Maximum Likelihood Estimation of State Space Models from Frequency Domain Data

Adrian Wills Brett Ninness Stuart Gibson

Abstract— **This paper addresses the problem of estimating linear time invariant models from observed frequency domain data. Here an emphasis is placed on deriving numerically robust and efficient methods that can reliably deal with high order models over wide bandwidths. This involves a novel application of the Expectation-Maximisation (EM) algorithm in order to find Maximum Likelihood estimates of state space structures. An empirical study using both simulated and real measurement data is presented to illustrate the efficacy of the EM-based method derived here.**

I. Introduction

A widely used approach to the problem of linear dynamic system identification is to first obtain measurements of a system's frequency response, and then take as estimate the model of given order whose response most closely matches these measurements [13], [15]. This technique may be employed because the system observations naturally present themselves in the frequency domain, such as in the area of vibration analysis [9]. Alternatively, it may be used in order to obtain some of the natural advantages that flow, such as ease of combining separate data sets, ability to compress large data sets, and ease of accommodating continuous time models [16].

However, the approach is not without difficulties. In most cases finding the estimate involves solving a non-convex non-linearly parametrized optimisation problem. This necessitates some sort of iterative search such as a gradient based technique [5]. Furthermore, when the observed data is in the frequency domain, it can have a very wide dynamic range involving several orders of magnitude, with features at small magnitudes being as significant from a modelling viewpoint as features at large magnitudes. Finally, it may may be required for the estimated model to fit the observed data over several decades of frequency.

The combined effect of these issues is that solving a frequency domain estimation problem often implies solving a complex optimisation problem subject to significant numerical conditioning difficulties. A wide variety of techniques have been developed to cope with these difficulties, with a commonly used approach being one of Gauss–Newton type search with respect to models which are specifically parametrized to ameliorate numerical conditioning issues [16].

This paper explores a new approach of tackling the model fitting problem via the Expectation-Maximisation (EM) algorithm. As opposed to gradient-based search which applies to the general class of smooth cost functions, the EM method was specifically developed to tackle (possibly non-smooth) Maximum-Likelihood (ML) criteria. It was originally developed in the statistics literature [4], and enjoys a reputation in that field of providing a particularly robust means for iteratively computing ML estimates.

However, it has not (to the authors knowledge) previously been applied to the frequency domain estimation problems studied here, although the current authors have studied the method for both linear and non-linear estimation from time domain data in [6], [7]. In that context, it was found to provide a method that was particularly robust against capture in local minima, which enjoyed favourable numerical conditioning of component calculations, and which implied reasonable computational load which scaled only modestly with increasing model complexity.

These advantages in the time domain case motivated the frequency domain study here, and as will be illustrated, the same benefits apply in this new setting. Like gradient based search, the algorithm involved is still iterative in nature and consists of two key steps. In the first, a smoothed state estimate based on the current iterate model parameters are computed. In the second, this state estimate is used to update the model parameters via solving two maximisation problem. Unlike the time domain cases studied in [6], [7], this maximisation step cannot be found in closed form and a gradient based search is required. However, the problems involved are convex or close to, and hence simply solved.

The efficacy of the method is illustrated empirically on a range of simulation studies of varying system dimension, and on two real world problems involving a cantilevered beam and a power transformer.

II. Model Structure and Available Data

This paper assumes the availability of a set of measurements $\{Y(\omega_k)\}$ of a system's frequency response at a set $\{\omega_k\}$ of frequencies which need not be regularly spaced, and addresses the problem of finding matrices A, B, C, D in a state space model such that the associated frequency response

$$G(\gamma_k) = C(\gamma_k I - A)^{-1}B + D; \quad \gamma_k = e^{j\omega_k} \text{ or } j\omega_k \quad (1)$$

This work was supported by the Australian Research Council.

A. Wills is with the School of Electrical and Computer Science, University of Newcastle, Australia and can be contacted at email:adrian.wills@newcastle.edu.au

B. Ninness is with the School of Electrical and Computer Engineering, University of Newcastle, Australia and can be contacted at email:brett.ninness@newcastle.edu.au

S. Gibson is with Lehman Brothers, London, UK email:stuart.gibson@lehman.com

matches $Y(\omega_k)$ as closely as possible. Note that in (1) both discrete time and continuous time models are considered.

Of course, it is necessary to be precise about how the quality of fit will be measured. As will presently be detailed, a Maximum-Likelihood criterion will be employed. This arises from certain stochastic assumptions, which may be motivated in several ways. One of these involves employing an underlying time domain description

$$\left[\begin{array}{c} x_{t+1} \\ y_t \end{array}\right] = \left[\begin{array}{cc} A & B \\ C & D \end{array}\right]\left[\begin{array}{c} x_t \\ u_t \end{array}\right] + \left[\begin{array}{c} w_t \\ e_t \end{array}\right] \tag{2}$$

together with an assumption that the measurement $Y(\omega)$ is formed as an N point DFT

$$Y(\omega) \triangleq \frac{1}{\sqrt{N}} \sum_{t=1}^{N} y_t e^{-j\omega t} \tag{3}$$

with input excitation $u_t = e^{j\omega t}/\sqrt{N}$. If $\{w_t\}$ and $\{e_t\}$ in (2) are assumed to be zero mean i.i.d. processes that satisfy

$$\left[\begin{array}{c} w_t \\ v_t \end{array}\right] \sim \mathcal{N}(0, \Pi), \qquad \Pi \triangleq \left[\begin{array}{cc} Q & 0 \\ 0 & R \end{array}\right] \tag{4}$$

then this implies

$$\left[\begin{array}{c} \widetilde{X}(\omega) \\ Y(\omega) \end{array}\right] = \left[\begin{array}{cc} A & B \\ C & D \end{array}\right]\left[\begin{array}{c} X(\omega) \\ 1 \end{array}\right] + \left[\begin{array}{c} W(\omega) \\ E(\omega) \end{array}\right] \tag{5}$$

where $X(\omega)$, $W(\omega)$, $E(\omega)$ are the DFT's of $\{x_t\}$, $\{w_t\}$, $\{e_t\}$ and $\widetilde{X}(\omega)$ is the DFT of $\{x_{t+1}\}$:

$$\begin{aligned} \widetilde{X}(\omega) &\triangleq \frac{1}{\sqrt{N}} \sum_{t=1}^{N} x_{t+1} e^{-j\omega t}, \\ &= e^{j\omega} X(\omega) + \frac{1}{\sqrt{N}} \left[x_{N+1} e^{-j\omega N} - x_1\right]. \end{aligned} \tag{6}$$

Therefore, as N increases $\widetilde{X}(\omega) \to e^{j\omega} X(\omega)$ and the distributional convergence

$$\left[\begin{array}{c} W(\omega) \\ E(\omega) \end{array}\right] \xrightarrow{\mathcal{D}} \mathcal{N}_c(0, \Pi) \tag{7}$$

occurs [3]. Also $E(\omega)$ is asymptotically independent of $E(\lambda)$ for $\omega \neq \lambda$ and similarly for $W(\omega)$. Here $\mathcal{N}_c$ denotes the complex Normal distribution which is defined such that if $Z \sim \mathcal{N}_c(\mu, P)$ for a vector $Z \in \mathbf{C}^n$ then Z has probability density function

$$p(Z) = \frac{1}{\pi^n |P|} \exp\left(-(Z-\mu)^\star P^{-1} (Z-\mu)\right). \tag{8}$$

This leads to the following model for observed frequency domain data which will be used in the remainder of this paper.

$$\left[\begin{array}{c} e^{j\omega} X(\omega) \\ Y(\omega) \end{array}\right] = \left[\begin{array}{cc} A & B \\ C & D \end{array}\right]\left[\begin{array}{c} X(\omega) \\ 1 \end{array}\right] + \left[\begin{array}{c} W(\omega) \\ E(\omega) \end{array}\right] \tag{9}$$

with $E(\omega)$ and $W(\omega)$ being i.i.d

$$\left[\begin{array}{c} W(\omega) \\ E(\omega) \end{array}\right] \sim \mathcal{N}_c(0, \Pi). \tag{10}$$

Note that while (9), (10) have been motivated here via (2) and assumptions of DFT employment, this is just one example of how its utility can be argued [16].

III. Maximum Likelihood Estimation

Given the model structure (9), (10), the work here supposes measurements $Y_k \triangleq Y(\omega_k)$ at M frequencies $\omega_1, \cdots, \omega_M$ and addresses the problem of using this data to form an estimate $\widehat{\theta}$ of the system parametrization

$$\theta \triangleq \text{vec}\left\{[\beta, \eta, \Pi]\right\}, \quad \beta \triangleq \text{vec}\left\{\Theta\right\}, \quad \eta \triangleq \text{vec}\left\{\Gamma\right\} \tag{11}$$

where

$$\Theta \triangleq [A, B], \quad \Gamma \triangleq [C, D]. \tag{12}$$

For this purpose, the paper employs the well-known maximum likelihood criterion

$$\widehat{\theta} \triangleq \arg\max_{\theta} p_\theta(Y_1, \cdots, Y_M)$$

which has well known properties of statistical optimality [16]. This requires the computation of the likelihood $p_\theta(Y_1, \cdots, Y_M)$. To achieve this, note that via (5)

$$Y_k = Y(\omega_k) = G(e^{j\omega_k}) + T(e^{j\omega_k}) W(\omega_k) + E(\omega_k)$$

where

$$T(z) \triangleq C(zI - A)^{-1}, \qquad G(z) \triangleq T(z)B + D. \tag{13}$$

Now assume, according to the motivation of the previous section, that $W(\omega)$ and $E(\omega)$ have complex Normal distributions given by (10). Then

$$Y_k \sim \mathcal{N}_c\left(G(e^{j\omega_k}), P_k\right) \tag{14}$$

where

$$P_k \triangleq C A_k^{-1} Q A_k^{-\star} C^T + R, \qquad A_k \triangleq e^{j\omega_k} I - A. \tag{15}$$

Therefore, by the independence of $W(\omega_k)$ and $W(\omega_\ell)$ for $k \neq \ell$, and similarly for $E(\omega_k)$, $E(\omega_\ell)$

$$\begin{aligned} p_\theta(Y_1, \cdots, Y_M) = \prod_{k=1}^{M} p_\theta(Y_k) = \\ \prod_{k=1}^{M} \frac{1}{\pi^n |P_k|} \exp\left\{-\left\|P_k^{-1/2}(Y_k - G(\omega_k)\right\|^2\right\}. \end{aligned} \tag{16}$$

Actually, since $\log x$ is monotonic in x, then the maximum likelihood estimate can be equivalently defined as

$$\widehat{\theta} = \arg\max_{\theta} L(\theta) \tag{17}$$

where $L(\theta)$ is the log-likelihood function:

$$L(\theta) = \log p_\theta(Y_1, \cdots, Y_M). \tag{18}$$

In this frequency domain case, equation (16) makes it clear that (after dropping terms independent of θ)

$$L(\theta) = \sum_{k=1}^{M} \log |P_k| + \left\|P_k^{-1/2}(Y_k - G(\omega_k)\right\|^2. \tag{19}$$

The essential difficulty now is that $L(\theta)$ is parametrized in quite a complicated way by θ via the formulation of $G(z)$, $T(z)$ and P_k. As such, the computation of $\widehat{\theta}$ via minimisation

of $L(\theta)$ is not straightforward, and certainly it requires some sort of iterative search procedure to be used.

Previous work has addressed this via the employment of a gradient-based search method such as Gauss–Newton iteration or a variant thereof [12], [16]. This paper explores an alternative approach of employing the so-called 'Expectation-Maximisation' (EM) algorithm.

IV. EM Algorithm

To explain and motivate the EM Algorithm it will be convenient to define the full output and full state information sequences as

$$\bar{Y} \triangleq \{Y_1, \ldots, Y_M\}, \quad \bar{X} \triangleq \{X_1, \ldots, X_M\}, \quad X_k \triangleq X(\omega_k). \tag{20}$$

Employing this, an essential ingredient in the understanding of the EM Algorithm is the recognition that via Bayes' rule $p(\bar{Y}) = p(\bar{X}, \bar{Y})/p(\bar{X} \mid \bar{Y})$ and hence the log likelihood $L(\theta)$ can be decomposed as

$$L(\theta) = \log p_\theta(\bar{Y}) = \log p_\theta(\bar{X}, \bar{Y}) - \log p_\theta(\bar{X}|\bar{Y}). \tag{21}$$

Note that in general, the component $\bar{X}$ is termed the 'missing data' and its choice is the main design variable in the deployment of the EM method. The selection here of it being the DFT of the states in an underlying state space model is therefore part of the adaptation of an EM approach to the specific frequency domain estimation setting.

A second essential point is that the expectation of $L(\theta)$ given the output measurements $\bar{Y}$ and with respect to a probability density function implied by parameters θ' is

$$\mathbf{E}_{\theta'}\left\{L(\theta)|\bar{Y}\right\} \quad = \quad \int \log p_\theta(\bar{Y}) p_{\theta'}(\bar{Y}|\bar{Y}) d\bar{Y} = \log p_\theta(\bar{Y}),$$

so that via (21)

$$L(\theta) = \underbrace{\mathbf{E}_{\theta'}\left\{\log p_\theta(\bar{X}, \bar{Y})|\bar{Y}\right\}}_{\mathcal{Q}(\theta,\theta')} - \underbrace{\mathbf{E}_{\theta'}\left\{\log p_\theta(\bar{X}|\bar{Y})|\bar{Y}\right\}}_{\mathcal{V}(\theta,\theta')}. \tag{22}$$

The final key component is that via (22)

$$L(\theta) - L(\theta') = \mathcal{Q}(\theta, \theta') - \mathcal{Q}(\theta', \theta') + \underbrace{\mathcal{V}(\theta', \theta') - \mathcal{V}(\theta, \theta')}_{\geq 0}, \tag{23}$$

where the indicated positivity arises since the term involved is easily established via the definition in (22) as the Kullback-Leibler divergence metric between $p_\theta(\bar{X}|\bar{Y})$ and $p_{\theta'}(\bar{X}|\bar{Y})$.

It follows immediately that if $\mathcal{Q}(\theta, \theta') > \mathcal{Q}(\theta', \theta')$ then $L(\theta) > L(\theta')$ which naturally suggests what is known as the EM algorithm which takes an approximation $\widehat{\theta}_k$ of the Maximum Likelihood estimate $\widehat{\theta}$ given by (17) and updates it to a better one $\widehat{\theta}_{k+1}$ according to:

1) **E Step**

$$\text{Calculate:} \qquad \mathcal{Q}(\theta, \widehat{\theta}_k); \tag{24}$$

2) **M Step**

$$\text{Compute:} \qquad \widehat{\theta}_{k+1} = \arg\max_{\theta \in \Theta} \mathcal{Q}(\theta, \widehat{\theta}_k). \tag{25}$$

Like gradient based search, a single iteration is unlikely to find the optimiser $\widehat{\theta}$ and hence the above steps are iterated multiple times to produce a sequence $\{\widehat{\theta}_0, \widehat{\theta}_1, \widehat{\theta}_2, \ldots\}$ of increasingly good approximations to $\widehat{\theta}$. The iterations are usually terminated using a standard criterion[5] such as the relative decrease of $L(\theta)$ falling below a pre-defined threshold.

V. Computation of $\mathcal{Q}(\theta, \widehat{\theta}_k)$

To implement the EM Algorithm for frequency domain estimation using the missing data choice $\bar{X}$ given by (20) it is clearly necessary to derive a method for computing $\mathcal{Q}(\theta, \widehat{\theta}_k)$. This may be simply achieved via the results of the following two Lemmas.

Lemma 5.1: The function $\mathcal{Q}(\theta, \widehat{\theta}_k)$ defined via (22) may be expressed as

$$\mathcal{Q}(\theta, \widehat{\theta}_k) = -\mathcal{Q}_\Gamma(\eta, R, \widehat{\theta}_k) - \mathcal{Q}_\Theta(\beta, Q, \widehat{\theta}_k) \tag{26}$$

where

$$\mathcal{Q}_\Gamma(\eta, R, \widehat{\theta}_k) \triangleq M \log|R| + \mathrm{Tr}\left\{R^{-1}\left[\Phi - \Gamma\Psi^\star - \Psi\Gamma^T + \Gamma\Sigma\Gamma^T\right]\right\}$$

$$\mathcal{Q}_\Theta(\beta, Q, \widehat{\theta}_k) \triangleq M \log|Q| - 2\sum_{n=1}^{M} \mathrm{Re}\left\{\log|A_n|\right\} + \mathrm{Tr}\left\{Q^{-1}\left[\Lambda - \Theta\Omega^\star - \Omega\Theta^T + \Theta\Sigma\Theta^T\right]\right\}$$

with A_n has been defined in (15) and

$$\Phi \triangleq \sum_{n=1}^{M} Y_n Y_n^\star, \qquad \Omega \triangleq \sum_{n=1}^{N} \mathbf{E}_{\widehat{\theta}_k}\left\{e^{j\omega_n} X_n Z_n^\star \mid Y_n\right\},$$

$$\Psi \triangleq \sum_{n=1}^{M} \mathbf{E}_{\widehat{\theta}_k}\left\{Y_n Z_n^\star \mid Y_n\right\}, \qquad \Lambda \triangleq \sum_{n=1}^{N} \mathbf{E}_{\widehat{\theta}_k}\left\{X_n X_n^\star \mid Y_n\right\},$$

$$\Sigma \triangleq \sum_{n=1}^{M} \mathbf{E}_{\widehat{\theta}_k}\left\{Z_n Z_n^\star \mid Y_n\right\}, \qquad Z_n \triangleq \begin{bmatrix} X_n \\ I \end{bmatrix}.$$

Proof: See [1] for details, which are not presented here due to space restrictions. ■

This reduces the problem of computing $\mathcal{Q}(\theta, \widehat{\theta}_k)$ to one of computing a smoothed state estimate and its covariance, which itself may be achieved by the results of the following Lemma.

Lemma 5.2: Assume that observed data $Y_n = Y(\omega_n)$ obeys the model (9),(10) which is parametrised by a given vector θ according to (11),(12). Then

$$\widehat{X}_n \triangleq \mathbf{E}_\theta\left\{X_n \mid Y_n\right\} = A_n^{-1}B + S_n R_n^{-1}(Y_n - CA_n^{-1}B - D),$$

$$\mathbf{E}_\theta\left\{X_n X_n^\star \mid Y_n\right\} = \widehat{X}_n \widehat{X}_n^\star + Q_n - S_n R_n^{-1} S_n^\star$$

where A_n has been defined in (15) and

$$Q_n \triangleq A_n^{-1} Q A_n^{-\star}, \quad S_n \triangleq Q_n C^T, \quad R_n \triangleq C Q_n C^T + R.$$

Proof: See [1] for details, which are not presented here due to space restrictions. ■

VI. MAXIMISATION OF $\mathcal{Q}(\theta, \widehat{\theta}_k)$

Together, Lemmas 5.1 and 5.2 provide a simple means for the 'expectation step' (24) of computing $\mathcal{Q}(\theta, \widehat{\theta}_k)$. This leads to consideration of the 'maximisation step' (25). Unfortunately, this is not so straightforward since it cannot be achieved by closed form expressions as was just possible in the expectation step.

In order to address this, recall that via (26), $\mathcal{Q}(\theta, \widehat{\theta}_k)$ can be decomposed into two components $\mathcal{Q}_\Gamma(\eta, R, \widehat{\theta}_k)$ and $\mathcal{Q}_\Theta(\beta, Q, \widehat{\theta}_k)$ which depend in a non-coupled fashion on particular elements of the state space description A, B, C, D, Q, R as defined via the parametrization in (11),(12).

This allows the optimisation problem (25) to be decomposed into smaller ones. The first of these involves maximisation (minimisation of $\mathcal{Q}_\Gamma(\eta, R, \widehat{\theta}_k)$ and $\mathcal{Q}_\Theta(\beta, Q, \widehat{\theta}_k)$) with respect to the noise parametrizations R and Q, which does have a closed form solution.

Lemma 6.1: For any $\eta = \text{vec}\{[C, D]\}$

$$\widehat{R}(\eta) \triangleq \frac{1}{M}\left[\Phi - \Gamma\Psi^\star - \Psi\Gamma^T + \Gamma\Sigma\Gamma^T\right] \tag{27}$$

is a stationary point of $\mathcal{Q}_\Gamma(\eta, R)$. Likewise, for any $\beta = \text{vec}\{[A, B]\}$

$$\widehat{Q}(\beta) \triangleq \frac{1}{M}\left[\Lambda - \Theta\Omega^\star - \Omega\Theta^T + \Theta\Pi\Theta^T\right] \tag{28}$$

is a stationary point of $\mathcal{Q}_\Theta(\beta, Q)$.

Proof: See [1] for details, which are not presented here due to space restrictions. ∎

Substitution of these minimisers for the Q and R terms in $\mathcal{Q}_\Gamma(\eta, R, \widehat{\theta}_k)$ and $\mathcal{Q}_\Theta(\beta, Q, \widehat{\theta}_k)$) then leads to 'concentrated' versions that depend only on the system parameters $\beta = \text{vec}\{\Theta\} = \text{vec}\{[A, B]\}$ and $\eta = \text{vec}\{\Gamma\} = \text{vec}\{[C, D]\}$ according to

$$\mathcal{Q}_\Gamma(\eta, \widehat{R}(\eta)) = M \log\left|\frac{1}{M}\left[\Phi - \Gamma\Psi^\star - \Psi\Gamma^T + \Gamma\Sigma\Gamma^T\right]\right| + Mp, \tag{29}$$

and

$$\begin{aligned}\mathcal{Q}_\Theta(\beta, \widehat{Q}(\beta)) =& M \log\left|\frac{1}{M}\left[\Lambda - \Theta\Omega^\star - \Omega\Theta^T + \Theta\Pi\Theta^T\right]\right| + \\ & Mn - 2\sum_{n=1}^{M} \text{Re}\{\log|A_n|\}.\end{aligned} \tag{30}$$

Considering (29), since

$$\begin{aligned}&\Phi - \Gamma\Psi^\star - \Psi\Gamma^T + \Gamma\Sigma\Gamma^T = \\ &\quad (\Gamma - \Psi\Sigma^{-1})\Sigma(\Gamma - \Psi\Sigma^{-1})^\star + \Phi - \Psi\Sigma^{-1}\Psi^\star\end{aligned} \tag{31}$$

then it is tempting to conclude via Minkowski's inequality for determinants of positive definite matrices [8] that

$$\widehat{\eta} = \text{vec}\left\{\widehat{\Gamma}\right\}, \quad \widehat{\Gamma} = \Psi\Sigma^{-1} \tag{32}$$

globally minimises (29). Unfortunately, this overlooks the constraint that $\widehat{\eta}$ must be real valued, and it is not clear how it can be simply accommodated while retaining a closed form solution. As well, (30) has an additional complicating factor due to the presence of the $\sum_{n=1}^{M} \text{Re}\{\log|A_n|\}$ term.

To address this, a Newton-type gradient based search, of the form

$$\widehat{\eta}_{k+1} = \widehat{\eta}_k - H_\Gamma^{-1}(\widehat{\eta}_k) g_\Gamma(\widehat{\eta}_k) \tag{33}$$

initialised at $\widehat{\eta}_0 = \text{vec}\left\{\text{Re}\left\{\Psi\Sigma^{-1}\right\}\right\}$ is proposed where $g_\Gamma(\eta)$ and $H_\Gamma(\eta)$ are (respectively) the gradient and Hessian of $\mathcal{Q}_\Gamma(\eta, \widehat{R}(\eta))$. The rationale is that since $\mathcal{Q}_\Gamma(\eta, \widehat{R}(\eta))$ is real valued, and hence its gradient and Hessian are real valued, then the search (33) automatically respects the realness constraint on $\widehat{\eta}$. Furthermore, since (29) is convex in η, then the Newton based search (33) can be expected to rapidly converge to a global minimiser of $\mathcal{Q}_\Gamma(\eta, \widehat{R}(\eta))$ [5].

An identical argument and approach applies to deriving a real valued minimiser $\widehat{\beta}$ of $\mathcal{Q}_\Theta(\beta, Q, \widehat{\theta}_k)$. Furthermore, the associated gradients and Hessians necessary to implement the Newton based search for $\widehat{\eta}$ and $\widehat{\beta}$ may be straightforwardly computed via the results of the following Lemmas.

Lemma 6.2: The gradients for $\mathcal{Q}_\Gamma(\eta, \widehat{R}(\eta))$ and $\mathcal{Q}_\Theta(\beta, \widehat{Q}(\beta))$ are given by

$$\begin{aligned} g_\Gamma(\eta) &\triangleq \frac{\partial \mathcal{Q}_\Gamma(\eta, \widehat{R}(\eta))}{\partial \eta} \\ &= 2\text{Re}\left\{\text{vec}\left\{\widehat{R}(\eta)^{-1}\left[\Gamma\Sigma - \Psi\right]\right\}\right\}, \\ g_\Theta(\beta) &\triangleq \frac{\partial \mathcal{Q}_\Theta(\beta, \widehat{Q}(\beta))}{\partial \beta} \\ &= 2\text{Re}\left\{\text{vec}\left\{\widehat{Q}(\beta)^{-1}\left[\Theta\Pi - \Omega\right]\right\} + \sum_{k=1}^{M}\begin{bmatrix}\text{vec}\left\{A_k^{-T}\right\} \\ \varnothing_{nm}\end{bmatrix}\right\}, \end{aligned}$$

respectively.

Proof: See [1] for details, which are not presented here due to space restrictions.7 ∎

Lemma 6.3: The Hessian matrices for $\mathcal{Q}_\Gamma(\eta, \widehat{R}(\eta))$ and $\mathcal{Q}_\Theta(\beta, \widehat{Q}(\beta))$ are given by

$$\begin{aligned} \frac{\partial^2 \mathcal{Q}_\Gamma(\eta, \widehat{R}(\eta))}{\partial\eta\partial\eta^T} &= 2\text{Re}\left\{\Sigma \otimes \widehat{R}(\eta)^{-1}\right\} \\ &\quad - M J_\Gamma^T(\eta)\left[\widehat{R}(\eta)^{-1} \otimes \widehat{R}(\eta)^{-T}\right] J_\Gamma^C(\eta) \\ \frac{\partial^2 \mathcal{Q}_\Theta(\beta, \widehat{Q}(\beta))}{\partial\beta\partial\beta^T} &= 2\text{Re}\left\{\Pi \otimes \widehat{Q}(\beta)^{-1}\right\} \\ &\quad - M J_\Theta^T(\beta)\left[\widehat{Q}(\beta)^{-1} \otimes \widehat{Q}(\beta)^{-T}\right] J_\Theta^C(\beta) \\ &\quad + \begin{bmatrix} 2\text{Re}\left(\sum_{k=1}^{M} A_k^{-1} \otimes A_k^{-T}\right) K & \varnothing_{n^2 \times nm} \\ \varnothing_{nm \times n^2} & \varnothing_{nm \times nm}\end{bmatrix} \end{aligned}$$

where

$$\begin{aligned} J_\Gamma(\eta) &\triangleq \frac{\partial \text{vec}\left\{\widehat{R}(\eta)\right\}}{\partial \eta^T} \\ &= \frac{1}{M}\left(\left[\Gamma\Sigma^T - \Psi^C\right] \otimes I_p + (I_p \otimes [\Gamma\Sigma - \Psi])\, K\right), \\ J_\Theta(\beta) &\triangleq \frac{\partial \text{vec}\left\{\widehat{Q}(\beta)\right\}}{\partial \beta^T} \\ &= \frac{1}{M}\left(\left[\Theta\Pi^T - \Omega^C\right] \otimes I_n + (I_n \otimes [\Theta\Pi - \Omega])\, K\right), \end{aligned}$$

and K is a commutation matrix (see p46 of [11]), which satisfies

$$K\text{vec}\{\cdot\} = \text{vec}\left\{\cdot^T\right\}.$$

Proof: See [1] for details, which are not presented here due to space restrictions. ∎

VII. Algorithm

The preceding arguments and derivations are summarised here via a formal specification of the EM-based algorithm proposed in this paper.

Given an initial guess $\widehat{\theta}_0$ and initialisation $k = 0$ perform the following steps.

1) **E-STEP**
   a) Compute $\mathbf{E}_{\widehat{\theta}_k}\{X_n \,|Y_n\}$ and $\mathbf{E}_{\widehat{\theta}_k}\{X_n X_n^\star \,|Y_n\}$ for $n = 1, \ldots, M$ via Lemma 5.2;
   b) Compute $\Phi, \Psi, \Sigma, \Lambda$ and Ω via Lemma 5.1;
2) **M-STEP**
   a) Determine a value $\eta = \widehat{\eta}_{k+1}$ that reduces $\mathcal{Q}_\Gamma(\eta, \widehat{R}(\eta), \widehat{\theta}_k)$ defined in (29) using the gradient and Hessian information from Lemmas 6.2 and 6.3 to implement the Newton search (33);
   b) Determine a value $\beta = \widehat{\beta}_{k+1}$ that reduces $\mathcal{Q}_\Theta(\beta, \widehat{Q}(\beta), \widehat{\theta}_k)$ defined in (30) using the gradient and Hessian information from Lemmas 6.2 and 6.3 to implement the Newton search (33);
   c) Determine $\widehat{R}(\widehat{\eta}_{k+1})$ and $\widehat{Q}(\widehat{\beta}_{k+1})$ from (27) and (28), respectively;
   d) Set

$$\widehat{\theta}_{k+1} = \left\{\widehat{\beta}_{k+1}, \widehat{\eta}_{k+1}, \text{vec}\left\{\widehat{Q}(\widehat{\beta}_{k+1})\right\}, \text{vec}\left\{\widehat{R}(\widehat{\eta}_{k+1})\right\}\right\}.$$

3) If converged, then stop. Otherwise set $k = k + 1$ and repeat.

VIII. Simulation

In this section we demonstrate the potential utility of the EM algorithm by profiling it against state-of-the-art and widely accepted gradient-based methods. To achieve this we consider a number of Monte-Carlo simulations with systems ranging from sixth-order, single-input single-output to 18th-order, three-input three-output. More precisely, three algorithms were considered, namely:

- The Expectation-Maximisation algorithm from section IV - this is denoted EM throughout;
- A standard implementation of the Levenberg-Marquardt algorithm for non-linear least squares , where the objective is to minimise a prediction error defined presently. This is denoted as LM throughout.
- Matlab's System Identification Toolbox [10] implementation of the Levenberg-Marquardt algorithm, where the objective is also to minimise the prediction error. The is denoted as SIT throughout.

The first example we consider is a single-input single-output (SISO) system, whose transfer function is given by

$$\begin{aligned} G(s) &= \frac{2.021 \times 10^{-3}}{(s+0.1)(s+0.2)} \\ &\quad \times \frac{1}{(s+0.01+0.1j)(s+0.01-0.1j)} \\ &\quad \times \frac{1}{(s+0.02+j)(s+0.02-j)}. \end{aligned} \tag{34}$$

This system was discretised using a zero-order-hold function with one-second sample time resulting in system $\bar{G}(z)$. Output measurements $\{Y_1, \ldots, Y_M\}$, where $M = 500$, were obtained via

$$\begin{aligned} Y_k &= \bar{G}(e^{j\omega_k}) + V_k, \quad \omega_k \triangleq 10^{-3+\frac{(k-1)(\log_{10}(\pi)+3)}{M-1}}, \\ V_k &\sim \mathcal{N}_c(0, 0.01), \quad k = 1, \ldots, M. \end{aligned} \tag{35}$$

We performed 100 simulations, where the noise and initial guess at parameter values were regenerated at the beginning of each run. In particular, the initial guess was randomly generated, but then ensured to correspond to a stable system.

To measure the success of each algorithm, we compute the prediction error of the final estimate $\widehat{\theta}$, denoted $e(\widehat{\theta})$, via

$$\begin{aligned} e(\widehat{\theta}) &\triangleq \text{Trace}\left(\frac{1}{M}\sum_{k=1}^{M}, E_k(\widehat{\theta})^\star E_k(\widehat{\theta})\right), \\ E_k(\widehat{\theta}) &\triangleq Y_k - \widehat{G}(e^{j\omega_k}), \\ \widehat{G}(e^{j\omega_k}) &\triangleq \widehat{C}(e^{j\omega_k} I - \widehat{A})^{-1}\widehat{B} + \widehat{D}, \end{aligned} \tag{36}$$

and compare it with the 2-norm of the sample variance

$$v \triangleq \text{Trace}\left(\frac{1}{M}\sum_{k=1}^{M} V_k^\star V_k\right). \tag{37}$$

From (35) we note that if $\widehat{\theta}$ is a good estimate then $e(\widehat{\theta}) \approx v$. With this in mind, a failure is determined if

$$e(\widehat{\theta}) > 1.3v. \tag{38}$$

Using this criteria, the number of failures for the first simulation are recorded in Table I under the epithet s_1. Note that both the EM and LM approaches proved robustly successful, despite all starting from initialisations which were 'failures' according the the criterion (38).

In consideration of these initial encouraging results, further Monte–Carlo simulations of increased system dimension were conducted as follows (as before n denotes number of states, m inputs and p outputs)

- $n = 2$, $m = 1$ and $p = 1$ - denoted s_2;
- $n = 3$, $m = 1$ and $p = 1$ - denoted s_3;
- $n = 8$, $m = 2$ and $p = 2$ - denoted s_4;
- $n = 18$, $m = 3$ and $p = 3$ - denoted s_5.

For each of the above Monte–Carlo simulations, 100 systems were randomly generated. Different noise realisations and randomly chosen initial guesses for the parameters were selected for each run. The number of failures are recorded in Table I under the columns labelled with epithets $s_2 - s_5$. For the lower order systems we see that EM performs slightly better than LM and both perform substantially better than SIT. As the system order is increased, the gap (in terms of number of failures) begins to grow between EM and LM indicating a clear performance advantage offered by the EM approach developed here for the case of higher dimension problems.

While the above simulations indicate the potential utility of the EM algorithm, it is of course interesting to assess performance on a real physical system. For this purpose, a 2×2 MIMO case arising from an aluminium cantilevered

	s_1	s_2	s_3	s_4	s_5
EM	0	2	4	8	13
LM	0	5	7	12	28
SIT	27	23	42	57	65

TABLE I

Number of failures for different algorithms (rows) and for different simulations (columns).

beam with piezoelectric actuators and sensors as shown in Figure 1 is considered.

Frequency domain measurements of the response of this apparatus at $M = 6337$ non-uniformly spaced frequency points between 5Hz and 500Hz were obtained by DFT analysis of experiments involving a periodic chirp signal excitation. For this experiment, the most natural model structure is continuous rather than discrete. However, this is handled by first applying a bilinear transform and then fitting a discrete model – see e.g. [14].

Figure 2 shows the measured frequency response for one of the transfer functions from this two-input, two-output system (the others are omitted for brevity). Figure 2 also shows the model fit for each algorithm and Figure 3 shows its associated error plot. In each case the system order was selected as $n = 12$, with $m = 2$ and $p = 2$. Furthermore, each algorithm was initialised with the same parameter values as determined using a frequency-domain subspace algorithm [14]. Note that the EM algorithm clearly outperforms the gradient-based algorithms in this case in that according to Figure 3, the modelling error is between 10 and 30 dB smaller when employing the EM approach. In particular, note from Figure 2 how the EM method is able to accurately model two lightly damped zeros between 40 and 80 rad/s that are completely missed by both gradient search based implementations considered here. Note that not capturing these modelling aspects can have significant implications for the success of subsequent feedback control design.

As a final test, again using data from a real physical system, but of higher state dimension and with estimation to be performed over a wider bandwidth, the modelling of a 132/66/11kV, 30MVA transformer was considered [2]. The frequency range used in the experiment was from 50Hz to 200kHz and $M = 123$ measurements were obtained. Figure 4 shows the measured data combined with two 31st order models. One is obtained via a frequency domain subspace algorithm [14], and the other is obtained via the EM algorithm in Section VII (starting from the subspace estimate). Again, despite the high model order, wide bandwidth over which a fit is required, and white dynamic range of measurements, a very close fit is achieved which is almost indistinguishable from the measurements.

IX. Conclusions

Still to be written.

Acknowledgements: The authors would like to thank Prof. Reza Moheimani and Dr. Andrew Fleming for providing the cantilevered beam experimental data and for their insights into aspects of its modelling.

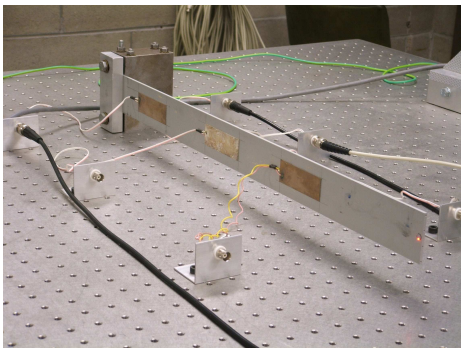

Fig. 1. Experimental apparatus.

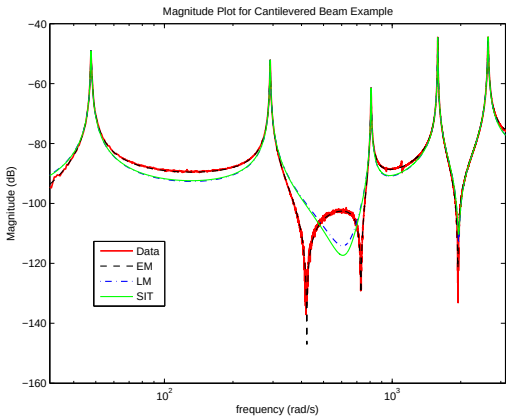

Fig. 2. Frequency response functions for cantilevered beam example.

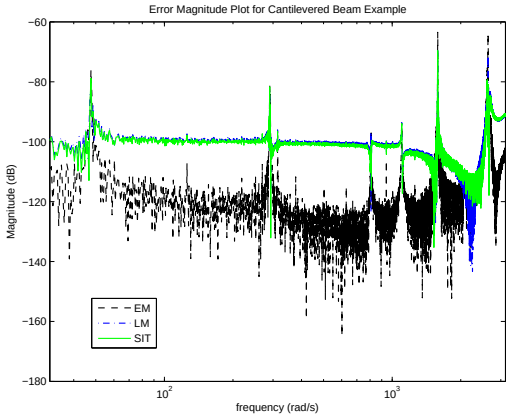

Fig. 3. Frequency response functions for cantilevered beam example.

References

[1] Brett Ninness Adrian Wills and Stuart Gibson. Robust maximum likelihood estimation from frequency domain data. *Preprint - Available at* `http://sigpromu.org/publications.html`, 2006.

[2] Hüseyin Akçay, Syed Islam, and Brett Ninness. Identification of power transformer models from frequency response data : A case study. *Signal Processing*, 68(3), 1998.

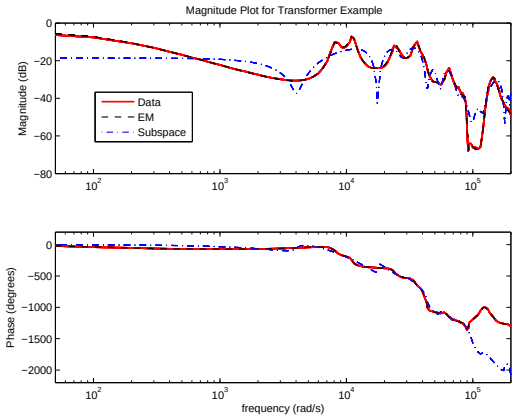

Fig. 4. Frequency response for transformer example.

[3] David R. Brillinger. *Time Series: Data Analysis and Theory*. Holden-Day, 1981.

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

[5] J.E. Dennis and Robert B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall, 1983.

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

[7] Stuart Gibson, Adrian Wills, and Brett Ninness. Maximum-likelihood parameter estimation of bilinear systems. *IEEE Trans. Automat. Control*, 50(10):1581–1596, 2005.

[8] Horn and Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

[9] J.N. Juang. *Applied System Identification*. Prentice Hall, 1994.

[10] Lennart Ljung. *MATLAB System Identification Toolbox Users Guide, Version 6*. The Mathworks, 2004.

[11] J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley, 1988.

[12] Tomas McKelvey. Frequency domain identification. In *Proceedings of SYSID 2000*, 2000. Santa Barbara, California.

[13] Tomas McKelvey. Frequency domain identification methods. *Circuits Systems Signal Process.*, 21(1):39–55, 2002. Special tutorial issue on system identification.

[14] Tomas McKelvey, Hüseyin Akçay, and Lennart Ljung. Subspace-based multivariable system identification from frequency response data. *IEEE Transactions on Automatic Control*, 41:960–979, 1996.

[15] R. Pintelon, P. Guillaume, Y. Rolain, J. Schoukens, and H. Van hamme. Parametric identification of transfer functions in the frequency domain—a survey. *IEEE Trans. Automat. Control*, 39(11):2245–2260, 1994.

[16] R. Pintelon and J. Schoukens. *System Identification: A Frequency Domain Approach*. IEEE Press, 2001.