Eindhoven University of Technology
Systems & Control
Control System Group of Electrical Engineering

Identification of Continuous-time ARX-Models Subject to Missing Data

IDNR:1033891
Yang Song

Supervisors: M.C.F.Donkers, G. Bottegal

September 14, 2018

TU/e Technische Universiteit
Eindhoven
University of Technology

Declaration concerning the TU/e Code of Scientific Conduct for the Master's thesis

I have read the TU/e Code of Scientific Conduct[i].

I hereby declare that my Master's thesis has been carried out in accordance with the rules of the TU/e Code of Scientific Conduct

Date

13-09-2018

Name

Yang Song

ID-number

1033893

Signature

Submit the signed declaration to the student administration of your department.

[i] See: http://www.tue.nl/en/university/about-the-university/integrity/scientific-integrity/
The Netherlands Code of Conduct for Academic Practice of the VSNU can be found here also.
More information about scientific integrity is published on the websites of TU/e and VSNU

Identification of Continuous-time ARX-Models Subject to Missing Data

Yang Song
Eindhoven University of Technology
Control Systems Group, Department of Electrical Engineering

Abstract—**This report presents algorithms, theory, and validation results for identification of continuous-time ARX models from incomplete sampled data. The missing data that are expected to have impact on the system identification include both input and output, and require a new state-space model for the estimation. Therefore, different algorithms are developed based on different methods to formulate the equivalent state-space models adapted to the continuous-time context. These algorithms are tested and compared with several identification methods, including Maximum-likelihood and Expectation-maximization. All the simulation are implemented using Monte-Carlo simulations in MATLAB.**

Index Terms—**system identification, maximum-likelihood, expectation-maximization, quasi-newton method, missing data, state-space, sampling, ARX-models**

I. Introduction

SYSTEM Identification is a major field in control and signal processing on building mathematical models of dynamical systems using input and output measurements. Most of the system identification literature [1] [2] deals with discrete-time models due to the advent of the digital computer and the suitability of measured data for digital control systems. Dynamical systems in the physical world are usually described by the differential equations derived from the physical laws, e.g., Newton's first law, Newton's second law, which indicate a continuous-time nature. Continuous-time models provide more advantages than discrete-time models because they provide stronger connection to system properties and the identified parameters of the continuous-time model usually have an immediate physical interpretation. We refer to [3] for some motivating examples for identifying continuous time models from sampled data.

Most of methods developed within this field assume that input-output measurement data are available at every sampling instant and at constant sampling intervals. In some cases, however, there are practical reasons for having incomplete data, e.g., infrequent/ scarce output measurements due to failing sensor, irregularly sampled systems and unexpected interruption in regularly sampling, which may lead to inaccurate estimates. To study the impact of the 'irregular' data and the way to obtain accurate estimates, many studies have been presented. In discrete-time, 'irregular' data show up as missing measurements. Since some measurements are missing, the simplest way is to estimate the parameters only with the available data. Other strategies require a reconstruction process, which means to rebuild the 'complete' data, using methods such as, the Kalman filter and the fixed-interval smoothing. This method was discussed in [4] together with the Expectation-Maximization (EM) algorithm. Other papers, such as [5], [6] also presented a frequency domain solution to the system identification problem with missing data.

In contrast, identification of continuous-time systems require more involved steps, and 'irregular' data show up in different forms. Irregular sampling is often studied; in [9], [11], continuous-time models without input are considered. In [15], a method is considered to estimate continuous-time auto-regressive models with exogenous inputs (ARX) using the approximation of the differentiation operator under time-varying sampling rates. Missing data have been discussed in [10] in frequency domain by treating them as unknown parameters in the identification problem. In general, few contributions were made for continuous-time system identification with incomplete data in time domain.

In this thesis, continuous-time ARX (auto-regressive models with exogenous variables) are chosen to study the identification problem since ARX models are the simplest choice to describe a dynamic process driven by an input with uncertainties. As mentioned above, digital signals are used nowadays to control physical dynamic systems, so we derive several algorithms to formulate the estimation model based on continuous-time inputs and discrete-time inputs. Moreover two methods will be discussed to accommodate for missing data.

The report is organized as follows. In Section II, we present the objective of the estimation and problem formulation. Then in Section III, we introduce two different methods to formulate state-space equivalent model. How to reconstruct the missing data using Kalman filter is discussed in Section IV. Maximum-Likelihood method is presented in Section V. Then two methods of dealing with missing data will be discussed in Section VI.Section VII presents some numerical illustrations. Concluding remarks are finally given in Section VIII.

II. Problem Formulation

Let us consider a continuous-time ARX (CARX) model that can be described using the general-linear polynomial form. This model provides flexibility for both system dynamics and stochastic dynamics, using the equation:

$$A(p)y(t) = B(p)u(t) + e(t) \tag{1}$$

Supervisors: M.C.F.Donkers, G.Bottegal

with

$$A(p) = p^n + a_1 p^{n-1} + a_2 p^{n-2} + \cdots + a_n$$
$$B(p) = b_1 p^{n-1} + b_2 p^{n-2} + \cdots + b_n,$$

where $t \in \mathbb{R}$ represents time, $u(t)$ is the input, $y(t)$ the output, $e(t)$ a zero mean white noise process with $\mathbb{E}\left\{e^2(t)\right\} = \sigma_e^2$ and p represents derivative so that $p^n u(t) = \frac{d^n u(t)}{dt^n}$. The objective is to identify this CARX plant, assuming zero-order hold, which indicates that a constant value $u(t_k)$ holds for $t_k \leq t < t_{k+1}$, $k \in \{0, 1, ..., N-1\}$. The input and output are collected at time $t_1, t_2, t_3, ..., t_N$.

A. Input Model

As mentioned in the introduction, we consider the case that both input and output data can be incomplete. It has been studied in Isaksson's work [4] that a new variable $z(t) = \begin{bmatrix} y(t) & u(t) \end{bmatrix}^T$ is the key to convert the problem to a form suitable for any of the missing-output methods. Therefore, it is reasonable to assume that the system operates in open loop and input $u(t)$ is also generated via an autoregressive model. A natural assumption would be to build a continuous-time model since we want to identify a continuous-time system. We can model the input as an auto-regressive stochastic process of the form

$$C^c(p)u(t) = v(t) \tag{2}$$

with

$$C^c(p) = p^{m_1} + c_1^c p^{m_1-1} + c_2^c p^{m_1-2} + \cdots + c_{m_1}^c,$$

and where $v(t)$ is continuous-time stationary white process noise, which is assumed to be Gaussian with zero mean and intensity zero mean white noise process with $\mathbb{E}\left\{v^2(t)\right\} = \sigma_v^2$.

On the other hand, we can also consider the case where the input is generated by a computer so that we can assume a discrete-time model for input, assuming zero-order hold. Then, we have that

$$C^d(q)u(k) = w(k) \tag{3}$$

with

$$C^d(q) = 1 + c_1^d q^{-1} + c_2^d q^{-2} + \cdots + c_{m_2}^d q^{-m_2}$$

where $k \in \mathbb{N}$ represents time, $w(k)$ is a zero mean white noise sequence with $\mathbb{E}\left\{w^2(k)\right\} = \sigma_w^2$, and q represents a time shift operator so that $u(k-1) = q^{-1}u(k)$.

The problem we want to solve is to find the coefficients a_k, b_k, and c_k (c^d or c^c) based on incomplete sampled output $y(k)$ and input $u(k)$. If we just ignore the incomplete data we can only consider the available data without reconstruction, using the least-squares method. The obvious drawback is that with an unfavorable pattern of missing data, the number of samples with complete information may be very small, leading to biased estimates. The details can be found in next part.

B. Motivating example

As mentioned above, we can estimate the system with the available data using least-squares method. The strategy is to identify a discrete-time ARX model first using sampled data then convert it to continuous time. Introduce the discrete-time ARX model

$$A^l(q)y(k) = B^l(q)u(k) + e^l(k), \tag{4}$$

with:

$$\begin{aligned} A^l(q) &= 1 - a_1^l q^{-1} - a_2^l q^{-2} - a_3^l q^{-3} - ... - a_n^l q^{-n} \\ B^l(q) &= b_1^l q^{-1} + b_2^l q^{-2} + b_3^l q^{-3} + ... + b_m^l q^{-m}, \end{aligned} \tag{5}$$

where $k \in \mathbb{N}$ represents time, $u(k)$ is the input, $y(k)$ the output, $e^l(k)$ a white noise process. The DARX can also be written as:

$$\begin{aligned} y(k) &= a_1^l y(k-1) + ... + a_n^l y(k-n) \\ &\quad + b_1^l u(k-1) + ... + b_m^l u(k-m) + e^l(k). \end{aligned} \tag{6}$$

The parameters and measurements are formulated as vectors

$$\theta = \begin{bmatrix} a_1^l \\ \vdots \\ a_n^l \\ b_1^l \\ \vdots \\ b_m^l \end{bmatrix} \qquad \phi(k) = \begin{bmatrix} y(k-1) \\ \vdots \\ y(k-n) \\ u(k-1) \\ \vdots \\ u(k-m) \end{bmatrix}$$

With these new vectors, the model in (6) can be rewritten as

$$y(k) = \phi^T(k)\theta + e^l(k) \tag{7}$$

We call the the parameter we want to estimate $\hat{\theta}$, The best set of parameters $\hat{\theta}$ is the one that minimizes the sum of the squared values of e^l. This minimization problem has a unique solution, solved by the normal equation:

$$\left(\frac{1}{N}\sum_{k=1}^{N}\phi(k)\phi^T(k)\right)\hat{\theta} = \left(\frac{1}{N}\sum_{k=1}^{N}\phi(k)y(k)\right). \tag{8}$$

Finally the parameter vector is given by

$$\hat{\theta} = \left(\frac{1}{N}\sum_{k=1}^{N}\phi(k)\phi^T(k)\right)^{-1}\left(\frac{1}{N}\sum_{k=1}^{N}\phi(k)y(k)\right). \tag{9}$$

The parameters can be estimated if we have all the measurements. If there are missing observations, the method is to do the summation only over values such that both $y(k)$ and $\phi(k)$ are complete. After completing $\hat{\theta}$, we can obtain the parameters of the continuous-time model by doing 'd2c' in Matlab. Because $\phi(k)$ contains previous information of $y(k)$ and $u(t)$, when one point is missing, for example $u(5)$, then $\phi(6)$, $\phi(7)$, ..., $\phi(5+m)$ are also incomplete, which means that the number of samples with complete information may be very small if the missing data is in non-ideal pattern (e.g. lose data every n moment may cause disappearance of the complete $\phi(k)$) and lead to estimation that is not ideal.

Motivating example: Consider the following CARX process

$$(p+2)y(t) = 50u(t) + e(t) \tag{10}$$

with the parameter vector given by $\theta_0 = [2 \quad 25]^T$; the spectral power of $e(t)$ is $\sigma_v^2 = 1$. Assume that there are $N = 100$ samples under the sampling time $T_s = 1$.

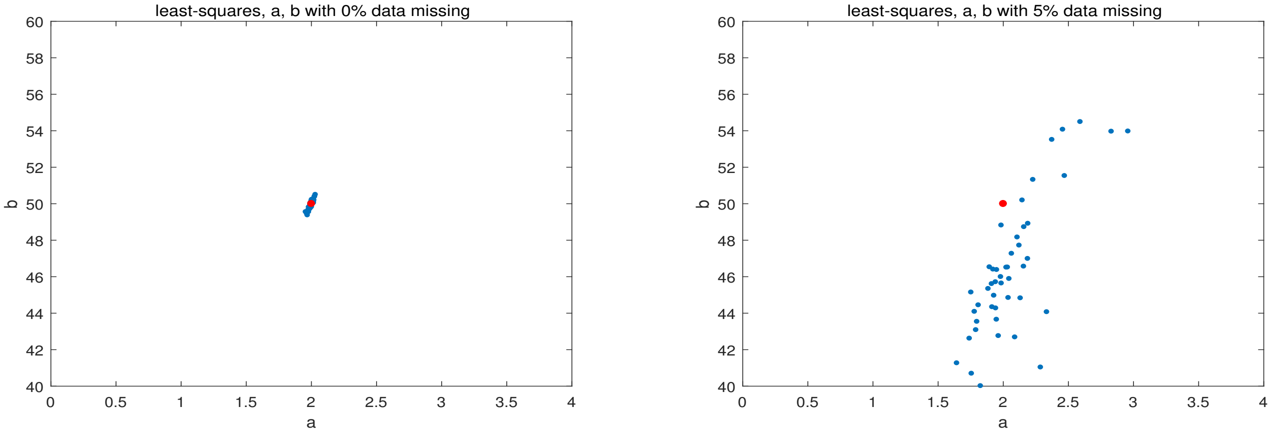

Fig. 1. Plots of the Least-squares method for system (10). Left - no missing data. Right - with 5% data missing

It can be seen in Fig. 1. that least-squares give a good estimation result when measurements are complete. When there is missing data, even with small percentage, least squares will lead to a biased estimates with high variance.

III. State-space formulation

We have already seen in the last section that considering only the available data will lead to an inaccurate estimation with big variance, so filling in the missing data seems necessary. In order to reconstruct the missing observations, we need to formulate a state-space model first. The CARX model (1) can be described in a state-space framework as

$$\begin{aligned}\dot{x}_y(t) &= A_y x_y(t) + B_y u(t) + E_y e(t)\\ y(t) &= C_y x_y(t)\end{aligned} \tag{11}$$

In this thesis, the observable canonical form is chosen, and hence $A_y \in \mathbb{R}^{n\times n}$, $B_y \in \mathbb{R}^{n\times 1}$, $E_y \in \mathbb{R}^{n\times 1}$ and $C_y \in \mathbb{R}^{1\times n}$ have the structures with

$$A = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0\\ -a_2 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ -a_{n-1} & 0 & 0 & \cdots & 1\\ -a_n & 0 & 0 & \cdots & 0 \end{bmatrix} B_y = \begin{bmatrix} b_1\\ b_2\\ b_3\\ \vdots\\ b_n \end{bmatrix} E_y = \begin{bmatrix} 1\\ 0\\ 0\\ \vdots\\ 0 \end{bmatrix}$$

$$C_y = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Here we consider two cases as mentioned in Section II-A, namely a continuous-time input model and a discrete-time input model. We will show later that the continuous-time input model leads to large variance if the inputs are generated assuming zero-order hold.

A. *Continuous-time u(t)*

We first consider a continuous-time input. The input signal in (2) is written in the observable canonical form

$$\begin{aligned}\dot{x}_u^c(t) &= A_u^c x_u^c(t) + B_u^c v(t)\\ u(t) &= C_u^c x_u^c(t)\end{aligned}$$

where $A_u^c \in \mathbb{R}^{m_1\times m_1}$, $B_u^c \in \mathbb{R}^{m_1\times 1}$ and $C_u^c \in \mathbb{R}^{1\times m_1}$ are defined as

$$A_u^c = \begin{bmatrix} -c_1^c & 1 & 0 & \cdots & 0\\ -c_2^c & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ -c_{m_1-1}^c & 0 & 0 & \cdots & 1\\ -c_{m_1}^c & 0 & 0 & \cdots & 0 \end{bmatrix} B_u^c = \begin{bmatrix} 1\\ 0\\ 0\\ \vdots\\ 0 \end{bmatrix} C_u^c = \begin{bmatrix} 1\\ 0\\ 0\\ \vdots\\ 0 \end{bmatrix}^T.$$

Let

$$x(t) = \begin{bmatrix} x_y(t)\\ x_u(t) \end{bmatrix}, \quad z(t) = \begin{bmatrix} y(t)\\ u(t) \end{bmatrix}, \quad \gamma(t) = \begin{bmatrix} e(t)\\ v(t) \end{bmatrix};$$

then we can obtain complete state-space model:

$$\begin{aligned}\dot{x}(t) &= A_c x(t) + K\gamma(t)\\ z(t) &= Cx(t),\end{aligned}$$

where

$$A_c = \begin{bmatrix} A_y & B_y C_u^c\\ 0 & A_u^c \end{bmatrix}, \quad K = \begin{bmatrix} E_y & 0\\ 0 & B_u^c \end{bmatrix}, \quad C = \begin{bmatrix} C_y & 0\\ 0 & C_u^c \end{bmatrix},$$

and the resulting noise $\gamma(t)$ becomes $[e(t) \quad v(t)]^T$, with covariance

$$\mathbb{E}\left\{\gamma(t)\gamma^T(t)\right\} = \Lambda = \begin{bmatrix} \sigma_e^2 & 0\\ 0 & \sigma_v^2 \end{bmatrix}. \tag{12}$$

Notice that the noise is ideally modeled as completely white in order to obtain the Markov property. However, in order to implement the proposed identification schemes, the continuous-time model has to be discretized based on the sampling interval h. Such discretization yields to the equivalent discrete-time model:

$$\begin{aligned}x(k+1) &= F_1 x(k) + \eta(k)\\ z(k) &= H_1 x(k)\end{aligned}, \tag{13}$$

where $F_1 = e^{A_c h}$ (see [16]), $H_1 = C$, and $\eta(k)$ is zero mean discrete-time white noise with covariance matrix Q_1 that includes solving an integral involving the matrix exponential on the form [16]:

$$Q_1 = \int_0^h e^{A_c\tau}\Lambda e^{A_c^T\tau} d\tau. \tag{14}$$

Notice that the discretization of process noise includes an integral, which is difficult to implement in simulation. Here we adopt the method present in [17] to compute it in an efficient way. First we define a square matrix T as follows:

$$T = \begin{bmatrix} -A_c & K\Lambda K^T \\ 0 & A_c^T \end{bmatrix}, \tag{15}$$

and denote its exponential matrix by

$$T_k = e^{Th} = \begin{bmatrix} \cdots & A_d^{-1}Q \\ 0 & A_d^T \end{bmatrix}. \tag{16}$$

Then we can obtain the discretized process noise Q evaluated by multiplying A_d with the upper-right partition of T_k.

B. Discrete-time u(k)

In the second case, we assume a discrete-time AR model for the input. The input signal in (3) is written on the form

$$\begin{aligned} x_u^d(k+1) &= A_u^d x_u^d(k) + B_u^d w(k) \\ u(k) &= C_u^d x^u(k) + w(k) \end{aligned},$$

where $A_u^d \in \mathbb{R}^{m_2 \times m_2}$, $B_u^d \in \mathbb{R}^{m_2 \times 1}$ and $C_u^d \in \mathbb{R}^{1 \times m_2}$ are defined as

$$A_u^d = \begin{bmatrix} -c_1^c & -c_2^c & \cdots & -c_{m_2-1}^c & -c_{m_2}^c \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \quad B_u^d = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$C_u^d = \begin{bmatrix} -c_1^c & -c_2^c & \cdots & -c_{m_2}^c \end{bmatrix}.$$

For the same reason, the continuous-time model (11) has to be discretized based on the sampling interval h, assuming zero-order hold for the input u and continuous integration for the noise e,

$$\begin{aligned} x_y(k+1) &= A_y^d x_y(k) + B_y^d u(k) + \xi(k) \\ y(k) &= C_y x_y(k) \end{aligned} \tag{17}$$

with $A_y^d = e^{A_y h}$, $B^d = A_y^{-1}(A_y^d - I)B_1$ ([16]), $\xi(k)$ is zero mean discrete-time white noise with covariance matrix given by

$$Q_2 = \int_0^h e^{A_y \tau} \sigma_e^2 e^{A_y^T \tau} d\tau.$$

Let

$$x(k) = \begin{bmatrix} x_y(k) \\ x_u^d(k) \end{bmatrix}, \quad z(k) = \begin{bmatrix} y(k) \\ u(k) \end{bmatrix}, \quad \delta(k) = \begin{bmatrix} \xi(k) \\ w(k) \end{bmatrix};$$

then we can obtain complete state-space model:

$$\begin{aligned} x(k+1) &= F_2 x(k) + G\delta(k) \\ z(k) &= H_2 x(k) + L\delta(k) \end{aligned}, \tag{18}$$

with

$$\begin{aligned} F_2 &= \begin{bmatrix} A_y^d & B_y^d C_u^d \\ 0 & A_u^d \end{bmatrix} & G &= \begin{bmatrix} I & B_y^d \\ 0 & B_u^d \end{bmatrix} \\ H_2 &= \begin{bmatrix} C_y & 0 \\ 0 & C_u^d \end{bmatrix} & L &= \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \end{aligned}. \tag{19}$$

We can see that the noise $\delta(k)$ acts in both the process and measurement models, which means that the process noise is correlated with the measurement noise. This correlation requires a full form of Kalman filter equations, which will be discussed in following section.

IV. Reconstruction

We have introduced the state-space formulation in the previous section, which can be used to reconstruct the missing observations. The reconstruction should be based on previous measurements, i.e., the estimates of a Kalman filter, which is a well-known tool to estimate the states of linear systems using the available measurements. We take model (18) as example, substituting the measurement noise $L\delta(k)$ with $\zeta(k) = \begin{bmatrix} 0 & w(k) \end{bmatrix}^T$. Then

$$\begin{aligned} x(k+1) &= F_2 x(k) + G\delta(k) \\ z(k) &= H_2 x(k) + \zeta(k) \end{aligned}.$$

We know that the system behaves according to the state-space equations, and we have measurements $z(k)$. Then the Kalman filter gives the estimate of $x(k)$ with smallest possible error variance. Notice that the process noise $\delta(k)$ and $\zeta(k)$ are assumed to be Gaussian with

$$\mathbb{E}\left\{ \begin{bmatrix} \delta(k) \\ \zeta(k) \end{bmatrix} \begin{bmatrix} \delta^T(k) & \zeta^T(k) \end{bmatrix} \right\} = \begin{bmatrix} R_1 & R_{12} \\ R_{21} & R_2 \end{bmatrix}.$$

We obtain $R_{12} \neq 0$ as mentioned before, so that the process noise is correlated with the measurement noise. The predicted and filtered estimates will be denoted as $\widehat{x}(k+1|k)$ and $\widehat{x}(k+1|k+1)$, corresponding covariance matrices are denoted by $P(k+1|k)$ and $P(k+1|k+1)$. The Kalman filter equation with correlated noise are shown as follows [19]:
TIME UPDATE:

$$\widehat{x}(k+1|k) = F_c(k)\widehat{x}(k|k) + G(k)R_{12}(k)R_2^{-1}(k)z(k) \tag{20}$$

and

$$\begin{aligned} P(k+1|k) &= F_c(k)P(k|k)F_c^T(k) \\ &\quad + G(k)(R_1(k) - R_{12}(k)R_2^{-1}(k)R_{12}(k)^T)G(k)^T, \end{aligned} \tag{21}$$

where

$$F_c(k) = (F_2(k) - G(k)R_{12}(k)R_2^{-1}(k)H_2(k)). \tag{22}$$

MEASUREMENT UPDATE:

$$\begin{aligned} \widehat{x}(k+1|k+1) &= \widehat{x}(k+1|k) + P(k+1|k)H_2^T(k+1) \\ &\quad \times S_c(k)^{-1}\epsilon(k+1) \end{aligned} \tag{23}$$

where $\epsilon(k)$ represents the innovation or prediction error

$$\epsilon(k) = (z(k) - H_2(k)\widehat{x}(k|k-1)) \tag{24}$$

and

$$\begin{aligned} P(k+1|k+1) &= P(k+1|k) - P(k+1|k) \\ &\quad \times H_2^T(k+1)S_c(k)^{-1}H_2(k+1)P(k+1|k), \end{aligned} \tag{25}$$

where

$$S_c(k) = H_2(k+1)P(k+1|k)H_2^T(k+1) + R_2(k+1). \tag{26}$$

The quantities $\epsilon(k)$ and $S_c(k)$ will be used in the next Section to derive the developed identification technique based on maximum likelihood.

V. Maximum Likelihood

In this section we will discuss the estimation approaches for the system parameters. The maximum likelihood method provides estimates of the parameter values based on an observed data set $Z_N = Z(1), Z(2), ..., Z(N)$ by maximizing a likelihood function. In order to use this method it is therefore necessary to first derive an expression for the likelihood function. We suppose that there is a series of independent identically distributed samples x_1, x_2, ..., x_N ($x_k \in \mathbb{R}^{n\times 1}$), coming from a Gaussian distribution with probability density function

$$f_\theta(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \tag{27}$$

where $\mu \in \mathbb{R}^{n\times 1}$ is the mean vector, $\Sigma \in \mathbb{R}^{n\times n}$ is the covariance matrix. A *likelihood* for a model is defined by the same equation expression as the probability density, but the roles of the data x and the parameter θ are interchanged.

$$L_{x_i}(\theta) = f_\theta(x_i). \tag{28}$$

Then the likelihood function for the full set of measurement can be shown as follows:

$$\begin{aligned} L(\theta|x_1, x_2, ..., x_N) &= f(x_1, x_2, ..., x_N|\theta) \\ &= f(x_1|\theta)\cdot(x_2|\theta)\cdot ...\cdot f(x_N|\theta)' \end{aligned} \tag{29}$$

The so-called method of maximum likelihood is an estimator of the unknown true parameter value θ. The point $\hat{\theta}$ that maximizes the likelihood function L is the final estimate we want. This estimator is called the maximum likelihood estimator (MLE). In practice, it is often more convenient when working with the natural logarithm of the likelihood function, (using the rule $\log a\cdot b = \log a + \log b$), namely the log-likelihood

$$\log L(\theta|x_1, x_2, ..., x_N) = \sum_{i=1}^{N} \log f(x_i|\theta), \tag{30}$$

where log is the natural logarithm. In our case, the probability density function $f(x_i|\theta)$ can be completed using the Kalman filter mentioned above. According to the papers [4], [13] and [14], the prediction errors $\epsilon(k)$ form a sequence uncorrelated in time. If the noise in the state-space model is Gaussian, then $\epsilon(k)$ is also Gaussian distributed. Let Z^N be the complete batch of observations. The likelihood function is then

$$L(\theta|Z^N) = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\det S_c(k)}} e^{-\frac{1}{2}\epsilon^T(k)S_c^{-1}(k)\epsilon(k)} \tag{31}$$

where $S_c(k)$ is the covariance matrix of the prediction error and $\epsilon(k)$ is the innovation as mention in last section.

$$S_c(k) = H(k)P(k|k-1)H^T(k) + R_2(k) \tag{32}$$

Then, we can calculate the log-likelihood function based on (30) as follows:

$$\begin{aligned} \log L(\theta|Z^N) = C - \frac{1}{2}\sum_{k=1}^{N} \log(\det S_c(k)) \\ -\frac{1}{2}\sum_{k=1}^{N} \epsilon^T(k)S_c^{-1}(k)\epsilon(k) \end{aligned} . \tag{33}$$

Let $W(k)$ be

$$W(k) = \frac{1}{2}\sum_{k=1}^{N} \log(\det S_c(k)) + \frac{1}{2}\sum_{k=1}^{N} \epsilon^T(k)S_c^{-1}(k)\epsilon(k). \tag{34}$$

Finding the maximum of $\log L(\theta|Z^N)$ is equivalent to finding the minimum of $W(k)$, which can be done, e.g, using a quasi-Newton method [20]. The key to quasi-Newton method is the calculation of the Hessian matrix H_k, which normally requires the calculation of second derivative. Instead of using the true Hessian matrix, a recursive process is chosen to do updating. In this thesis, we use BFGS formulation suggested by Broyden, Fletcher, Goldfarb, and Shanno [20] to calculate the update matrix.

VI. Missing data formulation

In this Section, we discuss two ways to incorporate missing data.

A. Isaksson's $D(k)$ together with doubling sampling time

We first consider to remove the data directly. Define $D(k)$ as a matrix reducing the number of rows in $z(k)$. If, for example if the measurement $u(k)$ is missing, we set

$$D(k) = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

Denoting the measured part of $z(k)$ by $z_m(k)$ the measurement equation is

$$z_m(k) = D(k)z(k) = y(k)$$

. In this way, we simply omit the row in the measurement equation corresponding to the missing observation. Similarly, when the measurement $y(k)$ is missing, we set

$$D(k) = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

so that

$$z_m(k) = D(k)z(k) = u(k)$$

If, for example, no measurement is missing, then

$$D(k) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{35}$$

How the incomplete data influence the likelihood function (33) is shown as follows:

$$\begin{aligned} R_1 &= \Lambda, \\ R_2(k) &= D(k)\Lambda D^T(k), \\ R_{12}(k) &= \Lambda D^T(k), \\ H(k) &= D(k)H, \end{aligned}$$

$$\epsilon(k) = z_m(k) - H(k)\widehat{x}(k|k-1).$$

We can see from above that the missing data change the value of $S_c(k)$ and $\epsilon(k)$. The tricky part is when both $y(k)$ and $u(k)$ are missing at certain sampled time k. Since one can not remove all the rows in $z(k)$ using this method, we consider skipping the missing measurements by doubling the sampling time. A different sampling time will change the value of state matrix and covariance in (13) or (17), which also influence the Kalman filter and the ML estimation. Notice that for the model with discrete-time input, we get a similar model to [4] after the discretization (18) by substituting the measurement noise with $\zeta(k) = \begin{bmatrix}0 & w(k)\end{bmatrix}^T$. Then

$$\begin{aligned} x(k+1) &= F_2 x(k) + G\delta(k) \\ z(k) &= H_2 x(k) + \zeta(k) \end{aligned}$$

The covariances for the process and the measurement noises are given by R_1 and R_2 respectively, and the cross covariance by R_{12}, so that

$$\begin{bmatrix} R_1 & R_{12} \\ R_{12} & R_2 \end{bmatrix} = \mathbb{E}\left\{ \begin{bmatrix} \delta(k) \\ \zeta(k) \end{bmatrix} \begin{bmatrix} \delta^T(k) & \zeta^T(k) \end{bmatrix} \right\},$$

$$R_1 = \mathbb{E}\left\{ \begin{bmatrix} \xi^2(k) & \xi(k)w(k) \\ w(k)\xi(k) & w^2(k) \end{bmatrix} \right\} = \begin{bmatrix} Q_2 & 0 \\ 0 & \sigma_w^2 \end{bmatrix},$$

$$R_{12} = \mathbb{E}\left\{ \begin{bmatrix} 0 & \xi(k)w(k) \\ 0 & w^2(k) \end{bmatrix} \right\} = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_w^2 \end{bmatrix},$$

$$R_{21} = \mathbb{E}\left\{ \begin{bmatrix} 0 & 0 \\ w(k)\xi(k) & w^2(k) \end{bmatrix} \right\} = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_w^2 \end{bmatrix},$$

$$R_2 = \mathbb{E}\left\{ \begin{bmatrix} 0 & 0 \\ 0 & w^2(k) \end{bmatrix} \right\} = \begin{bmatrix} 0 & 0 \\ 0 & \sigma_w^2 \end{bmatrix}.$$

B. Infinite Variance

The second method is to define $D(k)$ as a diagonal matrix with entries on the i-th diagonal element when i-th element of $z(k)$ is available. In our case the observation $z(k) = \begin{bmatrix} y(k) & u(k) \end{bmatrix}^T$, so $D(k) \in \mathbb{R}^{2\times 2}$, $D(k)$ is used to give infinite value to specific element of R_2 and R_{12}. Here we assume that a missing measurement can be regarded as a measurement with noise with infinite variance.

For example, if $y(k)$ is missing,

$$D(k) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \quad D^{-1}(k) = \begin{bmatrix} \infty & 0 \\ 0 & 1 \end{bmatrix}$$

where we have introduced a special notation for the inverse of $D(k)$. In this way we have that

$$R_2(k) = D^{-1}(k) R_2 D^{-1}(k)^T,$$

$$R_{12}(k) = R_{12} D^{-1}(k)^T.$$

. The ‘infinity’ element can not be implement in simulations but we noticed that R_2 always comes up with inverse in Kalman filter equations (20) $\sim$ (26),

$$R^{-1}(k) = D(k)^T R_2^{-1} D(k)$$

Similarly, if $u(k)$ is missing, then

$$D(k) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

if both $u(k)$ and $y(k)$ are missing, then

$$D(k) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

The tricky part is when both $y(k)$ and $u(k)$ are missing, it yields to a zero determinant $S_c(k)$, which can not be calculated under the natural logarithm. Hence, we implement this method only with Expectation-Maximization algorithm, which is shown in Appendix A.

VII. SIMULATION RESULTS

Two cases mentioned in Section III are simulated with a simple continuous-time ARX model. The true system model is chosen as

$$A(p)y(t) = B(p)u(t) + e(t) \tag{36}$$

with

$$\begin{aligned} A(p) &= p + a \\ B(p) &= b, \end{aligned}$$

where the parameter vector is $\theta_0 = \begin{bmatrix} a & b \end{bmatrix}^T = \begin{bmatrix} 2 & 50 \end{bmatrix}^T$. $e(t)$ is zero mean continuous-time white noise with a priori known variance $\mathbb{E}\left\{v^2(t)\right\} = 1$. The system was sampled with sampling time $0.5s$, assuming zero-order hold for input and it was simulated for 50 noise realizations, each with N data points and $M\%$ missing data.

A. Simulation 1: continuous-time input model

In simulation 1, we considered a continuous-time input model with $N = 100$, $M = 0, 10, 20, 50$. A simple continuous-time AR model is chosen:

$$C(p)u(t) = v(t), \tag{37}$$

with

$$C(p) = p + c.$$

$w(t)$ is zero mean continuous-time white noise with spectral power $\mathbb{E}\left\{v^2(t)\right\} = 1$. Then the parameter vector is $\theta_0 = \begin{bmatrix} a & b & c \end{bmatrix}^T = \begin{bmatrix} 2 & 50 & 1 \end{bmatrix}^T$. The estimation result is shown in Figure. 2. Mean and variance of the estimates can be found in Table I. The figure shows how this model response to different percentage of missing data.

In this case, continuous-time input model formulation yields to unbiased estimation with big variance for both a and b even under no missing data. Moreover, the estimation result remains accurate when the missing percentage reaches 50%.

B. Simulation 2: discrete-time input model

In simulation 2, we considered a discrete-time input model with $N = 100$, $M = 0, 10, 20, 50$. A simple discrete-time AR model for input:

$$C(q)u(k) = w(k), \tag{38}$$

with

$$C(q) = 1 + cq^{-1}.$$

$w(k)$ is zero mean discrete-time white noise with spectral power $\mathbb{E}\left\{w^2(k)\right\} = 0.6$. Then the parameter vector is $\theta_0 = \begin{bmatrix} a & b & c \end{bmatrix}^T = \begin{bmatrix} 2 & 50 & -0.18 \end{bmatrix}^T$. The estimation result is shown in Figure. 3. Mean and variance of the estimates can be found in Table II. The figure shows how well the model is identified when we have different percentage of missing data.

In this case, discrete-time input model apparently performs better than the other alternatives under modest amount of missing data. The variance of parameter a is one magnitude smaller than the continuous-time model when 0% and 10 % data are removed. On the other hand, this model is more sensitive to incomplete data, it can been seen that the variance increases greatly when we add 10% missing measurements every time. Moreover, the estimation can be regarded as failure (inaccurate mean and big variance) when there are 50% data missing.

C. *Simulation 3: Discrete-time ARX model*

In simulation 3, we identified a discrete-time ARX model using the sampled data (same data set as simulation 2). The estimation was done by applying the EM algorithm, which is discussed in Appendix A. The estimation result is shown in Figure. 4. Mean and variance of the estimates can be found in Table III. We can see from the figure that the EM algorithm estimation perform equally well compared to quasi-Newton optimization of the maximum likelihood (see simulation 2) when data are complete. This also corresponds to the conclusion in [4] that EM is an alternative way to calculate the maximum of likelihood function. We also noticed a jump in variance when measurements start missing, the variance of b increased from 0.0592 to 5.2796 within 10% missing measurements.

D. *Simulation 4: Same data points*

In simulation 4, we took the model in simulation 2 and investigated the influence of missing data by considering data sets that have the same number of data points, but different time spans. The result is shown in Figure. 5 and Table IV, which contains 100 sampled data with no missing percentage, 112 sampled data with 10% missing, 125 sampled data with 20% missing and 200 sampled data with 50% missing respectively. These four cases result in the same amount of available data (100). It is clear that the accuracy depends more on the missing percentage. 200 sampled measurements with 50% missing data gives worse estimation. On the other hand, we can also compare the result under same proportion of missing data, e.g. under 50% missing condition, 200 sampled data performs better than 100 sampled data.

VIII. Conclusion

The problem of estimating the parameters in CARX processes with input generated with ZOH from incomplete sampled data is considered and a direct approach under two different input models is suggested as solution. We first show that

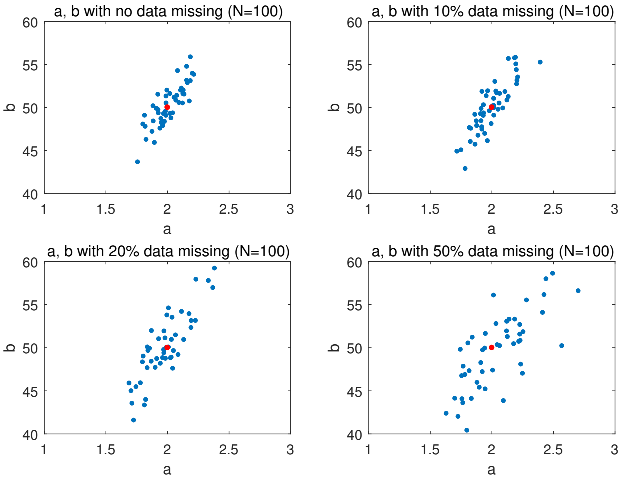

Fig. 2. Continuous-time input model with different percentage of data missing.

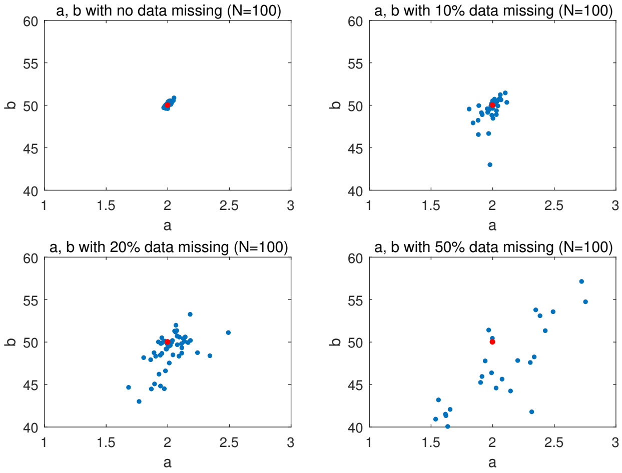

Fig. 3. Discrete-time input model with different percentage of data missing.

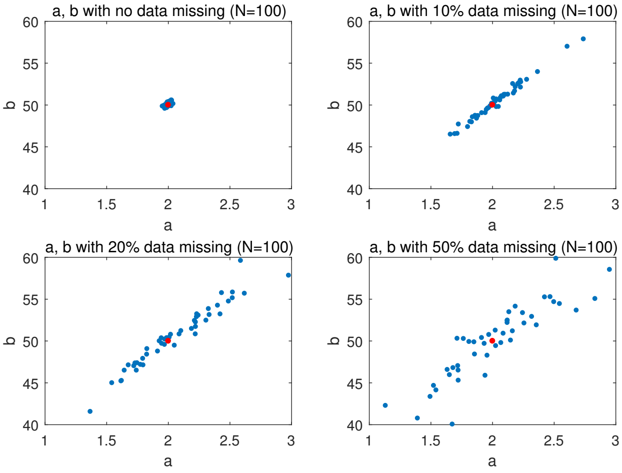

Fig. 4. Discrete-time ARX model with different percentage of data missing.

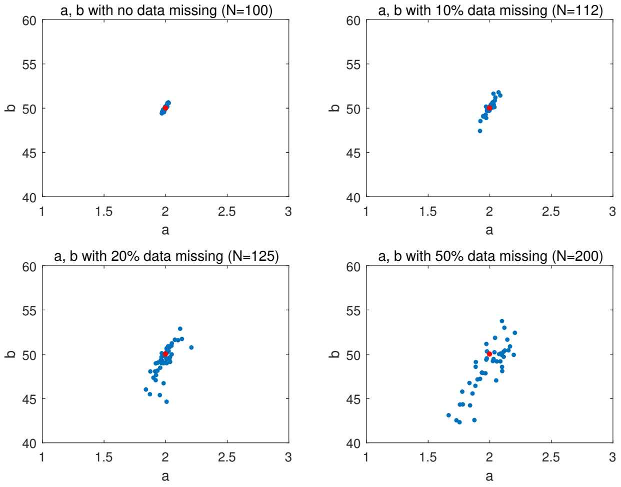

Fig. 5. Discrete-time input model with different percentage of data missing under same data points.

TABLE I
SIMULATION RESULTS OF CONTINUOUS-TIME INPUT MODEL

CAR input		a	b	c
No missing	mean	2.0177	50.3370	1.0632
	variance	0.0143	5.9528	0.0384
10 % missing	mean	1.9989	50.0539	0.9894
	variance	0.0193	9.0865	0.0386
20 % missing	mean	1.9796	50.0528	1.1069
	variance	0.0285	13.9407	0.0736
50 % missing	mean	1.9844	50.0311	1.0565
	variance	0.0728	27.4808	0.1085

TABLE II
SIMULATION RESULTS OF DISCRETE-TIME INPUT MODEL

DAR input		a	b	c
No missing	mean	2.0063	50.0733	-0.1818
	variance	$2.9335e^{-4}$	0.0799	0.0144
10 % missing	mean	1.9941	49.6341	-0.1616
	variance	0.0034	1.8406	0.0130
20 % missing	mean	2.0265	48.9531	-0.1784
	variance	0.0187	4.5063	0.0106
50 % missing	mean	1.6767	38.8121	-0.0759
	variance	0.3211	116.3561	0.0127

TABLE III
SIMULATION RESULTS OF DISCRETE-TIME ARX MODEL

DARX model		a	b	c
No missing	mean	1.9998	50.0627	*
	variance	$3.8404e^{-4}$	0.0592	*
10 % missing	mean	2.0326	50.4140	*
	variance	0.0432	5.2796	*
20 % missing	mean	2.0616	50.5909	*
	variance	0.1052	12.5770	*
50 % missing	mean	2.1454	52.6127	*
	variance	0.2893	41.9494	*

TABLE IV
SIMULATION RESULTS OF SAME DATA POINTS

Same data points		a	b	c
M=0, N=100	mean	2.0006	50.0046	-0.1885
	variance	$1.4966e^{-4}$	0.0745	0.0043
M=10, N=112	mean	2.0048	50.0531	-0.1853
	variance	$9.8574e^{-4}$	0.5187	0.0046
M=20, N=125	mean	1.9956	49.2763	-0.1766
	variance	0.0045	2.8745	0.0034
M=50, N=200	mean	1.9273	47.0091	-0.1605
	variance	0.0493	21.2840	0.0138

Least-squares leads to a biased result. The method presented then is based on that the missing data can be reconstructed by the estimates of a Kalman filter.

Numerical studies show that the discrete-time input model gives a better estimation when not too many data missing. The continuous-time input model yields to a estimation with large variance even with little missing data. For sake of comparison, we also propose to estimate the parameters by identifying a discrete-time model using sampled data and then convert it to continuous time, which can be regarded as indirect approach. This method gives a similar result to discrete-time input model under complete data. A disadvantage with discrete-time input model, though, is that its performance are more sensitive to missing data than continuous-time input model, giving inaccurate estimates when half measurements are lost. We have also verified that larger amount of data points lead to better performance.

REFERENCES

[1] Astrom K.J., T. Bohlin. Numerical identification if linear dynamic systems from normal operating records. *Theory of Self- Adaptive Control Systems*, Plenum Press, New York , 1966

[2] Box G.E.P., G.M. Jenkins. Time Series Analysis, Forecasting and Control. *Holden Day*, 1970

[3] H. Garnier, L. Wang, Identification of Continuous-time Models from Sampled Data. *Springer-Verlag London*, 2008, no. 1.

[4] A.J. Isaksson. Identification of ARX-models subject to missing data. *IEEE Transactions on Automatic Control* , vol. 38, Issue. 5, May 1993.

[5] R. Pintelon and J. Schoukens, Frequency domain system identification with missing data, *IEEE Transactions on Automatic Control*, vol. 45, no. 2, pp. 364-369, Feb. 2000.

[6] W. Tang, X. Zheng, J. Wu and L. Wang, State-space identification in frequency domain from missing measurements, *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, Beijing, 2017, pp. 6933-6939.

[7] Rolf Johansson. Continuous-Time Model Identification and State Estimation Using Non-Uniformly Sampled Data. *IFAC Proceedings Volumes*, vol 42, Issue 10, 2009, pages 1163-1168

[8] Feng Ding and Jie Ding. Least-squares parameter estimation for systems with irregularly missing data. *Wiley InterScience*, DOI: 10.1002/ asc. 1141, 18 August 2009

[9] Fengwei Chen, Juan C.Aguero, Marion Gilson, Hugues Garnier and Tao Liu. EM-based identification of continuous-time ARMA Models from irregularly sampled data. *Automatica*, vol 77, March 2017, pages: 293-301

[10] R. Pintelon and J. Schoukens, Identification of continuous-time systems with missing data, *IEEE Transactions on Instrumentation and Measurement*, vol. 48, no. 3, pp. 736-740, June 1999.

[11] Jeng-Ming Chen and Bor-Sen Chen.A.J. System Parameter Estimation with Input/Output Noisy Data and Missing Measurements. *IEEE TRANSACTIONS ON SIGNAL PROCESSING* , vol. 48, no. 6, JUNE 2000

[12] G. C. Goodwin and A. Feuer. Estimation with missing data. *Mathematical and Computer Modelling of Dynamical Systems*, vol. 5, no.3, pp. 220C244, 1998.

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

[14] R. H. Jones, Maximum likelihood fitting of ARMA models to times series with missing observations, *Technometrics*, vol. 22, pp. 389-395, Aug. 1980.

[15] Erik K. Larsson, Magnus Mossberg and Torsten Soderstrom. Identification of Continuous-Time ARX Models From Irregularly Sampled Data. *IEEE Transactions on Automatic Control*, vol: 52, no: 3, March 2007

[16] Raymond DeCarlo, Linear Systems: A State Variable Approach with Numerical Implementation, Prentice Hall, NJ, 1989

[17] Charles Van Loan, Computing integrals involving the matrix exponential,*IEEE Transactions on Automatic Control*. 23 (3): 395404, 1978

[18] T. Soderstrom, Wei Xing Zheng and P. Stoica, Comments on ”On a least-squares-based algorithm for identification of stochastic linear systems”, *IEEE Transactions on Signal Processing*, vol. 47, no. 5, pp. 1395-1396, May 1999.

[19] Ma Lili, Wang Huiran, Chen Jinguang, Analysis of Kalman Filter with Correlated Noises under Different Dependence. *Journal of Information and Computational Science* 7: 5 (2010) 11471154

[20] Papalambros, P., Wilde, D. (2000). Principles of Optimal Design: Modeling and Computation. Cambridge: Cambridge University Press.

[21] Rauch, H.E.; Tung, F.; Striebel, C. T. (August 1965). Maximum likelihood estimates of linear dynamic systems. AIAA Journal. 3 (8): 14451450.

[22] B. D. O. Anderson and J.B.Moore, Optimal Filtering. *Englewood Cliffs, NJ, Prentice-Hall*, 1979.

[23] Lei Chen, Lili Han, Biao Huang and Fei Liu. Parameter estimation for a dual-rate system with time delay. *ISA Transactions*, vol: 53, no: 5, September 2014, pages 1368-1376

[24] Ulf Soderstrom. Monetary Policy with Uncertain Parameters. March 2002

[25] L. Ljung, System Identification: Theoly for the User. *Englewood Cliffs*, NJ: Prentice-Hall, 1983

Appendix A

As mentioned in Section VII-D, we use EM algorithm to estimate a discrete-time ARX model first then do 'd2c' to obtain the parameters of continuous-time model. The EM algorithm is used to find the maximum likelihood parameters of a model in cases where the equations cannot be solved directly. The basis for the EM algorithm is two data sets X and Y, which represent complete data and incomplete data respectively. In this way, if X is observed, an observation of Y is available too, but not vice versa. We suppose that the likelihood function of the complete data X is $L(\theta|X)$ and the parameter $\hat{\theta}$ is obtained

$$\hat{\theta} = \arg\max_{\theta} \log L(\theta|X)$$

The EM algorithm seeks to find the estimate by iteratively applying two steps [4]. Firstly, we start with the 'E-step', calculating the expected value of the log likelihood function $Q(\theta|\theta^{(0)})$ with respect to incomplete data set Y given X under the current $\theta^{(0)}$. This Q is then used to obtain a new estimate $\theta^{(1)}$ in the so called 'M-step'. The new $\theta^{(1)}$ is then feedback to the 'E-step', iterating until convergence. In this thesis the estimate $\theta^{(n)}$ is said to have converaged if $\parallel \theta^{(n+1)} - \theta^{(n)} \parallel < 0.001$.

The difficulty of EM algorithm is at the E-step where we need to compute the conditional expectation $Q(\theta|\theta^{(n)})$ and this will be discussed in following part.

We considered a DARX model as (4) in the motivating example in Section II and a DAR model for the input u.

$$\begin{aligned} A^l(q)y(k) &= B^l(q)u(k) + v^l(k) \\ C^l((1)u(k) &= w^l(k), \end{aligned} \tag{39}$$

with

$$\begin{aligned} A^l(q) &= 1 - a_1^l q^{-1} - a_2^l q^{-2} - ... - a_n^l q^{-n}, \\ B^l(q) &= b_1^l q^{-1} + b_2^l q^{-2} + ... + b_m^l q^{-m}, \\ C^l(q) &= 1 - c_1^l q^{-1} - c_2^l q^{-2} - ... - a_n^l q^{-n}. \end{aligned} \tag{40}$$

where $v(k)$ and $w(k)$ are zero mean white noise sequences with $\mathbb{E}\left\{v^2(k)\right\} = \lambda_1$ and $\mathbb{E}\left\{w^2(k)\right\} = \lambda_2$ respectively. Define $\phi_1(k)$ and $\phi_2(k)$ as

$$\phi_1(k) = \begin{bmatrix} y(k-1) \\ u(k-1) \\ y(k-2) \\ u(k-2) \\ \vdots \\ y(k-n) \\ u(k-n) \end{bmatrix} = \begin{bmatrix} z(k-1) \\ z(k-2) \\ \vdots \\ z(k-n) \end{bmatrix} \qquad \theta_1 = \begin{bmatrix} a_1 \\ b_1 \\ \vdots \\ a_n \\ b_n \end{bmatrix}$$

$$\phi_2(k) = \begin{bmatrix} u(k-1) \\ u(k-2) \\ \vdots \\ u(k-n) \end{bmatrix} \qquad \theta_2 = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

With these new vectors, model (39) can be represented as

$$\begin{aligned} y(k) &= \phi_1^T(k)\theta_1 + v(k) \\ u(k) &= \phi_2^T(k)\theta_2 + w(k) \end{aligned}$$

To estimate θ_1 and θ_2, we first write the log-likelihood function based on Section III and [25]

$$\log L(\theta, \lambda_1, \lambda_2) = C - \frac{N}{2}\log\lambda_1 - \frac{N}{2}\log\lambda_2 - \frac{1}{2\lambda_1}\sum_{k=1}^{N}(y(k) - \phi_1^T(k)\theta_1)^2 - \frac{1}{2\lambda_2}\sum_{k=1}^{N}(u(k) - \phi_2^T(k)\theta_2)^2$$

Then we need to calculate $Q(\theta, \lambda_1, \lambda_2|\theta^{(n)}, \lambda_1^{(n)}, \lambda_2^{(n)})$. We denote the conditional expectation based on the all the observations $\mathbb{E}\{|Z^N\}$ by $\mathbb{E}_N$.

$$\begin{aligned} &Q(\theta, \lambda_1, \lambda_2|\theta^{(n)}, \lambda_1^{(n)}, \lambda_2^{(n)}) = \mathbb{E}_N\{\log L(\theta, \lambda_1, \lambda_2)\} \\ &= -\frac{N}{2}\log\lambda_1 - \frac{N}{2}\log\lambda_2 - \frac{1}{2\lambda_1}\sum_{k=1}^{N}\mathbb{E}_N(y(k) - \phi_1^T(k)\theta_1)^2 \\ &- \frac{1}{2\lambda_2}\sum_{k=1}^{N}\mathbb{E}_N(u(k) - \phi_2^T(k)\theta_2)^2 \end{aligned} \tag{41}$$

Setting derivatives of $Q(\theta, \lambda_1, \lambda_2|\theta^{(n)}, \lambda_1^{(n)}, \lambda_2^{(n)})$ with respect to λ_1 and λ_2 to zero

$$\frac{\partial Q}{\partial \lambda_1} = 0, \quad \frac{\partial Q}{\partial \lambda_2} = 0 \tag{42}$$

which lead to

$$\begin{aligned} \lambda_1 &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_N(y(k) - \phi_1^T(k)\theta_1)^2 \\ \lambda_2 &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_N(u(k) - \phi_2^T(k)\theta_2)^2 \end{aligned} \tag{43}$$

Then substitute λ_1 and λ_2 in (10) with (12) yields to

$$\begin{aligned} &Q(\theta, \lambda_1, \lambda_2|\theta^{(n)}, \lambda_1^{(n)}, \lambda_2^{(n)}) = \\ &C - \frac{2}{N}\log\underbrace{\left(\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_N(y(k) - \phi_1^T(k)\theta_1)^2\right)}_{V_1(\theta_1)} \\ &-\frac{N}{2}\log\underbrace{\left(\frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_N(u(k) - \phi_2^T(k)\theta_2)^2\right)}_{V_2(\theta_2)} \end{aligned} \tag{44}$$

Now we need to maximize $V_1(\theta_1)$ and $V_2(\theta_2)$, which becomes a least square problem.

$$\theta_1^{(n+1)} = \left(\sum_{k=1}^{N}\underbrace{\mathbb{E}_N\{\phi_1(k)\phi_1^T(k)\}}_{①}\right)^{-1}\sum_{k=1}^{N}\underbrace{\mathbb{E}_N\{\phi_1(k)y(k)\}}_{②}$$

$$\theta_2^{(n+1)} = \left(\sum_{k=1}^{N}\underbrace{\mathbb{E}_N\{\phi_2(k)\phi_2^T(k)\}}_{③}\right)^{-1}\sum_{k=1}^{N}\underbrace{\mathbb{E}_N\{\phi_2(k)y(k)\}}_{④}$$

and the update of the variance

$$\begin{aligned}\lambda_1^{(n+1)} &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_N(y(k)-\phi_1^T(k)\theta_1^{(n+1)})^2\\ \lambda_2^{(n+1)} &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}_N(u(k)-\phi_2^T(k)\theta_2^{(n+1)})^2\end{aligned} \tag{45}$$

$$\begin{aligned}\lambda_1^{(n+1)} = \frac{1}{N}\sum_{k=1}^{N}&\mathbb{E}_N(y^2(k)-2y(k)\phi_1^T(k)\theta_1^{(n+1)}\\ &+\theta_1^{(n+1)^T}\phi_1(k)\phi_1^T(k)\theta_1^{(n+1)})\\ \lambda_2^{(n+1)} = \frac{1}{N}\sum_{k=1}^{N}&\mathbb{E}_N(u^2(k)-2u(k)\phi_2^T(k)\theta_2^{(n+1)}\\ &+\theta_2^{(n+1)^T}\phi_2(k)\phi_2^T(k)\theta_2^{(n+1)})\end{aligned} \tag{46}$$

Therefore, to apply the EM algorithm we should calculate the expectations above. We have already obtained the filtered Kalman estimate $\widehat{x}(k|k)$ in the previous section. According to [4], a better reconstruction can be obtained using a fixed interval smoother with all observations. The optimal fixed-interval smoother provides $\widehat{x}(k|N)$ for $k < N$, and there are several smoothing algorithms in common use. According to [4], a so called Rauch-Tung-Striebel (RTS) smoother is considered, which is an efficient algorithm for fixed interval smoothing [21]. In order to correspond to the equations of Kalman filter in Section III, we choose same the notations based on model (18) The equations of the smoothing process are as follows [22] [4]:

$$\begin{aligned}\widehat{x}(k-1|N) = \widehat{x}&(k-1|k-1)\\ &+P(k-1|k-2)\overline{F}^T(k-1)S_1(k)\end{aligned} \tag{47}$$

$$\begin{aligned}P(k-1|N) = P(k-1|k-1) - P(k-1|k-2)\\ \cdot\overline{F}^T(k-1)S_2(k)\overline{F}(k-1)P(k-1|k-2)\end{aligned} \tag{48}$$

where

$$\begin{aligned}&\overline{F}(k) = F_2(k) - K(k)H_2(k)\\ &K(k) = (F_2P(k|k-1)H_2^T(k) + GR_{12}(k))\\ &\qquad\qquad\cdot[H_2(k)P(k|k-1)H_2^T(k) + R_2(k)]^{-1}\end{aligned}$$

Besides, S_1 and S_2 in the smoothing equations are obtained by backwards recursively updating by taking $j = N, N-1, N-2, \cdots, 1$.

$$\begin{aligned}S_1(j) = &\overline{F}^T(j)S_1(j+1)\\ &+H_2^T(j)[H_2(j)P(j|j-1)H_2^T(j) + R_2(j)]^{-1}\epsilon(j)\\ S_2(j) = &\overline{F}^T(j)S_2(j+1)\overline{F}(j) + H_2^T(j)\\ &\cdot(H_2(j)P(j|j-1)H_2^T(j) + R_2(j))^{-1}H_2(j)\end{aligned}$$

The initial condition should be considered as $S_1(N+1) = 0$ and $S_1(N+1) = 0$, which leads to $\widehat{x}(N|N) = \widehat{x}(N|N)$ in (47) and $P(N|N) = P(N|N)$ in (48).

To evaluate ① and ③ we use the state-space formulation to perform the expectation.

$$\mathbb{E}_N\{\phi_1(k)\} = \mathbb{E}_N\{x(k)(1:(n-2))\} = \hat{x}(k|N)(1:(n-2))$$

$$\mathbb{E}_N\{\phi_2(k)\} = \mathbb{E}_N\{x(k)(2:2:2n{-}2)\} = \hat{x}(k|N)(2:2:2n{-}2)$$

$1:(n-2)$ means the first row to the $(n-2)$th row. According to [4], the standard formula

$$\mathbb{E}(x-\mathbb{E}\,x)(x-\mathbb{E}\,x)^T = \mathbb{E}\,xx^T - \mathbb{E}\,x\,\mathbb{E}\,x^T$$

We know that the definition of covariance matrix $P(k|N)$ is

$$P(k|N) = \mathbb{E}_N\{(x(k)-\hat{x}(k|N))(x(k)-\hat{x}(k|N))^T\}$$

Hence, ①

$$\mathbb{E}_N\{\phi_1(k)\phi_1^T(k)\} = \hat{x}(k|N)\hat{x}(k|N)^T(1:(n-2)) + P_1(k|N)$$

where $P_1(k|N)$ is $P(k|N)((2n-1)\times(2n-1))$. Similarly, $P_2(k|N) = P(k|N)(2:2:2n-2)\times(2:2:2n-2)$
③

$$\mathbb{E}_N\{\phi_2(k)\phi_2^T(k)\} = \hat{x}(k|N)\hat{x}(k|N)^T[2,4,...,2n{-}2] + P_2(k|N)$$

Moreover, we need to compute ② and ④ which depend on $y(k)$ and $u(k)$. $\phi_1(k)$ can be found in the one step ahead vector $x(k+1)$ as element 3 to $2n+2$, and $y(k)$ as first element of $x(k+1)$. The corresponding covariance matrix is donated as $P_2(k+1|N)$. $\phi_2(k)$ can be found in the one step ahead vector $x(k+1)$ as elements $(4:2:2n)$, and $u(k)$ as second element of $x(k+1)$. The corresponding covariance matrix is donated as $P_4(k+1|N)$.