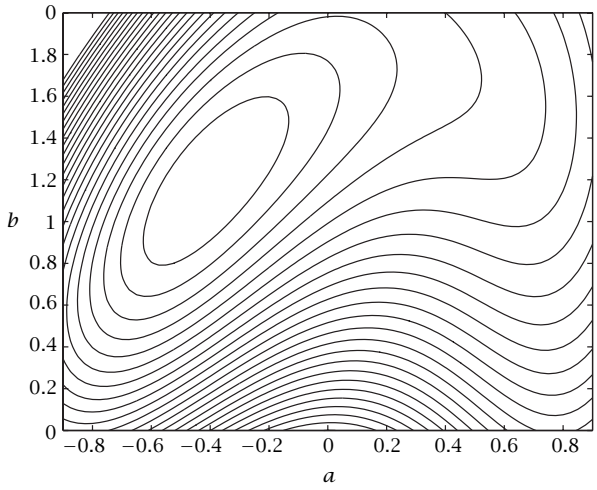

FIGURE 8: Contour plot of the log-likelihood values, as a function of a and b.

only outputs are missing we can estimate all time lags of the autocorrelation of the input, all time lags of the cross-correlation between the input and the output but only some time lags of the autocorrelation of the output. We can, however, always form the system of equations

$$\begin{bmatrix} \hat{r}_{yu}(1) \\ \vdots \\ \hat{r}_{yu}(na+nb) \end{bmatrix} = \begin{bmatrix} -\hat{r}_{yu}(0) & \cdots & \hat{r}_u(-nk-nb) \\ \vdots & \ddots & \vdots \\ -\hat{r}_{yu}(na+nb-1) & \cdots & \hat{r}_u(na-nk+1) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ b_{nb} \end{bmatrix}. \tag{33}$$

These are the same type of equations as we get when we use an instrumental variable method [23] and use old inputs as instruments. We know that we will get a unique consistent estimate under mild assumptions on the input. Basically, we need persistent excitation and an open loop identification experiment.

As we can get a unique consistent estimate from a subset of the sufficient statistic we can, of course, do something better using the entire statistic (less parameter variance). There can be, thus, only one global optimum of the likelihood function.

7. CONCLUSIONS

In this paper, we have studied the existence of multiple global optima of the likelihood function when identifying parameters of AR and ARX models. It is shown that the parameter estimation problem should be looked upon as a sampling of the correlation and cross-correlation functions of the input and output signals rather than a sampling of data. Hence, randomly missing data should not cause any problem as asymptotically all sample correlations and sample cross-correlations can be computed eventually.

It is established that two parameter sets yield identical values of the likelihood function if they fit the obtainable lags of the sample correlation and sample cross-correlation functions equally well. Also, it is shown that ARX models with all input data observed will not result in several optima.

REFERENCES

[1] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, New York, 1987.

[2] C. F. Ansley and R. Kohn, "Exact likelihood of a vector autoregressive moving average process with missing or aggregated data," *Biometrika*, vol. 70, no. 1, pp. 275–278, 1983.

[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] Y. Rosen and B. Porat, "Optimal ARMA parameter estimation based on the sample covariances for data with missing observations," *IEEE Transactions on Information Theory*, vol. 35, no. 12, pp. 342–349, 1989.

[5] A. C. Harvey and C. R. McKenzie, "Missing observations in dynamic econometric models: A partial synthesis," in *Time Series Analysis of Irregularly Observed Data*, E. Parzen, Ed., pp. 108–133, Springer-Verlag, New York, 1984.

[6] P. B. McGiffin and D. N. Murthy, "Parameter estimation for autoregressive systems with missing observations," *International Journal of Systems Science*, vol. 11, no. 9, pp. 1021–1034, 1980.

[7] P. B. McGiffin and D. N. Murthy, "Parameter estimation for autoregressive systems with missing observations-Part II," *International Journal of Systems Science*, vol. 12, no. 6, pp. 657–663, 1981.

[8] R. B. Miller and O. Ferreiro, "A strategy to complete a time series with missing observations," in *Time Series Analysis of Irregularly Observed Data*, E. Parzen, Ed., pp. 251–275, Springer-Verlag, New York, 1984.

[9] M. Tanaka and T. Katayama, "Robust identification and smoothing for linear system with outliers and missing data," in *Proceedings IFAC 11th Triennial World Congress*, vol. 3, pp. 160–165, Tallinn, Estonia, USSR, August 1990.

[10] M. Tanaka, "Identification of nonlinear systems with missing data using stochastic neural network," in *Proceedings of 35th IEEE Conference on Decision and Control*, vol. 1, pp. 933–934, 1996.

[11] E. Weinstein, M. Feder, and A. V. Oppenheim, "Sequential algorithms for parameter estimation based on the kullback-liebler information measure," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, no. 9, pp. 1652–1654, 1990.

[12] S. Mirsaidi and J. Oksman, "A class of real-time AR identification algorithms in the case of missing observations," in *Proceedings of EUSIPCO-96*, vol. 2, pp. 803–806, Trieste, Italy, 1996.

[13] R. Pintelon and J. Schoukens, "Identification of continuous-time systems with missing data," in *Proceedings of the 16th IEEE Instrumentation and Measurement Technology Conference*, vol. 2, pp. 1081–1085, Venice, Italy, 1999.

[14] A. J. Isaksson, "Identification of ARX models subject to missing data," *IEEE Trans. Automatic Control*, vol. 38, no. 5, pp. 813–819, 1993.

[15] A. J. Isaksson, "A recursive EM algorithm for identification subject to missing data," in *Postprint Volume from the IFAC Symposium on System Identification (SYSID '94)*, vol. 2, pp. 953–958, 1995.

[16] R. Wallin, A. J. Isaksson, and L. Ljung, "An iterative method for identification of ARX models subject to missing data," in *Proceedings of the 39th IEEE Conference on Decision and Control*, Sydney, Australia, December 2000.

[17] P. Stoica and T. Söderström, "Uniqueness of the maximum likelihood estimates of ARMA model parameters—an elementary proof," *IEEE Trans. Automatic Control*, vol. 27, no. 3, pp. 736–738, 1982.
[18] K. J. Åström and B. Wittenmark, *Computer-Controlled Systems*, Prentice Hall, New Jersey, 3rd edition, 1997.
[19] W. J. Rugh, *Linear System Theory*, Prentice Hall, New Jersey, 2nd edition, 1996.
[20] R. S. Bucy and L. A. Campbell, "Determination of steady state behavior for periodic discrete filtering problems," *Computers and Mathematics with Applications*, vol. 15, no. 2, 1988.
[21] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, Addison-Wesley, Massachusetts, 1991.
[22] F. Marvasti, "Nonuniform sampling theorems for bandpass signals at or below the Nyquist density," *IEEE Trans. Signal Processing*, vol. 44, no. 3, pp. 572–576, 1996.
[23] T. Söderström and P. Stoica, *Instrumental Variable Methods for System Identification*, vol. 57 of *Lecture notes in control and information sciences*, Springer-Verlag, Berlin, 1983.

Ragnar Wallin received his M.Sc. degree in electrical engineering and the Licentiat degree in automatic control in 1998 and 2000, respectively, all from the Royal Institute of Technology, Stockholm, Sweden. His research interests are in the systems and control field. The main interest is in system identification.

Alf J. Isaksson received his M.Sc. degree in computer engineering in 1983, and the Licentiat and Ph.D. degrees in automatic control in 1986 and 1988, respectively, all from Linköping University, Sweden. From 1988 to 1992 he was a Lecturer with the Department of Electrical Engineering, Linköping University. Between 1992 and 1998 he was a Lecturer at the Royal Institute of Technology, Stockholm, Sweden. Since 1998 he has been Professor in automatic control at the Royal Institute of Technology. His research interests are general in the systems and control area, including system identification, fault detection and applications of Kalman filtering. Recently his work has been directed more towards process control.