PARAMETER ESTIMATION WITH MISSING DATA VIA EQUALIZATION-MAXIMIZATION

Petre Stoica

Dept. of Information Technology,
Uppsala University, P. O. Box 337,
SE-75105, Uppsala, Sweden.

Luzhou Xu, Jian Li

Dept. of Electrical and Computer Engineering,
P. O. Box 116130, University of Florida,
Gainesville, FL 32611, USA.

ABSTRACT

The expectation-maximization (EM) algorithm is often used in maximum likelihood (ML) estimation problems with missing data. However, EM can be rather slow to converge. In this paper, we introduce a new algorithm for parameter estimation problems with missing data , which we call Equalization-Maximization (EqM) (for reasons to be explained later). We derive the EqM algorithm in a general context and illustrate its use in the specific case of Gaussian autoregressive time series with a varying amount of missing observations. In the presented examples, EqM outperforms EM in terms of computational speed, at a comparable estimation performance.

1. INTRODUCTION AND PRELIMINARIES

Consider a parameter estimation problem in which $\boldsymbol{\gamma}$ denotes the $N_{\boldsymbol{\gamma}} \times 1$ vector of available data samples, and $\boldsymbol{\theta}$ the $n \times 1$ vector of unknown parameters. Let $\boldsymbol{\mu}$ be an $N_{\boldsymbol{\mu}} \times 1$ vector which is such that if it were available then solving the ML estimation problem based on $\boldsymbol{\gamma}$ *and* $\boldsymbol{\mu}$ would be relatively easy. To be more specific, let $p_{\boldsymbol{\gamma}}(\boldsymbol{\gamma};\boldsymbol{\theta})$ denote the probability density function (pdf) of the available data; when viewed as a function of $\boldsymbol{\theta}$, for given $\boldsymbol{\gamma}$, $p_{\boldsymbol{\gamma}}(\boldsymbol{\gamma};\boldsymbol{\theta})$ is the so-called likelihood function. Similarly, let $p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\boldsymbol{\mu};\boldsymbol{\theta})$ denote the joint pdf of $\boldsymbol{\gamma}$ and $\boldsymbol{\mu}$. Generally, $\boldsymbol{\mu}$ is assumed to have the property that solving the problem

$$\max_{\boldsymbol{\theta}} p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\boldsymbol{\mu};\boldsymbol{\theta}) \tag{1}$$

is much easier than maximizing $p_{\boldsymbol{\gamma}}(\boldsymbol{\gamma};\boldsymbol{\theta})$ to obtain the ML estimate of $\boldsymbol{\theta}$, i.e.,

$$\max_{\boldsymbol{\theta}} p_{\boldsymbol{\gamma}}(\boldsymbol{\gamma};\boldsymbol{\theta}). \tag{2}$$

This work was supported in part by the Swedish Science Council (VR) and the National Science Foundation Grant CCR-0104887. Please address all correspondence to: Dr. Jian Li, Department of Electrical and Computer Engineering, P. O. Box 116130, University of Florida, Gainesville, FL 32611, USA. Phone: (352) 392-2642. Fax: (352) 392-0044. E-mail: li@dsp.ufl.edu.

In particular, in several cases of interest the solution of (1) can be obtained in closed form, whereas the maximization of (2) would require usage of nonlinear programming algorithms. Note that, although this scenario is most relevant to missing data problems, $\boldsymbol{\mu}$ can be in principle any vector of variables with the above property, not necessarily a vector comprising physically missing samples of a data string.

The EM algorithm is an iterative solver of (2) that consists of an E-step and M-step (see., e.g., [1], [2]). In many cases (but not always) the E-step of EM is relatively easy to perform and the complexity of the M-step is comparable with that of solving the complete-data problem in (1). Furthermore, EM has the desirable property of increasing the available-data likelihood at each iteration. However, EM may converge rather slowly and, moreover, it may not converge to the global maximum of $p_{\boldsymbol{\gamma}}(\boldsymbol{\gamma};\boldsymbol{\theta})$.

In this paper, we introduce a different type of algorithm that appears to be faster than EM, at a comparable estimation performance, and which additionally does not require an E-step as in EM. In the next section, we present briefly the new algorithm along with two other less successful attempts to enhance the computational performance of EM. In Section 3, we illustrate the usage and performance of these algorithms in the specific case of autoregressive (AR) time series with missing observations.

2. CYCLIC-MAXIMIZATION (CM) AND EQUALIZATION-MAXIMIZATION (EQM)

2.1. CM

A conceptually simpler algorithm than EM, which is sometimes termed as cyclic-maximization (CM) or Pseudo-EM (PEM) (see, e.g., [3] [4]), consists of the following main steps:

- Given $\hat{\boldsymbol{\theta}}^0$, do for $k = 1, 2, \ldots$ until convergence the following steps:

- Obtain $\hat{\boldsymbol{\mu}}^k$ via

$$\max_{\boldsymbol{\mu}} p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\boldsymbol{\mu};\hat{\boldsymbol{\theta}}^{k-1}) \tag{3}$$

- Obtain $\hat{\boldsymbol{\theta}}^k$ via

$$\max_{\boldsymbol{\theta}} p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\hat{\boldsymbol{\mu}}^k;\boldsymbol{\theta}) \tag{4}$$

While CM typically converges faster than EM, the parameter estimates obtained with CM may be significantly less accurate than the ML estimate; in particular, the CM estimates may be heavily biased, unless the ratio $N_{\boldsymbol{\mu}}/N_{\boldsymbol{\gamma}}$ is rather small (see, e.g., [4], [5] and also the next section). The reason for the inferior accuracy of CM is not difficult to understand. By a well-known property of cyclic maximization, CM monotonically increases the complete-data likelihood function, $p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\boldsymbol{\mu};\boldsymbol{\theta})$, at each iteration, and hence it is a solver of the following problem:

$$\max_{\boldsymbol{\mu},\boldsymbol{\theta}} p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\boldsymbol{\mu};\boldsymbol{\theta}). \tag{5}$$

Because

$$\ln p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\boldsymbol{\mu};\boldsymbol{\theta}) = \ln p_{\boldsymbol{\mu}|\boldsymbol{\gamma}}(\boldsymbol{\mu}|\boldsymbol{\gamma};\boldsymbol{\theta}) + \ln p_{\boldsymbol{\gamma}}(\boldsymbol{\gamma};\boldsymbol{\theta}) \tag{6}$$

the maximization of the function in (5) with respect to $\boldsymbol{\mu}$, for $\boldsymbol{\theta}$ fixed, is equivalent to

$$\max_{\boldsymbol{\mu}} \ln p_{\boldsymbol{\mu}|\boldsymbol{\gamma}}(\boldsymbol{\mu}|\boldsymbol{\gamma};\boldsymbol{\theta}). \tag{7}$$

Let $\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})$ denote the solution of (7), and let

$$\Delta(\boldsymbol{\theta}) = \ln\left[\, p_{\boldsymbol{\mu}|\boldsymbol{\gamma}}(\boldsymbol{\mu}|\boldsymbol{\gamma};\boldsymbol{\theta})\,\right]|_{\boldsymbol{\mu}=\hat{\boldsymbol{\mu}}(\boldsymbol{\theta})}. \tag{8}$$

It follows from the previous discussion that the CM estimate of $\boldsymbol{\theta}$ is given by the solution to the problem:

$$\max_{\boldsymbol{\theta}} \ln p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\hat{\boldsymbol{\mu}}(\boldsymbol{\theta});\boldsymbol{\theta}) \Leftrightarrow \max_{\boldsymbol{\theta}} [\ln p_{\boldsymbol{\gamma}}(\boldsymbol{\gamma},\boldsymbol{\theta}) + \Delta(\boldsymbol{\theta})\,]. \tag{9}$$

As an example, if the conditional pdf $p_{\boldsymbol{\mu}|\boldsymbol{\gamma}}(\boldsymbol{\mu}|\boldsymbol{\gamma};\boldsymbol{\theta})$ is Gaussian, i.e.,

$$p_{\boldsymbol{\mu}|\boldsymbol{\gamma}}(\boldsymbol{\mu}|\boldsymbol{\gamma};\boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{N_{\boldsymbol{\mu}}}{2}}|\mathbf{C}(\boldsymbol{\theta})|^{\frac{1}{2}}} e^{-\frac{1}{2}[\boldsymbol{\mu}-\mathbf{m}(\boldsymbol{\theta})]^T\mathbf{C}^{-1}(\boldsymbol{\theta})[\boldsymbol{\mu}-\mathbf{m}(\boldsymbol{\theta})]} \tag{10}$$

where $\mathbf{m}(\boldsymbol{\theta})$ is the conditional mean and $\mathbf{C}(\boldsymbol{\theta})$ is the conditional covariance of $\boldsymbol{\mu}$ given $\boldsymbol{\gamma}$, then we have:

$$\hat{\boldsymbol{\mu}}(\boldsymbol{\theta}) = \mathbf{m}(\boldsymbol{\theta}) \tag{11}$$

and (to within an additive constant)

$$\Delta(\boldsymbol{\theta}) = -\frac{1}{2}\ln|\mathbf{C}(\boldsymbol{\theta})|. \tag{12}$$

The presence of the second term in (9) clearly shows that in general the CM estimate of $\boldsymbol{\theta}$ is not a maximizer of $p_{\boldsymbol{\gamma}}(\boldsymbol{\gamma};\boldsymbol{\theta})$. It is precisely this term, which typically is independent of the available data (see, e.g., (12)), that makes CM be significantly less accurate than ML (in particular, the CM estimate may be heavily biased), unless $N_{\boldsymbol{\mu}}/N_{\boldsymbol{\gamma}}$ is "small" (in which case the second term in (9) becomes negligible). Hence, despite the fact that CM is both conceptually and computationally much simpler than EM, its usage can be recommended only if the ratio $N_{\boldsymbol{\mu}}/N_{\boldsymbol{\gamma}}$ is fairly small.

Following the previous discussion on CM, a natural question is whether it would be possible to enhance the estimation performance of CM without sacrificing its conceptional and computational simplicity. It follows from (9) that the problem (2) can be reformulated as:

$$\max_{\boldsymbol{\theta},\boldsymbol{\mu}} [\ln p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\boldsymbol{\mu};\boldsymbol{\theta}) - \Delta(\boldsymbol{\theta})] \tag{13}$$

For fixed $\boldsymbol{\theta}$, the maximization of the above function with respect to $\boldsymbol{\mu}$ has exactly the same solution as in the corresponding step of CM. However, for given $\boldsymbol{\mu}$, the maximization of (13) with respect to $\boldsymbol{\theta}$ is much more complicated than the maximization in the last step (4) of CM, owing to the second term in (13); in particular, the maximization of (13) with respect to $\boldsymbol{\theta}$, for fixed $\boldsymbol{\mu}$, usually does not have a closed-form solution. Consequently, the cyclic maximization of (13) with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\theta}$ is computationally more intensive than CM, and thus it is not the modification of CM we are seeking. A more appealing modification of CM, which leads to the EqM algorithm, is derived next.

2.2. EqM

The estimate of $\boldsymbol{\mu}$ used in the CM algorithm, for fixed $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{k-1}$, is given by $\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\theta}}^{k-1})$, which is the mode of the conditional pdf $p_{\boldsymbol{\mu}|\boldsymbol{\gamma}}(\boldsymbol{\mu}|\boldsymbol{\gamma};\hat{\boldsymbol{\theta}}^{k-1})$ (see (7)). This is a very sensible estimate, and hence it might seem difficult to find a better choice of $\boldsymbol{\mu}$, for given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{k-1}$. However, we can see from (6) that by choosing $\boldsymbol{\mu}$ as a function $\boldsymbol{\theta}$, let us say $\boldsymbol{\mu} = \mathbf{h}(\boldsymbol{\theta})$, such that

$$p_{\boldsymbol{\mu}|\boldsymbol{\gamma}}(\boldsymbol{\mu}|\boldsymbol{\gamma};\boldsymbol{\theta})|_{\boldsymbol{\mu}=\mathbf{h}(\boldsymbol{\theta})} = \text{const.}\,, \tag{14}$$

we get

$$\ln\, p_{\boldsymbol{\gamma}}(\boldsymbol{\gamma};\boldsymbol{\theta}) = \text{const.} + \ln\,[p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\boldsymbol{\mu};\boldsymbol{\theta})]|_{\boldsymbol{\mu}=\mathbf{h}(\boldsymbol{\theta})}. \tag{15}$$

The choice of the function $\mathbf{h}(\boldsymbol{\theta})$, which is not unique, will be discussed shortly. Using (15) we can reformulate the ML estimation problem in (2) as:

$$\max_{\boldsymbol{\theta}} p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma},\mathbf{h}(\boldsymbol{\theta});\boldsymbol{\theta}). \tag{16}$$

In general, this maximization problem is not easier to solve than the original ML problem (2); in particular, usually (16) does not admit a closed-form solution. However,

we can use the following fairly natural iterative algorithm, which we call EqM (for reasons explained below), to approximate the solution of (16):

- Given $\hat{\boldsymbol{\theta}}^0$, do for $k = 1, 2, \ldots$ until convergence the following steps:
- **Eq-step**: Obtain $\hat{\boldsymbol{\mu}}^k$ via

$$\hat{\boldsymbol{\mu}}^k = \mathbf{h}(\hat{\boldsymbol{\theta}}^{k-1}) \tag{17}$$

- **M-step**: Obtain $\hat{\boldsymbol{\theta}}^k$ via

$$\max_{\boldsymbol{\theta}} p_{\boldsymbol{\gamma},\boldsymbol{\mu}}(\boldsymbol{\gamma}, \hat{\boldsymbol{\mu}}^k; \boldsymbol{\theta}). \tag{18}$$

Setting $\boldsymbol{\mu} = \mathbf{h}(\boldsymbol{\theta})$ in (6) *equalizes* the values of the conditional pdf $p_{\boldsymbol{\mu}|\boldsymbol{\gamma}}(\boldsymbol{\mu}|\boldsymbol{\gamma}; \boldsymbol{\theta})$ corresponding to different values of $\boldsymbol{\theta}$, and thus the name of the corresponding step of the above algorithm.

As an example, consider again the Gaussian conditional pdf in (13). Let

$$\boldsymbol{\mu} = \mathbf{h}(\boldsymbol{\theta}) = \mathbf{m}(\boldsymbol{\theta}) + \left[\frac{\ln\left(\frac{r}{|\mathbf{C}(\boldsymbol{\theta})|}\right)}{\mathbf{v}^T(\boldsymbol{\theta})\mathbf{C}^{-1}(\boldsymbol{\theta})\mathbf{v}(\boldsymbol{\theta})}\right]^{\frac{1}{2}} \mathbf{v}(\boldsymbol{\theta}) \tag{19}$$

where $\mathbf{v}(\boldsymbol{\theta})$ is an $N_{\boldsymbol{\mu}} \times 1$ vector (which possibly depends on $\boldsymbol{\theta}$), and r is a constant which satisfies:

$$r \geq |\mathbf{C}(\boldsymbol{\theta})|. \tag{20}$$

Then, a simple calculation shows that, for (19),

$$p_{\boldsymbol{\mu}|\boldsymbol{\gamma}}(\boldsymbol{\mu}|\boldsymbol{\gamma}; \boldsymbol{\theta})|_{\boldsymbol{\mu}=\mathbf{h}(\boldsymbol{\theta})} = \frac{1}{(2\pi)^{N_{\boldsymbol{\mu}}/2} r^{1/2}} = \text{const.} \tag{21}$$

To satisfy the condition in (20) on r we can somewhat arbitrarily choose:

$$r = |\mathbf{C}(\hat{\boldsymbol{\theta}}^0)| \tag{22}$$

(or slightly larger). The accuracy of EqM appears to depend on r in a mild way. However, in general, a smaller value of r is likely to have a beneficial effect on the performance of EqM. Consequently, we recommend setting r to as small a value as possible, such as in (22). Of course, choosing r in this way we may be at risk of violating the condition in (20) at some iteration of EqM. However, this is not a serious problem; if (20) does not hold at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^k$, we can increase r as necessary (e.g., to $r = |\mathbf{C}(\hat{\boldsymbol{\theta}}^k)|$ or slightly larger), and continue the iterative process with the new value of r; doing so is acceptable as long as EqM converges with a fixed value of r.

Regarding the choice of $\mathbf{v}(\boldsymbol{\theta})$ in (19), the accuracy of EqM appears to depend on $\mathbf{v}(\boldsymbol{\theta})$ in a relatively complicated manner. In the next section we show empirically that the estimation errors associated with EqM can be kept reasonably small if we set:

$$\mathbf{v}(\boldsymbol{\theta}) = \mathbf{C}_{1\cdot}(\boldsymbol{\theta}) \tag{23}$$

where $\mathbf{C}_{1\cdot}(\boldsymbol{\theta})$ denotes the first column of $\mathbf{C}(\boldsymbol{\theta})$. The function $\mathbf{h}(\boldsymbol{\theta})$, (19), corresponding to (23) is given by

$$\mathbf{h}(\boldsymbol{\theta}) = \mathbf{m}(\boldsymbol{\theta}) + \left[\frac{\ln\left(\frac{r}{|\mathbf{C}(\boldsymbol{\theta})|}\right)}{\mathbf{C}_{11}(\boldsymbol{\theta})}\right]^{\frac{1}{2}} \mathbf{C}_{1\cdot}(\boldsymbol{\theta}) \tag{24}$$

where $\mathbf{C}_{11}(\boldsymbol{\theta})$ denotes the $(1, 1)$ element of $\mathbf{C}(\boldsymbol{\theta})$ note, as a small bonus, that the computation of the inverse matrix $\mathbf{C}^{-1}(\boldsymbol{\theta})$ is not required in (24).

We should note that in the numerical examples of the next section we have also tested much larger values of r than that given by (22), as well as randomly generated $\mathbf{v}(\boldsymbol{\theta})$ vectors instead of (23), and have found that the performance of EqM was almost unchanged.

Next, we remark on the fact that the M-step of EqM is identical computationally to the last step, (4), of the CM, whereas the Eq-step of EqM is only slightly more involved than step (6) of CM. Consequently, the computational burdens per iteration associated with CM and EqM are quite similar to one another. Regarding the convergence speed, the empirical experience we have accumulated so far suggests that EqM, like CM, converges in a small number of iterations. Hence, like CM, EqM is usually faster than EM.

We have shown in [6] that EqM generally does not maximize the function in (16). However, the difference between EqM and ML can be made small by choosing the function $\mathbf{h}(\boldsymbol{\theta})$ appropriately. In the next section, we show that for the scenario considered there the choice of $\mathbf{h}(\boldsymbol{\theta})$ in (24) is satisfactory, in the sense that the corresponding accuracy of EqM is close to the ultimate accuracy associated with EM.

3. NUMERICAL ILLUSTRATIONS AND CONCLUDING REMARKS

Consider an AR time series, $\{y(t)\}_{t=1,2,\ldots}$, generated by the equation

$$y(t) + a_1 y(t-1) + \ldots + a_n y(t-n) = \varepsilon(t) \tag{25}$$

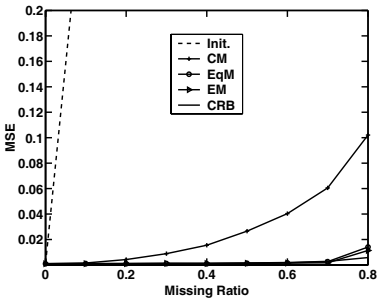
where $\{\varepsilon(t)\}_{t=1,2,\ldots}$ is a sequence of i.i.d. Gaussian random variables with mean zero and variance σ^2, and the coefficients $\{a_k\}$ are such that the polynomial $z^n + a_1 z^{n-1} + \ldots + a_n$ has all its zeros strictly inside the unit circle. We generate N observations with (25), out of which we randomly omit $N_{\boldsymbol{\mu}}$ (the locations of the omitted samples are uniformly distributed on $[1, N]$). The remaining $N_{\boldsymbol{\gamma}} = N - N_{\boldsymbol{\mu}}$ observations are to be used for estimating the parameters in (25); we will focus on the estimation of $\{a_k\}$ in what follows.

We consider the following second-order AR time series:

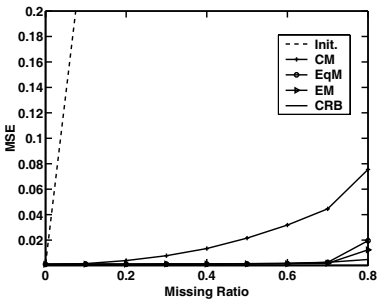
$$n = 2, \quad a_1 = -1.5, \quad a_2 = 0.7, \quad \sigma^2 = 1, \qquad (26)$$

and fix $N = 500$, but we vary the number of missing observations, $N_{\boldsymbol{\mu}}$, such that $N_{\boldsymbol{\mu}}/N$ takes on values in the interval $[0,\ 0.8]$. To measure the quality of an estimate we use MSE figures, which we plot versus the ratio $N_{\boldsymbol{\mu}}/N$. We estimate the MSE values by means of 1000 standard Monte-Carlo simulations, across which we vary both the noise sequence $\{\varepsilon(t)\}$ and the positions of the $N_{\boldsymbol{\mu}}$ missing observations.

We use EM, CM and EqM methods to estimate $\{a_k\}$ (see [6] for more details). As a comparistion, we also show the performance of the initial estimation where we set $\boldsymbol{\mu} = 0$ and estimate $\{a_k\}$ via the method of least-squares (LS). The so-obtained $\{\hat{a}_k\}$ are used as initial estimates in all the other (iterative) algorithms.



(a) $\hat{a}_1$



(b) $\hat{a}_2$

Fig. 1. MSE of $\{\hat{a}_1\ \hat{a}_2\}$, vs. $N_{\boldsymbol{\gamma}}/N$.

In Fig. 1 we show the MSE of the estimates $\{\hat{a}_1,\ \hat{a}_2\}$ obtained with the previous methods, along with the Cramér-Rao bound (CRB). In Fig. 2 we display the average number of flops per run for our Matlab implementations of the above estimation algorithms.

Based on our admittedly limited experience with the above parameter estimation methods we submit the following facts:

- The initial estimation method and CM are fast but, as $N_{\boldsymbol{\mu}}/N$ increases, both of them become increasingly biased, which leads to unacceptably large MSE values.
- EM yields accurate estimates whose MSE follows the CRB for much larger values of $N_{\boldsymbol{\mu}}/N$ (up to $N_{\boldsymbol{\mu}}/N = 0.7$ in the reported example). However, EM is more intensive computationally than CM.
- EqM is faster than EM, at a comparable estimation performance. In fact, it appears that EqM offers the EM's statistical performance at the CM's computational cost.

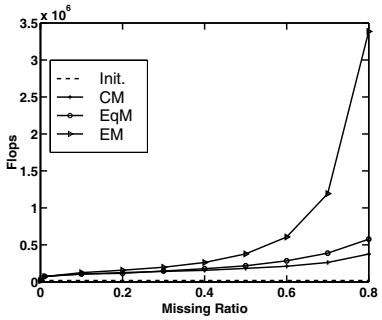


Fig. 2. Average number of flops per run, vs. $N_{\boldsymbol{\gamma}}/N$.

4. REFERENCES

[1] 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 B*, vol. 39, pp. 1–18, 1977.

[2] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 2002.

[3] J. D. Sargan and E. G. Drettakis, "Missing data in an autoregressive model," *International Economic Review*, vol. 15, pp. 39–58, February 1974.

[4] R. B. Miller and O. Ferreiro, "A strategy to complete a time series with missing observations," *In E. Parzen, Ed., Time Series Analysis of Irregularly Observed Data, College Station, TX*, pp. 251–275, 1984.

[5] R. Wallin, A. J. Isaksson, and L. Ljung, "An iterative method for identification of ARX models from incomplete data," *Proceedings of the 39th IEEE conference on Decision and Control. Sydney, Australia*, December 2000.

[6] P. Stoica, L. Xu, and J. Li, "A new type of parameter estimation algorithm for missing data problems," submitted to *Statistics and Probability Letters*, September 2004.