Dual time-frequency domain system identification

Juan C. Agüero [a] , Wei Tang [b] , Juan I. Yuz [c] , Ramón Delgado [a]
Graham C. Goodwin [a]

[a] *School of Electrical Engineering and Computer Science, The University of Newcastle, Australia*

[b] *School of Automation, Northwestern Polytechnical University, China*

[c] *Electronic Engineering Department, Universidad Técnica Federico Santa María, Chile*

Abstract

In this paper we obtain the maximum likelihood estimate of the parameters of discrete-time linear models by using a dual time-frequency domain approach. We propose a formulation that considers a (reduced-rank) linear transformation of the available data. Such a transformation may correspond to different options: selection of time-domain data, transformation to the frequency domain, or selection of frequency-domain data obtained from time-domain samples. We use the proposed approach to identify multivariate systems represented in state-space form by using the Expectation-Maximization algorithm. We illustrate the benefits of the approach via numerical examples.

Key words: Robust System Identification, Maximum Likelihood

1 Introduction

Identification of linear systems has been a topic of recurrent interest in the areas of time series analysis [24,26,8,35] and system identification [20,34,48]. Most of the proposed algorithms available in the current literature utilize data in the time domain. However, interest in developing methods that utilize data in the frequency domain has increased in the last decade (see [34,41,36,3] and the references therein). An important feature of frequency domain identification methods is that the model can be fitted using the data in a reduced frequency range. This has been used as a tool to *robustify* the algorithms to modelling errors and to fit *low-complexity* models for at least four decades (see e.g. [27,16,41,36,56,17,3] and the references therein).

Estimation algorithms that are able to deal with missing data have also been a topic of recurrent interest in the areas of Statistics, Time Series Analysis, Finance and Engineering (see e.g. [5,1,43,47,32,49,28,33,21,6]). There has been renewed interest in developing identification algorithms that consider missing measurements, for example, in networked control systems subject to data packet dropouts (see e.g. [44]).

In the current paper, we consider the problem of identifying a discrete-time linear model using frequency domain data in a reduced bandwidth, when we have missing data in the time domain. A similar problem has been addressed before in [40]. However, in that work missing measurements were considered as extra parameters to be estimated. Here, we consider the missing data as random variables.

The notation to be used in the remainder of the paper is defined below:

1.1 Notation

A^T denotes the transpose of the matrix A, $A^\dagger$ denotes the pseudoinverse (in the sense of Penrose) of the matrix A, $\operatorname{tr}\{A\}$ denotes the trace of A, $\det\{A\}$ denotes the determinant of A, $\operatorname{vec}\{A\}$ denotes the operator

[⋆] This paper was not presented at any IFAC meetings. Corresponding author: Juan C. Agüero. Tel. +61-2-49216151. Fax +61-2-49601712. Work partially supported by ARC, Australia, National Natural Science Foundation of China (No. 50905140), Research Fund for the Doctoral Program of Higher Education of China (No.20096102120039) and NPU "Aoxiang Star" Program, and CONICYT-Chile (Project ACT 53 and FONDECYT 1110586).

Email addresses: juan.aguero@newcastle.edu.au (Juan C. Agüero), tangwei@nwpu.edu.cn (Wei Tang), juan.yuz@usm.cl (Juan I. Yuz), ramon.delgado@uon.edu.au (Ramón Delgado), graham.goodwin@newcastle.edu.au (Graham C. Goodwin).

that transforms a matrix into a vector by stacking its columns , $A \otimes B$ denotes Kronecker product, $0_{a \times b}$ and I_a denote the zero matrix of dimensions $a \times b$ and the identity matrix of dimensions $a \times a$ respectively. For the sake of conciseness, we will use the following column vector to represent a sub-sequence of the signal x_t, $\vec{x}_{k:l} = \left[x_k^T \; x_{k+1}^T \; \dots \; x_l^T\right]^T$. The particular case $\vec{x}_{k:k}$ denotes the singleton x_k. We use $\vec{y}^{(o)}$ and $\vec{y}^{(m)}$ to denote the collection of observed and missing data over the time interval 0 to $N-1$, respectively. β_0 denotes the true parameter, and $\hat{\beta}$ denotes an estimate of β_0. $p_{\vec{x}}(\vec{\alpha})$ denotes the probability density function (pdf) of the random vector $\vec{x}$ evaluated at $\vec{\alpha}$. Whenever it is clear from the context, we will use $p(\vec{x})$ to denote the pdf of the random vector $\vec{x}$ evaluated at $\vec{x}$. $p(\vec{x}|\vec{y})$ denotes the conditional pdf of $\vec{x}$ for a given $\vec{y}$. $\mathbb{E}\{\cdot\}$ denotes the expectation operator. $\mathbb{E}\{x|y\}$ denotes the expected value of the random variable x given the random variable y. A selection matrix is denoted as $S_x \in \mathbb{R}^{n \times m}$ and contains, only one non-zero element, i.e. 1, in each row. Depending on the context z denotes either the forward shift operator, $zy_t = y_{t+1}$, or the $\mathcal{Z}$ transform variable.

1.2 The problem of interest

Consider a discrete-time linear system of the form:

$$y_t = G_o(z)u_t + H_o(z)\eta_t \tag{1}$$

where, $u_t \in \mathbb{R}^{n_u}$, $y_t \in \mathbb{R}^{n_y}$ are the input, and measured output signal respectively. The random processes $\eta_t \in \mathbb{R}^{n_\eta}$ is a discrete-time zero mean Gaussian white noise sequence with covariance matrix P.

We assume that the input, u_t, is a known deterministic signal and that some output measurements are missing. We define the following (stochastic) signal, which defines the instants where the data is missing:

$$s_t = \begin{cases} 1 & \text{if } y_t \text{ is measured,} \\ 0 & \text{if } y_t \text{ is missing.} \end{cases} \tag{2}$$

The observed and missing data are denoted by $\vec{y}^{(o)}$, and $\vec{y}^{(m)}$ respectively.

We combine the parameters to be estimated in a vector denoted by β. This vector of parameters includes the parameters that define models for $G_o(z)$ and $H_o(z)$, the statistical properties of the system initial condition, and the noise covariance matrix P. We assume that there exists a *true* value, β_o, for the parameter vector.

In order to translate the data to the frequency domain, we use the Discrete Fourier Transform (DFT) given by:

$$Y_k = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} y_t e^{-j\frac{2\pi}{N}kt} \quad ; k = 0, \dots, N-1$$

We use capital letters, e.g. Y_k, to represent the Fourier transform of the signal y_t. We utilize the following real transformation to represent the frequency domain data (see e.g. [3] for details):

$$\vec{Y}^R_{0:N-1} = M_R \vec{y}_{0:N-1} \tag{3}$$

where

$$\vec{Y}^R_{0:N-1} = \begin{cases} \text{If } N \text{ is even:} \\ [Y_0^T, \sqrt{2}\Re\{Y_1^T\}, \sqrt{2}\Im\{Y_1^T\}, \dots, Y_K^T]^T \\ \\ \text{If } N \text{ is odd:} \\ [Y_0^T, \sqrt{2}\Re\{Y_1^T\}, \sqrt{2}\Im\{Y_1^T\}, \dots, \sqrt{2}\Re\{Y_K^T\}, \sqrt{2}\Im\{Y_K^T\}]^T \end{cases}$$

where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ represent "the real and imaginary part of" respectively. Note that M_R in (3) is a square, full-rank, real, unitary matrix, and $K = \lfloor N/2 \rfloor$ is the largest integer less than or equal to $N/2$.

We are interested in developing a Maximum-Likelihood-based algorithm to estimate β_o using the available data. The available data considers the effect of missing data in the time domain, and the use of frequency domain data in a reduced frequency range (i.e. by selecting some components of $\vec{Y}^R_{0:N-1}$).

The layout of the remainder of the paper is as follows. In Section 2 we discuss relevant issues regarding ML estimation in both time and frequency domains. In Section 3, a dual time-frequency domain algorithm is presented which allows one to deal with incomplete data. In Section 4, an EM-based algorithm is developed in order to identify dynamic systems using a dual time-frequency domain approach. In Section 5, we present numerical examples. Finally, in Section 6, we draw conclusions.

2 Maximum Likelihood estimation in time and frequency domains

2.1 Maximum Likelihood estimation in the time domain

The problem of Maximum Likelihood (ML) estimation in the time domain is well documented and appears in many textbooks (see e.g. [20,48,34]). Indeed, for the system described in (1), the ML estimate obtained from the data $\vec{y}_{0:N-1}$ is given by the solution of the following optimisation problem:

$$\hat{\beta} = \arg\min_{\beta} l(\beta)$$

where the negative-log-likelihood function, $l(\beta)$, is given by, save for some constants, (see e.g. [10, page 140]):

$$l(\beta) = \sum_{t=0}^{N-1} (y_t - \hat{y}_{t|t-1})^T \Sigma_{t|t-1}^{-1} (y_t - \hat{y}_{t|t-1}) + \log\det\left\{\Sigma_{t|t-1}\right\}$$

where

$$\hat{y}_{t|t-1}(\beta) = \mathbb{E}\left\{y_t|\vec{y}_{0:t-1}, \beta\right\} \quad (4)$$

$$\Sigma_{t|t-1}(\beta) = \mathbb{E}\left\{(y_t - \hat{y}_{t|t-1})(y_t - \hat{y}_{t|t-1})^T|\vec{y}_{0:t-1}, \beta\right\} \quad (5)$$

and where y_0 is assumed to be normally distributed with mean $\hat{y}_{0|-1}(\beta)$ and variance $\Sigma_{0|-1}(\beta)$. The quantities in (4) and (5) can be obtained by representing the linear system in state-space form and using Kalman filtering techniques [1] . In order to reduce the computational load associated with the calculation of the gradient and Hessian of $l(\beta)$, iterative algorithms have also been developed, (see e.g. [9,22,46] and [10, chapter 4]).

We next present a result that shows the effect of employing a linear transformation of the data on the ML estimate.

Lemma 1 *The ML estimate for a vector of parameters β_0 obtained from the data $\vec{y}_{0:N-1}$ is invariant under any bijective linear transformation of the data, $g = L\vec{y}_{0:N-1}$, where $L \in \mathbb{R}^{m\times Nn_y}$, i.e.*

$$\hat{\beta} = \arg\max_{\beta} p(\vec{y}_{0:N-1}|\beta) = \arg\max_{\beta} p_g(L\vec{y}_{0:N-1}|\beta) \quad (6)$$

On the other hand, the ML estimate obtained from the data $L\vec{y}_{0:N-1}$ when rank $\{L\} = l \leq \min\{m, Nn_y\}$ is given by:

$$\hat{\beta} = \arg\max_{\beta} p_{g_r}(V_1^T\vec{y}_{0:N-1}|\beta) \quad (7)$$

where $g_r = V_1^T\vec{y}_{0:N-1}$ and the singular value decomposition of L is given by:

$$L = U\begin{bmatrix} S & 0_{l\times(Nn_y-l)} \\ 0_{(m-l)\times l} & 0_{(m-l)\times(Nn_y-l)}\end{bmatrix}\begin{bmatrix} V_1^T \\ V_2^T\end{bmatrix} \quad (8)$$

$U \in \mathbb{R}^{m\times m}$ and $V = \begin{bmatrix} V_1 & V_2\end{bmatrix}$ are unitary matrices where $V_1 \in \mathbb{R}^{Nn_y\times l}$ and $V_2 \in \mathbb{R}^{Nn_y\times(Nn_y-l)}$ and S is a diagonal matrix with elements given by the non-zero singular values of L, i.e. $S = \text{diag}\{\sigma_1(L), \sigma_2(L), \ldots, \sigma_l(L)\}$.

PROOF. *We analyse two cases:*

(i) L is a square invertible matrix: (6) *is immediately obtained by using the theorem of transformation of random variables (see e.g. [23, page 21]).*

(ii) rank $\{L\} = l \leq \min\{m, Nn_y\}$: *considering that* (8) *is the singular value decomposition of the matrix L (see e.g. [45]), and using the theorem of transformation of random variables we have that the ML estimate obtained from the data $L\vec{y}_{0:N-1}$ is the same as the ML estimate obtained from the data $\begin{bmatrix}(SV_1^T\vec{y})^T & 0^T_{(m-l)\times Nn_y}\end{bmatrix}^T$. The pdf of this latter vector is singular, and the ML estimate is given by the maximisation of the envelope of the singular pdf. Finally, considering that S is an invertible matrix, and using the theorem of transformation of random variables, we obtain the result in* (7). □□□

2.2 ML estimation in the frequency domain

The ML estimate in the frequency domain is obtained by solving the following optimisation problem:

$$\hat{\beta} = \arg\max_{\beta} p(\vec{Y}^R_{0:N-1}|\beta)$$

Notice that $\vec{Y}^R_{0:N-1}$ contains all the information in the data since there exists a bijective linear transformation from $\vec{y}_{0:N-1}$ to $\vec{Y}^R_{0:N-1}$. Using Lemma 1, we have that the ML estimates obtained in time and frequency domains are equivalent even for finite data length (see [3] for a discussion comparing the estimates obtained by the "usual" ML in the frequency domain with those obtained in the time domain).

In order to fit models using the frequency domain data in a reduced frequency range, the following likelihood is typically used:

$$\hat{\beta} = \arg\max_{\beta} p(g|\beta) \quad (9)$$

$$g = S_F\vec{Y}^R_{0:N-1} \quad (10)$$

and S_F is a (deterministic) full row rank frequency domain selection matrix.

3 ML estimation with incomplete data in time and frequency domain

In practice, it is common that only a *coarse* version of the data is available for estimation purposes. This could be due to a failure in the measurement device, quantisation of the measurements, or due to the fact that outliers have been discarded. As a consequence, a considerable amount of research on estimation with incomplete data has been developed in the last four decades. The main focus in this area has been on missing data in the time domain (see e.g. [13,33,29]). To the best of our knowledge, [40] is the only reference that deals with data in the frequency domain.

The estimation algorithm in (9)-(10) is a particular case of Maximum Likelihood for *data grouping* [29]. The concept of data grouping makes reference to two sample spaces $\mathscr{C}$ and $\mathscr{G}$ and a many-to-one mapping from $\mathscr{C}$

[1] Note that the expressions for $\hat{y}_{t|t-1}(\beta)$ and $\Sigma_{t|t-1}(\beta)$ are simpler when the model for $H_o(z)$ is invertible, stable and minimum phase.

to $\mathscr{G}$. The data used for estimation, g, is a realisation from $\mathscr{G}$. The corresponding data C in $\mathscr{C}$ is not directly used for estimation, but is used to define the likelihood function of interest. It is assumed that there is a (deterministic) coarsening mapping $C \to g(C)$ from $\mathscr{C}$ to $\mathscr{G}$. In this framework, C is called the complete data. The corresponding likelihood function is given by [29]:

$$L_G(\beta) = p(g|\beta) = \int_{C(g)} p(C|\beta)dC \tag{11}$$

where $C(g)$ is the set given by all the values of C that are consistent with the given data g, i.e.

$$C(g) = \{C \in \mathscr{C} | g = g(C)\}$$

For the case of maximum likelihood using a reduced frequency range, the complete data is given by $C = Y^R_{0:N-1}$ and the many-to-one mapping by $g = S_F \vec{Y}^R_{0:N-1}$.

On the other hand, in most estimation problems the coarsening mapping from $\mathscr{C}$ to $\mathscr{G}$ is, in general, a stochastic process. In this case, the likelihood of interest is given by [29]:

$$L_C(\beta, \psi) = \int_{C(g)} p(C|\beta)p(g|C,\psi)dC \tag{12}$$

where the vector ψ is used to parameterise the stochastic law for g given C. In particular, when the coarsening data mechanism is due to missing data, we have that the complete data is given by $C = \vec{y}_{0:N-1}$, and the data used for estimation is given by $g = \vec{y}^{(o)} = S_T \vec{y}_{0:N-1}$ where S_T is the selection matrix that defines the observed data $\vec{y}^{(o)}$. In this case, we have that the probability of $\vec{y}^{(o)}$ given $\vec{y}_{0:N-1}$ depends on the stochastic law for s_t in (2).

It is well known that the estimates provided by (11) and (12) are, in general, different [29]. In addition, it is more difficult to develop an algorithm to optimise (12) than to develop an algorithm to optimise (11). A special case is when both approaches provide the same estimates. This situation occurs if the stochastic law of the coarsening data mechanism satisfies an assumption, known as *Coarsening At Random* (CAR). A missing data mechanism is CAR if, for the fixed observed value of g and for each value of ψ, $p(g|C,\psi)$ takes the same value for all $C \in C(g)$.

A particular case when the coarsening data mechanism is CAR is when the coarsening is due to missing data in the time domain and $\vec{s}_{0:N-1}$ satisfies $p(\vec{s}_{0:N-1}|\vec{y}^{(o)}, \vec{y}^{(m)}, \psi) = p(\vec{s}_{0:N-1}|\psi)$. This latter condition is known in the literature as *missing completely at random* [43]. For example, if the data generating mechanism ($\vec{y}_{0:N-1}$) is independent of the missing data mechanism ($\vec{s}_{0:N-1}$), and $\vec{s}_{0:N-1}$ is obtained by the following Markov model:

$$p(s_t = 0|s_{t-1} = 0) = 1 - p(s_t = 1|s_{t-1} = 0) = \varsigma \tag{13}$$
$$p(s_t = 1|s_{t-1} = 1) = 1 - p(s_t = 0|s_{t-1} = 1) = \delta \tag{14}$$

where $0 \le \varsigma \le 1$, $0 \le \delta \le 1$ and the vector of parameters ψ is given by $\psi = \begin{pmatrix} \varsigma & \delta \end{pmatrix}^T$, then the coarsening data mechanism is CAR.

Within the scope of the current paper it suffices to understand that, under the assumption that the coarsening data mechanism is CAR, the maximum likelihood estimate is given by the solution of the optimisation problem in (11). We refer the reader to existing literature regarding the theory of estimation with coarse and missing data for further details (see e.g. [43,29,33]).

3.1 Incomplete data in time and frequency domain

For the problem of interest in this paper, the coarsening data mechanism is due to both missing data in time domain, and the choice of using the frequency domain data in a reduced frequency range. Since the selection of the reduced frequency range is deterministic (determined by the user and not based on the data) the only randomness in the coarsening procedure is due to the missing data mechanism in the time domain.

Obtaining models that fit the data over some specific frequency bandwidth is often motivated as a mechanism to obtain *robust* estimates and also because, in general, one is interested in obtaining a low-complexity model (see e.g. [17]). This can be treated by using a selection of the data in the frequency domain. This can be expressed as follows:

$$\vec{y}^{(r)} = S_F \vec{Y}^R_{0:N-1} \tag{15}$$

where $S_F \in \mathbb{R}^{m \times Nn_y}$ is a selection matrix. This selection matrix can be constructed in a similar fashion to the selection matrix S_T for missing data in the time domain. Combining equations (3) and (15) we have that the data to be used in our identification procedure is given by:

$$\vec{y}^{(r)} = S_F \vec{Y}^R_{0:N-1} = S_F M_R \vec{y}_{0:N-1} \tag{16}$$

Similarly, we can also build a matrix $S_F^\perp \in \mathbb{R}^{(Nn_y - m) \times Nn_y}$ that selects the data in the frequency outside the range of interest.

$$\vec{y}^{(nr)} = S_F^\perp \vec{Y}^R_{0:N-1} = S_F^\perp M_R \vec{y}_{0:N-1} \tag{17}$$

Combining (16) and (17) we have that:

$$\begin{bmatrix} y^{(r)} \\ y^{(nr)} \end{bmatrix} = \begin{bmatrix} S_F \\ S_F^\perp \end{bmatrix} M_R \vec{y}_{0:N-1}$$

where S_F and $S_F^\perp$ are selection matrices. Similarly, the observed and missing data in time domain satisfy the following:

$$\begin{bmatrix} y^{(o)} \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} S_T \\ S_T^\perp \end{bmatrix} \vec{y}_{0:N-1}$$

where the matrices $S_T \in \mathbb{R}^{p \times Nn_y}$ and $S_T^\perp \in \mathbb{R}^{(Nn_y - p) \times Nn_y}$ are selection matrices. Combining both equations we have that:

$$\begin{bmatrix} \vec{y}^{(r)} \\ \vec{y}^{(nr)} \end{bmatrix} = \begin{bmatrix} S_F \\ S_F^\perp \end{bmatrix} M_R \begin{bmatrix} S_T \\ S_T^\perp \end{bmatrix}^{-1} \begin{bmatrix} \vec{y}^{(o)} \\ \vec{y}^{(m)} \end{bmatrix}$$

Considering that $\begin{bmatrix} S_T^T & (S_T^\perp)^T \end{bmatrix}^T$ is a permutation matrix, we have that [45, page 330] $\begin{bmatrix} S_T \\ S_T^\perp \end{bmatrix}^{-1} = \begin{bmatrix} (S_T)^T & (S_T^\perp)^T \end{bmatrix}$. Thus, we have that the data in the frequency range of interest can be described by $\vec{y}^{(r)} = S_F M_R S_T^T \vec{y}^{(o)} + S_F M_R (S_T^\perp)^T \vec{y}^{(m)}$. Notice that only part of the data of interest is available, i.e. the data that, in this case, can be used for inference is given by $\vec{y}^{(x)} = S_F M_R S_T^T \vec{y}^{(o)}$. Since the rank of the matrix $S_F M_R S_T^T \in \mathbb{R}^{m \times p}$ is less or equal to $\min\{m, p, Nn_y\}$, we have that this matrix will not be, in general, full row rank. Then, using Lemma 1 (with $L = S_F M_R S_T^T$) we have that the data to be used for estimation can be defined as $g = V_1^T \vec{y}^{(o)}$, where V_1 is obtained from a singular value decomposition of $S_F M_R S_T^T$.

In order to calculate the likelihood function we use the relationship between $\vec{y}^{(o)}$ and $\vec{y}_{0:N-1}$, i.e. $g = V_1^T S_T \vec{y}_{0:N-1}$ and consider that $V_1^T S_T$ is a full-row rank matrix (see [45, page 38]). Notice that, if the coarsening data mechanism is CAR, then the likelihood function of interest is then given by (11) and the set $C(g)$ is given by:

$$C(g) = \{C \in \mathscr{C} | g = V_1^T \vec{y}^{(o)} = V_1^T S_T \vec{y}_{0:N-1}\}$$

Notice that, since S_T is full row rank, it is sufficient to consider $g = S_T^T \vec{y}^{(o)}$ and there is no need to obtain its singular value decomposition.

The complete data C and the set $\mathscr{C}$ might contain the output $\vec{y}_{0:N-1}$, but will, more generally, be defined depending on the estimation problem of interest.

In the case that the coarsening data mechanism is given by missing data in the time domain, the set $C(g)$ is given by:

$$\begin{aligned} C(g) &= \{C \in \mathscr{C} | V_1^T \vec{y}^{(o)} = V_1^T S_T \vec{y}_{0:N-1}\} \\ &= \{C \in \mathscr{C} | g = \vec{y}^{(o)} = S_T \vec{y}_{0:N-1}\} \end{aligned}$$

since V_1^T is a full row rank matrix (or S_T a full row rank selection matrix). The set $\mathscr{C}$ can be defined, for example, by the union of missing and observed data. This agrees with the analysis typically used for missing data in the time domain.

Remark 2 *Note that the random variable $\vec{y}^{(x)} = S_T^T \vec{y}^{(o)}$ is the same as would be obtained by replacing the missing data with zero entries in the vector $\vec{y}_{0:N-1}$. The likelihood function is, however, calculated by using the random variable $\vec{y}^{(x)} = S_T^T S_T \vec{y}_{0:N-1}$ which has a singular distribution. This likelihood function is, in general, different to the one obtained from the random variable $\vec{y}_{0:N-1}$ evaluated at the observed data with the missing data replaced by zero.* ▽▽▽

On the other hand, if the coarsening data mechanism is due to a selection of frequency domain data, then the set $C(g)$ is given by:

$$C(g) = \{C \in \mathscr{C} | g = S_F Y_{0:N-1}^R = S_F M_R \vec{y}_{0:N-1}\}$$

which agrees with the analysis typically used for maximum likelihood in the frequency domain.

3.2 The Expectation-Maximization algorithm

The Expectation-Maximisation (EM) algorithm is one of the main tools used to identify dynamic systems in the presence of missing data in the time domain (see e.g. [47,49,30,18,21]).

In the EM algorithm there is an underlying assumption that the missing data mechanism is somehow "independent" of the coarsening process [13]. The usual assumption for missing data in the time domain is that, in the missing data mechanism, $\vec{s}_{0:N-1}$ depends only on $\vec{y}^{(o)}$ and not on $\vec{y}^{(m)}$.

The EM algorithm finds the estimate for a vector of parameters β_o by maximising the likelihood function given in (11). This means that the EM algorithm can be utilised for general problems with incomplete data.

The EM algorithm may be summarised as follows [13]:

(1) Choose an initial estimate $\hat{\beta}_0 \in \Pi$, where Π is a constraint set in the parameter space. $i = 0$.

(2) Compute the auxiliary function $\mathbf{Q}(\beta, \hat{\beta}_i)$ which is the expected value of the complete data log-likelihood given the observed data g and the estimate $\hat{\beta}_i$, i.e. $\mathbf{Q}(\beta, \hat{\beta}_i) = \mathbb{E}\left\{\log[p(C|\beta)] \,|g, \hat{\beta}_i\right\}$.

(3) Set $\hat{\beta}_{i+1} = \arg\max_{\beta \in \Pi} \mathbf{Q}(\beta, \hat{\beta}_i)$.

(4) $i = i+1$. Go to step 2, and repeat until convergence.

Steps 2 and 3 are usually known as the E-step and M-step, respectively. Under quite general conditions

[13,55,50], the EM algorithm can be proven to converge to a stationary point of the likelihood function. In many practical applications this will be a local maximum of the likelihood function [38].

4 Dual time-frequency identification of linear state-space models via the EM algorithm

Identification of multivariate systems has received considerable attention in the last four decades in the area of time series analysis [24,42,35,8] and system identification [20,48,34].

Multivariate systems can be identified by using different estimation methods such as Maximum Likelihood (ML), the Prediction Error Method (PEM), and Subspace identification methods. All of these methods can be formulated in both time- and frequency- domains (see e.g. [20,34,12,51,41,7,2]).

Systems represented in state-space form are used in many areas including automatic control [19] and econometrics [25]. State-space models provide a concise and flexible representation of a multivariate system [14]. We use the following state-space formulation:

$$x_{t+1} = Ax_t + Bu_t + w_t \tag{18}$$

$$y_t = Cx_t + Du_t + v_t \tag{19}$$

where $x_t \in \mathbb{R}^n$, $u_t \in \mathbb{R}^{n_u}$, $y_t \in \mathbb{R}^{n_y}$ are the state, the input, and measured output signal respectively. The random processes $\eta_t = \begin{bmatrix} w_t^T & v_t^T \end{bmatrix}^T$ is a discrete-time zero mean Gaussian white noise sequence with covariance matrix given by:

$$\mathbb{E}\left\{\eta_\ell \, {\eta_k}^T\right\} = P\delta_K[\ell - k], \quad P = \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \tag{20}$$

where $\delta_K[\ell]$ is the Kronecker delta function, and x_0 is a Gaussian random variable with mean μ_0 and covariance matrix P_0 i.e. $x_0 \sim \mathrm{N}(\mu_0, P_0)$. Furthermore, we assume that x_0, and η_t are jointly independent and that the system is operating in open loop, i.e. the input signal u_t and the noise sequence η_t are jointly independent. Note that when the "real" system does not satisfy the assumptions made to find the maximum likelihood estimate, the resultant estimates correspond to estimates based on Quasi-maximum likelihood [53]. In fact, the analysis when the noise η_t is non Gaussian distributed is well known (see [26] for further details).

We are interested in obtaining ML estimates of the state-space matrices $\{A, B, C, D\}$ in (18), (19), of P in (20), and μ_0 and P_0, given the measurements of the input and output data sets $\{u_t\}$ and $\{y_t\}$ of length N, *i.e.*, $t = 0, \ldots, N-1$. Note that, the system is not parameter identifiable since a linear transformation of the states will define an equivalent state-space representation for the system. On the other hand, the main difficulty in identifying a system expressed in state-space form is that the likelihood function is, in general, non-convex. Thus, the success of the optimisation procedure utilised to maximise the likelihood function strongly depends on which parameterisation is used. In [37] it is suggested that one should use over-parameterised state-space models in order to overcome the numerical issues that arise when using canonical parameterisations. These over-parameterised models have also been successfully used in the context of EM-based algorithms (see e.g. [47,39,18,15,2]). In this paper we also use an over-parameterised state-space model. The extension to non-over-parameterised state-space model is straightforward.

ML estimation in the presence of time domain missing data has previously been addressed by using the Expectation-Maximization (EM) algorithm (see e.g. [30,21]).

4.1 Calculation of the intermediate function $Q(\beta, \hat{\beta}_i)$

We first calculate the pdf of the complete data given by $C = \{\vec{x}_{0:N-1}, \vec{y}_{0:N-1}\}$:

Lemma 3 *The joint pdf of $\vec{x}_{0:N}$ and $\vec{y}_{0:N-1}$ is given by*

$$p(\vec{x}_{0:N}, \vec{y}_{0:N-1}) = N(\mu, \Sigma) \tag{21}$$

where $N(\mu, \Sigma)$ denotes the pdf of a normal distribution with mean μ and covariance matrix Σ where:

$$\mu = \Gamma\mu_0 + \Lambda\vec{u}_{0:N-1}$$
$$\Sigma = \Gamma P_0 \Gamma^T + \Omega\mathcal{P}[I_N \otimes P]\mathcal{P}^T\Omega^T$$

where $\mathcal{P} \in \mathbb{R}^{N(n+n_y)\times N(n+n_y)}$ is a permutation matrix that transforms $\vec{\eta}_{0:N-1}$ into the stacking of $\vec{w}_{0:N-1}$ and $\vec{v}_{0:N-1}$, and $\Gamma \in \mathbb{R}^{(n+N(n_y+n))\times n}$, $\Lambda \in \mathbb{R}^{(n+N(n_y+n))\times Nn_u}$ and $\Omega \in \mathbb{R}^{(n+N(n_y+n))\times N(n+n_y)}$ are given by:

$$\Gamma = \left[\begin{array}{c} I_n \\ A \\ A^2 \\ \vdots \\ A^N \\ \hdashline C \\ CA \\ \vdots \\ CA^{N-1} \end{array}\right] \tag{22}$$

$$\Lambda = \left[\begin{array}{ccccc} \multicolumn{5}{c}{0_{n\times Nn_u}} \\ B & 0 & \cdots & \cdots & 0 \\ AB & B & 0 & \cdots & 0 \\ A^2B & AB & B & 0 & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ A^{N-1}B & A^{N-2}B & \cdots & AB & B \\ \hline D & 0 & \cdots & 0 & 0 \\ CB & D & 0 & \vdots & \vdots \\ CAB & CB & D & 0 & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ CA^{N-2}B & CA^{N-3}B & \cdots & CB & D \end{array}\right]$$

$$\Omega = \left[\begin{array}{ccccc|c} \multicolumn{5}{c|}{0_{n\times Nn}} & \\ I_n & \cdots & \cdots & \cdots & 0 & \\ A & I_n & 0 & \cdots & \cdots & \\ \vdots & \ddots & \ddots & \ddots & \vdots & 0_{(Nn+n)\times Nn_y} \\ A^{N-1} & \cdots & \cdots & A & I_n & \\ \hline \multicolumn{5}{c|}{0_{n_y\times Nn}} & \\ C & 0 & \cdots & \cdots & 0 & \\ CA & C & 0 & \cdots & 0 & \\ CA^2 & CA & C & \cdots & 0 & I_{Nn_y} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ CA^{N-2} & \cdots & \cdots & C & 0 & \end{array}\right] \quad (23)$$

Moreover, the logarithm of the joint pdf of $\vec{x}_{0:N}$ *and* $\vec{y}_{0:N-1}$ *is given by:*

$$\begin{aligned} l^c(\beta) &= \log p(\vec{x}_{0:N}, \vec{y}_{0:N-1}) \\ &= -\frac{1}{2}\log\det\{P_0\} - \frac{N}{2}\log\det\{P\} \\ &-\frac{1}{2}(x_0-\mu_0)^T P_0^{-1}(x_0-\mu_0) \\ &-\frac{1}{2}\sum_{t=0}^{N-1}(\zeta_t - \Theta\xi_t)^T P^{-1}(\zeta_t - \Theta\xi_t) \\ &-[N(n+n_y)+n]\log 2\pi \end{aligned} \quad (24)$$

where

$$\Theta = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \quad \zeta_t = \begin{bmatrix} x_{t+1} \\ y_t \end{bmatrix}, \quad \xi_t = \begin{bmatrix} x_t \\ u_t \end{bmatrix}$$

PROOF. *Using* (1) *and* (19)*, we have that*

$$\begin{bmatrix} \vec{x}_{0:N} \\ \vec{y}_{0:N-1} \end{bmatrix} = \Gamma x_0 + \Lambda \vec{u}_{0:N-1} + \Omega \begin{bmatrix} w_{0:N-1} \\ v_{0:N-1} \end{bmatrix}$$

where Γ, Λ *and* Ω *are given in* (22)-(23)*. Then, using a permutation matrix* $\mathcal{P}$ *we have that:*

$$\begin{bmatrix} w_{0:N-1} \\ v_{0:N-1} \end{bmatrix} = \mathcal{P}\vec{\eta}_{0:N-1}$$

Finally, on noting that η_t *is an i.i.d. Gaussian white noise sequence, the result in* (21) *follows.*

On the other hand, using Bayes' rule, we have that the joint pdf of $\vec{x}_{0:N-1}$ *and* $\vec{y}_{0:N}$ *is given by:*

$$p(\vec{x}_{0:N-1}, \vec{y}_{0:N}) = p(x_0)\prod_{t=0}^{N-1} p(x_{t+1}, y_t | x_t) \quad (25)$$

Considering that ζ_t *given* ξ_t *is Gaussian with mean* $\Theta\xi_t$ *and variance* P*, that* $x_0 \sim N(\mu_0, P_0)$*, and applying the logarithm to both sides of* (25)*, we obtain* (24). □□□

Lemma 4 *For the dynamic system represented in state-space form in* (18)-(19)*, and considering the complete data is given by* $C = \left[\vec{x}_{0:N-1}^T \ \vec{y}_{0:N-1}^T\right]^T$*, the intermediate function* $\mathbf{Q}(\beta, \beta_i)$ *in the EM algorithm for the ML problem in* (11) *with incomplete data* g *(a function of the data* $\vec{y}_{0:N-1}$*) is given by:*

$$\begin{aligned} \mathbf{Q}(\beta,\beta_i) &= \mathbb{E}\left\{\log\left[p(\vec{x}_{0:N}, \vec{y}_{0:N-1}|\beta)\right] | g, \beta_i\right\} \\ &= -\frac{1}{2}\log\det\{P_0\} - \frac{N}{2}\log\det\{P\} \\ &-\frac{1}{2}\mathrm{tr}\left\{P_0^{-1}[(\hat{x}_{0|g}-\mu_0)(\hat{x}_{0|g}-\mu_0)^T + \Sigma_{0|g}]\right\} \\ &-\frac{1}{2}\mathrm{tr}\left\{P^{-1}[\Xi - \Psi\Theta^T - \Theta\Psi^T + \Theta\Delta\Theta^T]\right\} \\ &-[N(n+n_y)+n]\log 2\pi \end{aligned}$$

where

$$\Xi = \sum_{t=0}^{N-1}\Sigma_{\zeta_t}, \ \Psi = \sum_{t=0}^{N-1}\Sigma_{\zeta_t\xi_t}, \ \Delta = \sum_{t=0}^{N-1}\Sigma_{\xi_t}$$

and

$$\begin{aligned} \hat{x}_{0|g} &= \mathbb{E}\{x_0|g,\beta_i\} \\ \hat{\zeta}_t &= \mathbb{E}\{\zeta_t|g,\beta_i\} \\ \hat{\xi}_t &= \mathbb{E}\{\xi_t|g,\beta_i\} \end{aligned} \quad (26)$$

$$
\begin{aligned}
\Sigma_{0|g} &= \mathbb{E}\left\{(x_0 - \hat{x}_{0|g})(x_0 - \hat{x}_{0|g})^T|g, \beta_i\right\} \\
\Sigma_{\zeta_t} &= \mathbb{E}\left\{(\zeta_t - \hat{\zeta}_t)(\zeta_t - \hat{\zeta}_t)^T|g, \beta_i\right\} \\
\Sigma_{\zeta_t\xi_t} &= \mathbb{E}\left\{(\zeta_t - \hat{\zeta}_t)(\xi_t - \hat{\xi}_t)^T|g, \beta_i\right\} \\
\Sigma_{\xi_t} &= \mathbb{E}\left\{(\xi_t - \hat{\xi}_t)(\xi_t - \hat{\xi}_t)^T|g, \beta_i\right\} \qquad (27)
\end{aligned}
$$

PROOF. *The result is directly obtained by using Lemma 3, and Corollary 17.* □□□

Remark 5 *Notice that the result in Lemma 4 is valid for any estimation problem for the system* (18)-(19) *using incomplete data (provided the coarsening data mechanism is CAR). However, it is necessary to calculate the quantities in* (26) *to* (27) *for the specific problem of interest.* ▽▽▽

Remark 6 *If* $g = L\vec{y}_{0:N-1}$ *where* $L \in \mathbb{R}^{m\times N_y}$ *is a deterministic matrix then the quantities from* (26) *to* (27) *can be calculated using Lemma 16 with* $T = \begin{bmatrix} 0 & L \end{bmatrix}$ *and* $\vec{z} = \begin{bmatrix} \vec{x}_{0:N} \\ \vec{y}_{0:N-1} \end{bmatrix}$ ▽▽▽

We next give the details of the M-step in the EM algorithm.

Proposition 7 *The stationary points of* $\mathbf{Q}(\beta, \beta_i)$ *(with variables* Θ *and* P^{-1}*) are given by*

$$
\begin{aligned}
\Theta &= \Psi\Delta^{-1} \\
P^{-1} &= \left[\Xi - \Psi\Delta^{-1}\Psi\right]^{-1}
\end{aligned}
$$

PROOF. *Directly by using Lemma 14.* ▽▽▽

Remark 8 *Note that a Cholesky factorisation can be used to calculate* P *in a numerically robust fashion. This has been previously used in [15] for the time-domain EM algorithm.* ▽▽▽

Remark 9 *If the parameters are required to satisfy linear constraints, then it is possible to obtain an optimisation algorithm that freezes some parameters and optimises the non-frozen ones. This approach will lead to a generalised EM algorithm. See [4] for further details.* ▽▽▽

4.2 Direct maximisation of the likelihood function

An important property of the EM algorithm is that the first and second order derivatives of the log-likelihood function can be obtained by taking derivatives of the intermediate function $\mathbf{Q}(\beta, \beta_i)$ and the joint pdf of $\vec{x}_{0:N-1}, \vec{y}_{0:N-1}$. The following lemma formally states how these derivatives can be obtained:

Lemma 10 *The first and second order derivatives of the log-likelihood function* $l(\beta)$ *are given by the Fisher and Louis identities:*

$$
\begin{aligned}
\left.\frac{\partial l(\beta)}{\partial \beta}\right|_{\beta_i} &= \left.\frac{\partial \mathbf{Q}(\beta, \beta_i)}{\partial \beta}\right|_{\beta_i} \\
\left.\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}\right|_{\beta_i} &= M - vv^T =: H_i
\end{aligned}
$$

where

$$
\begin{aligned}
M &= \left.\frac{\partial^2 \mathbf{Q}(\beta, \beta_i)}{\partial \beta \partial \beta^T}\right|_{\beta_i} + \mathbb{E}\left\{zz^T|\vec{g}, \beta_i\right\} \\
v &= \left.\frac{\partial \mathbf{Q}(\beta, \beta_i)}{\partial \beta}\right|_{\beta_i} \\
z &= \left.\frac{\partial \log p(\vec{x}_{0:N}, \vec{y}_{0:N-1}|\beta)}{\partial \beta}\right|_{\beta_i}
\end{aligned}
$$

PROOF. *See e.g. [10, page 354].* □□□

Remark 11 *Notice that* $\mathbb{E}\left\{zz^T|\vec{g}, \beta_i\right\}$ *can be efficiently obtained by using Lemmas 14, 16 and 18.* ▽▽▽

Remark 12 *Notice that, using Lemma 10, it is possible to develop a Newton-like algorithm. The inverse of the Hessian of the log-likelihood function might be calculated by using the Matrix Inversion Lemma (see e.g. [45, Chapter 15]).* ▽▽▽

In [52] it was pointed out that the EM-based estimation algorithm typically behaves better than Newton-based algorithms when the initial estimates for the parameters are far from the true value. However, the EM algorithm slows down when the parameter estimate is close to the value that maximises the likelihood function. Thus, in [52] it is advocated that an algorithm that combines both approaches should be utilised. In addition, notice that the quantities used in the EM algorithm can be directly used in the implementation of a Newton-like algorithm (see Remark 12).

5 Numerical examples

5.1 Example: Comparison with other methods

In this numerical example, we compare the performance of the proposed algorithm with two previous approaches: (i) the method in [40] that develops an ML estimate for the parameters of a linear system in transfer function representation. The algorithm uses the traditional frequency domain identification approach and considers missing data as extra parameters to be estimated; and (ii) the methods in [2,54] where an EM algorithm for systems represented in state space form is used. These approaches utilise data in the frequency domain, and do not cover the case of missing data in the time domain.

We compare our algorithm with a combination of the two approaches above, i.e., we use the algorithm in [40] for a system represented in state space form as in [2,54]. We call this method the "current approach".

We consider the system given in (18)-(20), with

$$A = \begin{bmatrix} 1 & -0.8 \\ 1 & 0 \end{bmatrix}, B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, C = \begin{bmatrix} 1 & 0.6 \end{bmatrix}, D = 0 \quad (28)$$

$Q = 0.01 I_2$, $S = [0\ 0]^T$ and $R = 0.01$, and the input signal is zero mean Gaussian white noise with variance $\sigma_u^2 = 4$.

The transfer function representation for the system of interest is given by:

$$Y_k = G_o(z_k)U_k + H_o(z_k)W_k + V_k + T_o(z_k)$$

where

$$\begin{aligned} G_o(z_k) &= C(z_k I - A)^{-1}B + D \qquad (29) \\ H_o(z_k) &= C(z_k I - A)^{-1} \\ T_o(z_k) &= C(z_k I - A)^{-1}\alpha \end{aligned}$$

where $z_k = e^{j\omega_k}$, $\omega_k = \frac{2\pi}{N}k$, and $\alpha = x_0 - x_N$. We consider α as an extra parameter to be estimated (see [11,2,54,3]).

We use $N = 1024$ data points for each one of 100 Monte-Carlo simulations.

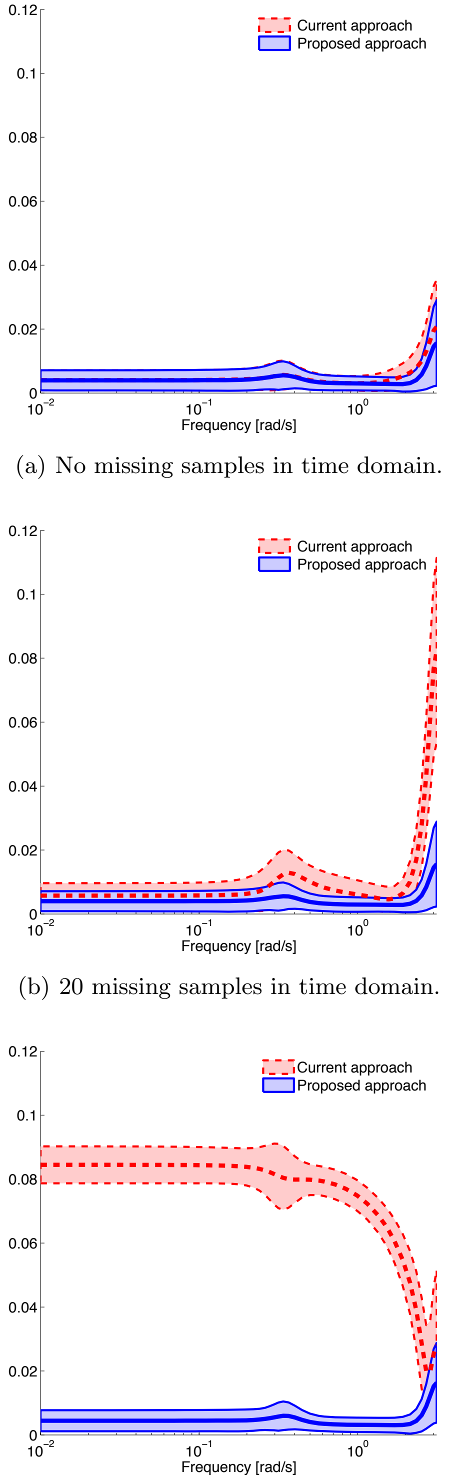
Three cases are analysed: (i) when all data is available for the estimation, (ii) when 20 consecutive samples of the output data are missing, (iii) when 200 consecutive missing samples of the output data are missing.

Figures 1(a)-1(c) show the mean and one standard deviation band (over all the Monte-Carlo simulations) of the magnitude of the relative error given by:

$$R_i(e^{j\omega_k}) := |(G_0(e^{j\omega_k}) - \hat{G}_i(e^{j\omega_k}))/G_0(e^{j\omega_k}))| \quad (30)$$

where G_0 is described in (28)-(29), and G_i is the corresponding transfer function $\hat{G}_i = \hat{C}_i(zI - \hat{A}_i)^{-1}\hat{B}_i + \hat{D}_i$ corresponding to the estimated model in the i^{th} Monte-Carlo simulation. The magnitude of the relative error is computed for each ω_k belonging to a finite set in the range $(0, \pi)$.

Figure 1(a) shows the magnitude of the relative error when all the data is available for estimation (no missing data). We observe that the proposed approach provides similar results to the combination of the previous approaches available in the literature.



(a) No missing samples in time domain.

(b) 20 missing samples in time domain.

(c) 200 missing samples in time domain.

Fig. 1. Relative error magnitude $|(G_0 - \hat{G}_i)/G_0|$ for both approaches. Estimating the missing data as extra parameters (red-dashed line) and as hidden random variables (blue-continuous line). Mean value of the relative error appears with a thick line. The maximum deviation of the relative error for all Monte-Carlo runs appears with a thin line.

Figure 1(b) shows the magnitude of the relative error when 20 consecutive samples of output data are missing. We observe that both methods provide good estimates. However, the estimates obtained with the proposed approach have a smaller relative error.

Figure 1(c) shows the magnitude of the relative error when 200 consecutive output samples are missing. We see that the proposed approach still provides good estimates. However, the "current approach" is not able to estimate the system due to the large amount of missing data.

5.2 Example: Impact of incomplete data for the identification of systems subject to under-modelling

Consider a first order discrete-time linear system represented in state-space form as in (18)-(20). The input signal is chosen as zero mean Gaussian white noise with variance $\sigma_u^2 = 4$. The corresponding input-output representation of the system is given by

$$G_0(z_k) = C(z_k I - A)^{-1}B + D = \frac{0.25}{z_k - 0.75}$$

$N = 1024$ data points are utilised to identify the system in the case where there is no missing data.

We run 100 Monte-Carlo simulations for different noise realisations and we estimate the state space matrices corresponding to a first order system.

We impose the constraint $D = 0$ to identify the system for all cases and obtain the estimates by using the proposed algorithm. The purpose of this numerical example is to illustrate the flexibility of the dual time-frequency domain approach. The computation of the expected values in (26)-(27) might increase the computational load of the complete algorithm. However, the complexity of the key steps of the numerical procedure is similar to the standard EM algorithm applied to state-space systems using time domain data. It is beyond of the scope of the current paper to delve further into numerically efficient algorithms.

We test our approach with **two different sets of data**:

Set (i): No under-modelling: Here we generate the data using a first order model with $A = 0.75$, $B = 0.5$, $C = 0.5$, $D = 0$, $Q = 0.01$, $R = 0.01$ and $S = 0$.

Set (ii): With under-modelling: Here, the data is generated by a second order state-space model with the following matrices:

$$A = \begin{bmatrix} 0.5 & 0.4472 \\ 0 & 0.75 \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \tag{31}$$

$$C = \begin{bmatrix} 0.1847 & 0.413 \end{bmatrix} \qquad D = 0$$

$$Q = 0.01 I_2 \quad R = 0.01 \qquad S = 0 \tag{32}$$

This model, compared to the model for data set (i), includes a *fast* pole located at $z = 0.5$ which has an effect for frequencies above 0.7 [rad/s], approximately.

We consider **three different cases for state-space model estimation**:

(i) Full data: In this case we perform estimation with the full set of data. Notice that this can be performed either in frequency or time domain, leading to the same system estimate.

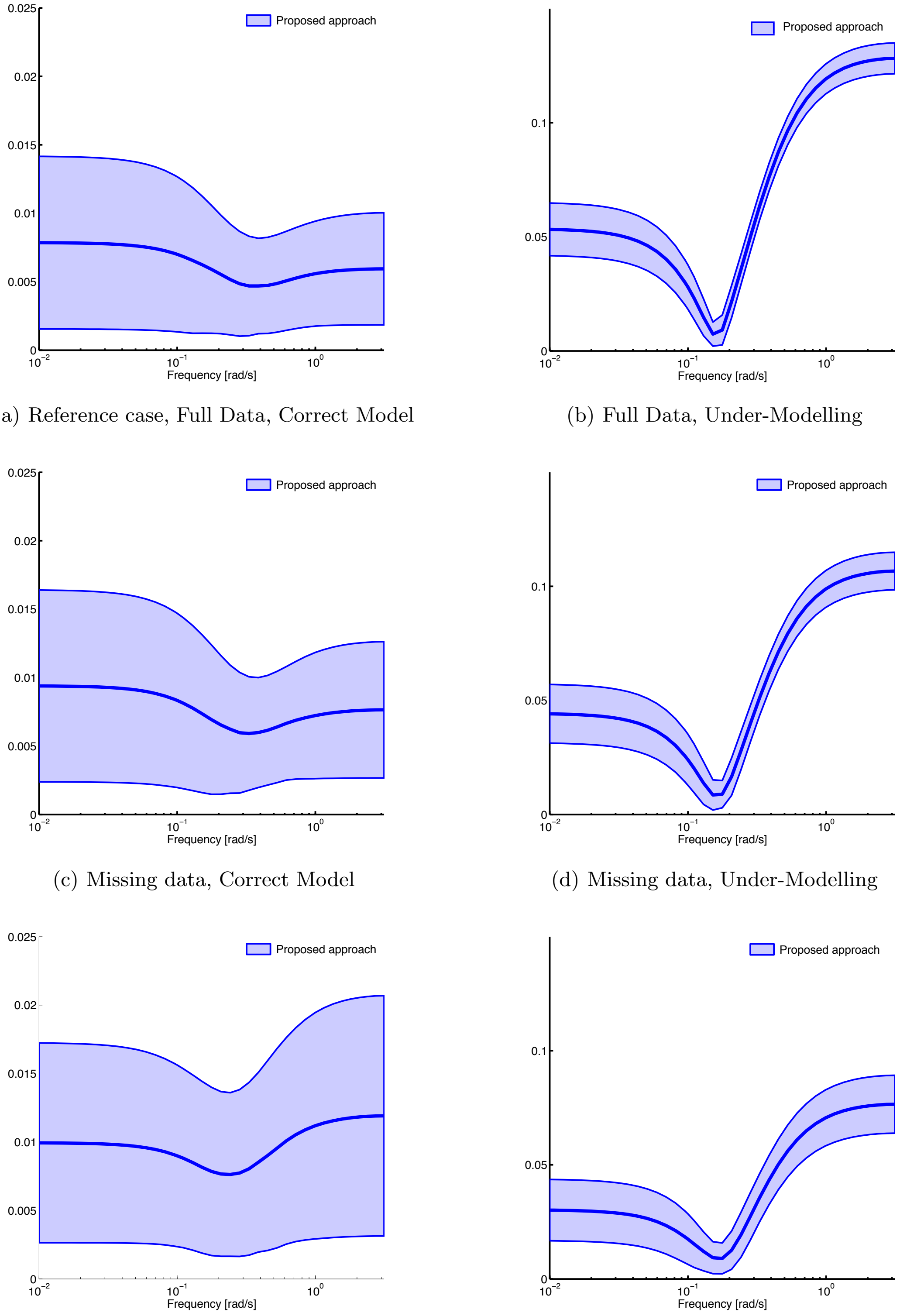
(ii) Missing data in the time domain: In this case, the missing data mechanism s_t is given by the model (13)-(14) with $\varsigma = 0.7$, and $\delta = 0.7$. We use the same realisation for the missing data mechanism s_t, corresponding to 557 missing samples, for all the Monte-Carlo simulations.

(iii) Robust identification with missing data: In this case, we consider the missing time domain data described in (ii) and, additionally, we restrict the frequency domain estimation bandwidth to the range from 0 to $\pi/5[rad/s] \approx 0.62[rad/s]$, i.e., below the frequencies where the *unmodeled fast pole* in (31)-(32) has an effect.

In summary, six different situations are considered as shown in Figure 2. This figure shows the mean ± one standard deviation of the relative error (30) over all Monte-Carlo simulations (with respect to the nominal first order model). The plots on the left hand side (i.e, Figures 2(a), 2(c), and 2(e)) correspond to the first set of data, i.e., when there is no under-modelling in the estimation process. The plots on the right hand side of the figure (i.e., Figures 2(b), 2(d), and 2(f)) correspond to the second set of data generated by the system (31)-(32) and we fit a first order model. Notice that the vertical scale between left and right plots is different since (as one would expect) when under-modelling is introduced, there is an increment (of one order of magnitude) in the relative error.

Figure 2(a), shows the magnitude of the relative error, when the full data is available and the system and model have the same structure (first order). We use the data corresponding to all the frequencies in the estimation which is equivalent to the usual case of time domain identification. We refer to this simulation as *Reference Case.* On the other hand, Figure 2(b) shows the relative error when under-modelling is present at high frequencies (i.e. the data is generated by the second order model). This leads to an increase in the relative error compared to the reference case, in particular, at high frequencies due to the unmodeled fast pole.

Figures 2(c) and 2(d) show the results when missing time domain data is considered. Figure 2(c) shows a slight increase in the relative error compared to the Reference case. On the other hand, Figure 2(d) shows the relative error when under-modelling is present at high frequencies. In this case, the missing data helps to obtain a bet-



(a) Reference case, Full Data, Correct Model

(b) Full Data, Under-Modelling

(c) Missing data, Correct Model

(d) Missing data, Under-Modelling

(e) Robust identification, Missing Data, Correct Model

(f) Robust identification, Missing Data, Under-Modelling

Fig. 2. Relative error magnitude $|(G_0 - \hat{G})/G_0|$ for all simulations corresponding to the example presented in Section 5.

ter result than those obtained when there is no missing data (see Figure 2(b)) but the system is identified using the structure corresponding to the nominal model (first order).

Finally, Figures 2(e) and 2(f) shows the relative error for the case when there is missing data, and we use frequency domain data in a reduced frequency range. Figure 2(e) shows that the relative error is greater than that obtained corresponding to the missing data case in Figure 2(c), specially at high frequencies. In this case, the variance of the estimates increases since only a part of the data is utilised for identification, but in fact, there is no under-modelling. On the other hand, Figure 2(f) shows the relative error when there is time domain missing data, we use frequency domain data in a reduced frequency range, and the system is identified using the structure corresponding to the nominal model (first order), but the data is generated by the second order system. Note, that the relative error at all frequencies is smaller than the relative errors obtained in all other cases where under-modelling is present (see Figures 2(b) and 2(d)). This confirms that using incomplete data (discarding or missing data) for identification may be beneficial in the presence of modelling errors.

6 Conclusions

In this paper, we have considered the problem of obtaining an ML estimate for the parameters of a linear discrete-time-invariant stochastic system. We have presented new ideas to perform *Robust System Identification* in the presence of *missing-data* in the time domain. These ideas have led to a new procedure which we refer to as *Dual Time-Frequency Domain Identification.*

We have also presented an estimation procedure based on the EM algorithm to identify a system represented in state-space form. We developed a Newton-like algorithm that calculates the derivatives of the likelihood function using the quantities obtained in the EM algorithm. The flexibility of the approach developed in this paper has been illustrated by several numerical examples.

It is important to note that the results show that, when the system structure is known (no under-modelling), using incomplete data for identification (missing data in time domain or data selection frequency domain) leads to deterioration of the estimation accuracy. On the other hand, when under-modelling is present (a very common case), using incomplete data for identification may improve the *quality* of the estimated models.

Acknowledgements

We thank the Associate Editor and the Reviewers for their helpful comments and suggestions.

References

[1] A. A. Afifi and R. M. Elasshoff. Missing observations in mutivariate statistics. *Journal of the American Statistical Association*, 61(315):595–604, 1966.

[2] J. C. Agüero, J. I. Yuz, and G. C. Goodwin. Frequency domain identification of MIMO state space models using the EM algorithm. In *European Control Conference ECC*, 2007.

[3] J. C. Agüero, J. I. Yuz, G. C. Goodwin, and R. A. Delgado. On the equivalence of time and frequency domain maximum likelihood estimation. *Automatica*, 46(2):260–270, 2010.

[4] J. C. Agüero, J. I. Yuz, G. C. Goodwin, and W. Tang. Identification of state-space systems using a dual time-frequency domain approach. In *Proceedings of the 49th IEEE Conference on Decision and Control, CDC*, 2010.

[5] T. W. Anderson. Maximum-likelihood estimates for a multivariate normal-distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278):200–203, 1957.

[6] M. Banbura, D. Giannone, and L. Reichlin. Nowcasting. *Working paper Series 1275, European Central Bank*, 2010.

[7] D. Bauer. Comparing the CCA subspace method to pseudo maximum likelihood methods in the case of no exogenous inputs. *Journal of Time Series Analysis*, 26(5):631–668, 2005.

[8] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time series analysis: Forecasting and control.* Wiley, fourth edition, 2008.

[9] R. F. Brown and G. C. Goodwin. Efficient calculation of curvature in state and parameter estimation. *Electronics letters*, 6(18):579–580, 1970.

[10] O. Cappé and T. Moulines, E. Rydén. *Inference in hidden markov models.* Springer, 2005.

[11] B. Cauberghe, P. Guillaume, R. Pintelon, and P. Verboven. Frequency-domain subspace identification using FRF data from arbitrary signals. *Journal of Sound and Vibration*, 290(3-5):555–571, 2006.

[12] M. Deistler, K. Peternell, and W. Scherrer. Consistency and relative efficiency of subspace methods. *Automatica*, 31(12):1865–1875, 1995.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from imcomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[14] J. Durbin and S. J. Koopman. *Time series analysis by state space methods.* Oxford University Press, 2005.

[15] S. Gibson and B. M. Ninness. Robust maximum-likelihood estimation of multivariable dynamic systems. *Automatica*, 41(10):1667–1682, 2005.

[16] G. C. Goodwin. Some observations on robust estimation and control. In *7th IFAC Symposium on Identification and System Parameter Estimation*, York, UK, 1985.

[17] G. C. Goodwin, J. C. Agüero, J. S. Welsh, J. I. Yuz, G. J. Adams, and C. R. Rojas. Robust identification of process models from plant data. *Journal of Process Control*, 18:810–820, 2008.

[18] G. C. Goodwin and A. Feuer. Estimation with missing data. *Mathematical and Computer Modelling of Dynamical Systems*, 5(3):220–244, 1998.

[19] G. C. Goodwin, S. F. Graebe, and M. E. Salgado. *Control System Design.* Prentice Hall, Upper Saddle River, NJ, 2001.

[20] G. C. Goodwin and R. Payne. *Dynamic system identification: Experiment design and data analysis.* Academic Press, 1977.

[21] R. B. Gopaluni. A particle filter approach to identification of nonlinear processes under missing observations. *The Canadian Journal of Chemical Engineering*, 86(6):1081–1092, 2008.

[22] N. K. Gupta and R. K. Mehra. Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations. *IEEE Transactions on Automatic Control*, AC-19(6):774–783, 1974.

[23] A. Gut. *An intermediate course in probability.* Springer, second edition, 2009.

[24] E. J. Hannan. *Multiple time series.* John Wiley and Sons, Inc., 1970.

[25] E. J. Hannan. The identification and parameterization of ARMAX and state space forms. *Econometrica*, 44(4):713–723, 1976.

[26] E. J. Hannan and M. Deistler. *The Statistical theory of linear systems.* John Wiley and Sons, Inc., 1988.

[27] E. J. Hannan and P. M. Robinson. Lagged regression with unknown lags. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(2):252–267, 1973.

[28] D. Heitjan and S. Basu. Distinguishing "missing at random" and "missing completely at random". *The American Statistician*, 50(3):207–213, 1996.

[29] D. Heitjan and D. B. Rubin. Ignorability and coarse data. *Annals of Statistics*, 19(4):2244–2253, 1991.

[30] A. Isaksson. Identification of ARX models subject to missing data. *IEEE Transactions on Automatic Control*, 38(5):813–819, 1993.

[31] R. Kan. From moments of sum to moments of product. *Journal of Multivariate Analysis*, 99(3):542–554, 2008.

[32] R. J. Little and D. B. Rubin. On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician*, 37(3):218–220, Aug 1983.

[33] R. J. Little and D. B. Rubin. *Statistical analysis with missing data.* John Wiley & Sons, second edition, 2002.

[34] L. Ljung. *System Identification: Theory for the user.* Prentice Hall, second edition, 1999.

[35] H. Lütkepohl. *New introduction to multiple time series analysis.* Springer, 2005.

[36] T. McKelvey. Frequency domain identification methods. *Circuits Systems Signal Processing*, 21(1):39–55, 2002.

[37] T. McKelvey and A. Helmersson. State space parameterization of multivariable linear systems using tridiagonal matrix form. *Proceedings of the 35th IEEE Conference on Decision and Control, CDC*, pages 3654–3659, 1996.

[38] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley, 1997.

[39] R. Pintelon, G. Schoukens, T. McKelvey, and Y. Rolain. Minimum variance bounds for overparameterized models. *IEEE Transactions on Automatic Control*, 41(5):719–720, 1996.

[40] R. Pintelon and J. Schoukens. Frequency domain system identification with missing data. *IEEE Transactions on Automatic Control*, 45(2):364–369, 2000.

[41] R. Pintelon and J. Schoukens. *System Identification: A frequency domain approach.* IEEE Press, 2001.

[42] G. C. Reinsel. *Elements of multivariable time series analysis.* Springer, 1997.

[43] D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[44] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. Sastry. Foundations of control and estimation over lossy networks. *Proceedings of the IEEE*, 95(1):163 – 187, 2007.

[45] G. A. F. Seber. *A matrix handbook for statisticians.* Wiley, 2008.

[46] M. Segal and E. Weinstein. A new method for evaluating the log-likelihood gradient, the Hessian and the Fisher information matrix for linear dynamic systems. *IEEE Transactions on Information Theory*, 35(3):682–687, 1989.

[47] R. Shumway and D. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 1982.

[48] T. Söderström and P. Stoica. *System identification.* Prentice-Hall International, 1989.

[49] M. Tanaka and T. Katayama. Robust identification and smoothing for linear system with outliers and missing data. In *11th IFAC World Congress*, volume 3, pages 160–165, 1990.

[50] F. Vaida. Parameter convergence for EM and MM algorithms. *Statistica Sinica*, 15:831–840, 2005.

[51] P. Van Overschee and B. De Moor. *Subspace identification for linear systems: Theory-implementation-applications.* Kluwer Academic Publishers, 1996.

[52] M. W. Watson and R. F. Engle. Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23:385–400, 1983.

[53] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, Vol. 50(1):1–25, 1982.

[54] A. Wills, B. Ninness, and S. Gibson. Maximum likelihood estimation of state space models from frequency domain data. *IEEE Transactions on Automatic Control*, 54(1):19 – 33, 2009.

[55] C. F. J. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

[56] J. I. Yuz and G. C. Goodwin. Robust identification of continuous-time systems from sampled data. In H. Garnier and L. Wang, editors, *Continuous-time model identification from sampled data*, chapter 3, pages 67–89. Springer, 2008.

7 Appendix

Lemma 13 *Let A, B, C, D, and X be matrices of appropriate dimensions, then*

(1) $\mathrm{vec}\{ABC\} = [C^T \otimes A]\mathrm{vec}\{B\}$.
(2) $[A \otimes B]^H = A^H \otimes B^H$.
(3) $[A \otimes B]^{-1} = A^{-1} \otimes B^{-1}$.
(4) $\mathrm{tr}\{ABCD\} = \mathrm{vec}\{D\}^T [A \otimes C^T]\mathrm{vec}\{B^T\}$.
(5) $\mathrm{tr}\{ABC\} = \mathrm{tr}\{CAB\} = \mathrm{tr}\{BCA\}$.
(6) $\frac{\partial \mathrm{tr}\{AXB\}}{\partial X} = A^T B^T$.
(7) $\frac{\partial \mathrm{tr}\left\{AXBX^T\right\}}{\partial X} = A^T X B^T + AXB$.
(8) $\frac{\partial \log\det\{X\}}{\partial X} = X^{-T}$.
(9) $\frac{\partial \mathrm{vec}\left\{X^{-1}\right\}}{\partial(\mathrm{vec}\{X\})^T} = -[X^{-T} \otimes X^{-1}]$.

PROOF. *See e.g. [45, Chapters 4, 11 and 17].* □□□

Lemma 14 *Let the cost function $J(X, Y)$ be*

$$J(X, Y) = \alpha \log\det\{X\} + \beta \mathrm{tr}\{XZ(Y)\}$$
$$Z(Y) = [\Xi - \Psi Y^T - Y\Psi^T + Y\Delta Y^T]$$

where $X = X^T \in \mathbb{R}^{m\times m}$, and $Y \in \mathbb{R}^{m\times n}$ are matrices, $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$ are constants, $\Xi = \Xi^T$, $\Delta = \Delta^T$, and Ψ are constant matrices of appropriate dimension. Then, the partial derivatives of $J(X, Y)$ with respect to the matrices X and Y are given by:

$$\frac{\partial J(X,Y)}{\partial X} = \alpha X^{-1} + \beta Z(Y)$$
$$\frac{\partial J(X,Y)}{\partial Y} = 2\beta X[Y\Delta - \Psi]$$

PROOF. *Directly by using Lemma 13.* □□□

7.1 Gaussian distributions

A random vector $x \in \mathbb{R}^{n\times 1}$ is said to be real (non-singular) Gaussian distributed (denoted $x \sim \mathrm{N}(\mu, \Sigma)$) if its pdf is given by [45, Chapter 20]:

$$p(x) = \det\{2\pi\Sigma\}^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}[x-\mu]^T\Sigma^{-1}[x-\mu]\right\}$$

where μ and Σ (positive definite) are the mean and covariance of x. In addition, if Σ is positive semidefinite (i.e. singular) the probability distribution still exists but not the density function. In this case, $x \in \mathbb{R}^{n\times 1}$ is real Gaussian distributed if and only if $\alpha^T x$ is univariate normal for all $\alpha \in \mathbb{R}^n$. If $x = a \in \mathbb{R}$, we define x to be $x \sim \mathrm{N}(a, 0)$.

Lemma 15 *Let $\begin{bmatrix} X^T & Y^T \end{bmatrix}^T$ be a real Gaussian distributed random vector with mean $\begin{bmatrix} \mu_X^T & \mu_Y^T \end{bmatrix}^T$ and variance $\begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}$ (possibly singular). Then, X given Y is a Gaussian distributed random vector with mean and variance given by:*

$$\mu_{X|Y} = \mu_x + \Sigma_{xy}\Sigma_y^{\dagger}(Y - \mu_y)$$
$$\Sigma_{X|Y} = \Sigma_x - \Sigma_{xy}\Sigma_y^{\dagger}\Sigma_{yx}$$

respectively.

PROOF. *See [45, Chapter 20].* □□□

Lemma 16 *If $\vec{z} \in \mathbb{R}^n$ is a Gaussian random vector with mean μ_1 and a non-singular covariance matrix Σ_1 (i.e. $\vec{z} \sim N(\mu_1, \Sigma_1)$), and $\vec{g} = T\vec{z}$ where $T \in \mathbb{R}^{m\times n}$ is a deterministic matrix ($m \le n$), then $\vec{z}$ given $\vec{g}$ is a Gaussian random vector with mean μ_2 and covariance Σ_2 given by $\vec{z}|\vec{g} \sim N(\mu_2, \Sigma_2)$, where*

$$\mu_2 = \mu_1 + \Sigma_1 T^T (T\Sigma_1 T^T)^{\dagger}(\vec{g} - T\mu_1)$$
$$\Sigma_2 = \Sigma_1 - \Sigma_1 T^T (T\Sigma_1 T^T)^{\dagger} T\Sigma_1$$

PROOF. *Considering that $\begin{bmatrix} \vec{z}^T & \vec{g}^T \end{bmatrix}^T = \begin{bmatrix} I & T^T \end{bmatrix}^T \vec{z}$, we have that $\begin{bmatrix} \vec{z}^T & \vec{g}^T \end{bmatrix}^T$ is a Gaussian random vector with mean $\begin{bmatrix} \mu_1^T & (T\mu_1)^T \end{bmatrix}^T$ and variance $\begin{bmatrix} \Sigma_1 & \Sigma_1 T^T \\ T\Sigma_1 & T\Sigma_1 T^T \end{bmatrix}$. The result then follows using Lemma 15.* □□□

7.2 Expected Value of a Quadratic term

Lemma 17 *Let $\zeta \in \mathbb{R}^m$ and $\xi \in \mathbb{R}^n$ be random vectors, and let $\Theta \in \mathbb{R}^{m\times n}$ and $P^{-1} \in \mathbb{R}^{m\times m}$ be deterministic matrices, then*

$$\mathbb{E}\left\{(\zeta - \Theta\xi)^T P^{-1}(\zeta - \Theta\xi)\right\} = (\hat{\zeta} - \Theta\hat{\xi})^T P^{-1}(\hat{\zeta} - \Theta\hat{\xi}) + \mathrm{tr}\left\{P^{-1}\left[\Sigma_\zeta - \Theta\Sigma_{\zeta\xi}^T - \Sigma_{\zeta\xi}\Theta^T + \Theta\Sigma_\xi\Theta^T\right]\right\}$$

where $\hat{\xi} = \mathbb{E}\{\xi\}$, $\hat{\zeta} = \mathbb{E}\{\zeta\}$, $\Sigma_{\zeta\xi} = \mathbb{E}\left\{(\zeta - \hat{\zeta})(\xi - \hat{\xi})^T\right\}$, $\Sigma_\xi = \mathbb{E}\left\{(\xi - \hat{\xi})(\xi - \hat{\xi})^T\right\}$, $\Sigma_\zeta = \mathbb{E}\left\{(\zeta - \hat{\zeta})(\zeta - \hat{\zeta})^T\right\}$.

PROOF. *The result is obtained by expanding the term on the left hand side and using the properties of the trace and expected value operators.* □□□

Lemma 18 *Let $\vec{z} = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix}^T \in \mathbb{R}^n$ be a Gaussian random vector with mean μ and covariance Σ (i.e. $\vec{z} \sim N(\mu, \Sigma)$), then for nonnegative integers s_1 to s_n, we have:*

$$\mathbb{E}\left\{\prod_{i=1}^{n} z_i^{s_i}\right\} = \sum_{\nu_1=0}^{s_1} \cdots \sum_{\nu_n=0}^{s_n} \sum_{r=0}^{\lfloor s/2 \rfloor} (-1)^m f(\nu_1, \cdots, \nu_n)$$
$$f(\nu_1, \cdots, \nu_n) = \begin{pmatrix} s_1 \\ \nu_1 \end{pmatrix} \cdots \begin{pmatrix} s_n \\ \nu_n \end{pmatrix} \frac{\left(\frac{h^T\Sigma h}{2}\right)(h^T\mu)^{s-2r}}{r!(s-2r)!}$$

where $m = \sum_{i=1}^n \nu_i$, z_i is the i-th component of the vector $\vec{z}$, $\lfloor s/2 \rfloor$ denotes the largest integer smaller than or equal to $s/2$, $s = s_1 + \cdots s_n$, $h = \begin{bmatrix} \frac{s_1}{2} - \nu_1 & \cdots & \frac{s_n}{2} - \nu_n \end{bmatrix}^T$, and $\begin{pmatrix} a \\ b \end{pmatrix}$ are binomial coefficients.

PROOF. *See [31].* □□□