MISSING DATA RECOVERY VIA A NONPARAMETRIC ITERATIVE ADAPTIVE APPROACH

Petre Stoica[†], *Jian Li*[‡], *Jun Ling*[‡] *and Yubo Cheng*[‡]

[†]Dept. of Information Technology, Uppsala University, Uppsala, Sweden
[‡]Dept. of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA

ABSTRACT

We introduce a missing data recovery methodology based on a weighted least squares iterative adaptive approach (IAA). The proposed method is referred to as the missing-data IAA (MIAA) and it can be used for uniform or non-uniform sampling as well as for arbitrary data missing patterns. MIAA uses the IAA spectrum estimates to retrieve the missing data, based on a spectral least squares criterion similar to that used by IAA. Numerical examples are presented to show the effectiveness of MIAA for missing data recovery. We also show that MIAA can outperform an existing competitive approach, and this at a much lower computational cost.

Index Terms— Missing Data Recovery, Spectral Estimation, Iterative Adaptive Approach, Weighted Least Squares

1. INTRODUCTION

Missing data problems occur in a wide range of applications (see, e.g., [1, 2, 3, 4, 5, 6]) and several studies have been carried out to investigate these problems (see, e.g., [5, 6, 7] and the references therein). For gapped-data sequences, where the missing samples appear clustered in groups, a competitive approach is the gapped-data amplitude and phase estimation (GAPES) algorithm [3]. GAPES works well for gapped data but not so well for arbitrary missing data patterns. The missing-data amplitude and phase estimation (MAPES) approach was developed using the expectation maximization algorithm [8]. MAPES works well even in the case where the missing samples occur at arbitrary positions of a uniform sampling grid. However, the performance of MAPES degrades relatively quickly when the percentage of missing samples increases. Moreover, the computational complexity of MAPES is rather high.

A central issue in the missing data recovery approaches mentioned above, as well as for the approach of this paper, is the spectral estimation from the available samples [9]. Recently, a non-parametric and user parameter free weighted least squares based iterative adaptive approach (IAA) was proposed for spatial spectral estimation (a.k.a. array processing) [10]. IAA can be used with few or even a single snapshot and with arbitrary array geometries. In this paper, we show how IAA can be extended to deal with the missing temporal data recovery problem; the IAA-based missing data recovery approach is referred to as MIAA. First MIAA uses IAA to obtain an accurate spectral estimate from the given samples. Then a similar spectral least squares criterion to that employed by IAA is used with the IAA spectrum estimate to recover the missing samples. MIAA works for arbitrary data missing patterns as well as for uniform or non-uniform sampling. Moreover, MIAA can be used for both interpolation and extrapolation of data sequences.

2. PROBLEM FORMULATION

Let

$$\mathbf{y}_g = \begin{bmatrix} y_{t_1} \\ \vdots \\ y_{t_N} \end{bmatrix}, \quad \mathbf{y}_m = \begin{bmatrix} y_{\bar{t}_1} \\ \vdots \\ y_{\bar{t}_{\bar{N}}} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_g \\ \mathbf{y}_m \end{bmatrix}, \quad (1)$$

where $\mathbf{y}_g$ is the vector of available (or given) samples, $\mathbf{y}_m$ is the vector of missing (or desired) samples, $\{t_n\}$ denote the sampling times of available samples, and $\{\bar{t}_{\bar{n}}\}$ are the sampling times of the missing samples. Usually $\{\bar{t}_{\bar{n}}\}$ are interleaved with $\{t_n\}$, so that the missing data recovery is basically an interpolation problem. However, we do not need to make this assumption as MIAA can be used for both interpolation and extrapolation problems. The problem of interest is to recover (or, rather, estimate) $\mathbf{y}_m$ from $\mathbf{y}_g$.

3. MIAA

The available and missing samples are assumed to be complex-valued. The modification of MIAA to deal with real-valued missing data recovery problems is relatively straightforward and is not discussed herein.

3.1. Using IAA for Spectral Estimation from Given Data

MIAA first uses IAA [10] for spectral estimation from the given data vector $\mathbf{y}_g$. Let K denote the number of grid points

This work was supported in part by the Swedish Research Council (VR), the European Research Council (ERC), and the National Science Foundation under Grant No. CCF-0634786 and ECCS-0729727.

in the frequency domain and let $\{\omega_k\}_{k=1}^K$ be the corresponding frequencies; usually K is chosen to be quite large and the frequency grid of interest $\{\omega_k\}$ is uniform. Let

$$\mathbf{a}_g(\omega_k) = \begin{bmatrix} e^{j\omega_k t_1} \\ \vdots \\ e^{j\omega_k t_N} \end{bmatrix}, \mathbf{a}_m(\omega_k) = \begin{bmatrix} e^{j\omega_k \bar{t}_1} \\ \vdots \\ e^{j\omega_k \bar{t}_{\bar{N}}} \end{bmatrix},$$

$$\mathbf{a}(\omega_k) = \begin{bmatrix} \mathbf{a}_g(\omega_k) \\ \mathbf{a}_m(\omega_k) \end{bmatrix}, \mathbf{A} = [\mathbf{a}(\omega_1) \cdots \mathbf{a}(\omega_K)], \quad (2)$$

where $\mathbf{a}_g(\omega_k)$ and $\mathbf{a}_m(\omega_k)$ are the Fourier vectors corresponding to $\mathbf{y}_g$ and to $\mathbf{y}_m$ at frequency ω_k. The data vector $\mathbf{y}$ can be modeled as:

$$\mathbf{y} = \mathbf{A}\boldsymbol{\alpha}, \quad (3)$$

where $\boldsymbol{\alpha} = [\alpha(\omega_1) \ldots \alpha(\omega_K)]^T$ is the complex-valued spectral amplitude vector for the chosen frequency grid $\{\omega_k\}$, and $(\cdot)^T$ denotes the transpose.

A possible noise term of the data vector is not modeled explicitly but implicitly via its contribution to the amplitude vector $\boldsymbol{\alpha}$. We also note that in many applications, the vector $\boldsymbol{\alpha}$ in (3) is "almost sparse" in the sense that, while few elements of it (if any) are equal to zero, many elements are quite small and only a small number of elements have significant magnitudes. This type of application suits IAA best (see, e.g., [10]), and it will be the one that we will focus on in this paper.

Let

$$P_k = |\alpha(\omega_k)|^2, \quad k = 1, \cdots, K, \quad (4)$$

denote the data power at frequency ω_k. For each frequency ω_k in the available data, the interference covariance matrix can be defined as:

$$\mathbf{Q}_g(\omega_k) = \mathbf{R}_g - P_k \mathbf{a}_g(\omega_k)\mathbf{a}_g^H(\omega_k), \quad (5)$$

where $(\cdot)^H$ denotes the conjugate transpose and $\mathbf{R}_g$ is the covariance matrix of the given data, i.e.,

$$\mathbf{R}_g = \sum_{k=1}^K P_k \mathbf{a}_g(\omega_k)\mathbf{a}_g^H(\omega_k). \quad (6)$$

By using a weighted least squares criterion, the spectral estimation problem at frequency ω_k can be formulated as:

$$\min_{\alpha(\omega_k)} [\mathbf{y}_g - \alpha(\omega_k)\mathbf{a}_g(\omega_k)]^H \mathbf{Q}_g^{-1}(\omega_k) [\mathbf{y}_g - \alpha(\omega_k)\mathbf{a}_g(\omega_k)]. \quad (7)$$

The solution to this optimization problem is given by:

$$\widehat{\alpha}(\omega_k) = \frac{\mathbf{a}_g^H(\omega_k)\mathbf{Q}_g^{-1}(\omega_k)\mathbf{y}_g}{\mathbf{a}_g^H(\omega_k)\mathbf{Q}_g^{-1}(\omega_k)\mathbf{a}_g(\omega_k)}. \quad (8)$$

By using the matrix inversion lemma (see, e.g., [9]), it can be shown that:

$$\mathbf{a}_g^H(\omega_k)\mathbf{Q}_g^{-1} = \frac{\mathbf{a}_g^H(\omega_k)\mathbf{R}_g^{-1}}{1 - P_k \mathbf{a}_g^H(\omega_k)\mathbf{R}_g^{-1}\mathbf{a}_g(\omega_k)}. \quad (9)$$

Therefore, we can replace $\mathbf{Q}_g^{-1}(\omega_k)$ in (8) by $\mathbf{R}_g^{-1}$ to avoid computing $\mathbf{Q}_g^{-1}(\omega_k)$ for each ω_k:

$$\widehat{\alpha}(\omega_k) = \frac{\mathbf{a}_g^H(\omega_k)\mathbf{R}_g^{-1}\mathbf{y}_g}{\mathbf{a}_g^H(\omega_k)\mathbf{R}_g^{-1}\mathbf{a}_g(\omega_k)}. \quad (10)$$

Note that $\mathbf{R}_g$ depends on the very spectral amplitudes $\{\alpha(\omega_k)\}_{k=1}^K$ that we want to estimate. Thus (10) must be implemented in an iterative manner. The initialization of IAA can be done by setting $\mathbf{R}_g$ in (10) to the identity matrix $\mathbf{I}$, which gives the least squares (LS) estimate of $\alpha(\omega_k)$.

3.2. Missing Data Recovery

From the estimated amplitude spectrum $\{\widehat{\alpha}(\omega_k)\}$, we can recover the missing data by using a least squares criterion somewhat similar to that used by IAA. Depending on the way in which this spectral fitting least squares criterion is defined, we obtain two different versions of MIAA, as explained below.

3.2.1. MIAA-1

This version estimates the missing data by fitting the IAA-like spectral estimates that would be obtained from $\mathbf{y}_m$, if it were available, to $\{\widehat{\alpha}(\omega_k)\}$ given by IAA (see (10)):

$$\min_{\mathbf{y}_m} \sum_{k=1}^K \left| \frac{\mathbf{a}_m^H(\omega_k)\mathbf{R}_m^{-1}\mathbf{y}_m}{\mathbf{a}_m^H(\omega_k)\mathbf{R}_m^{-1}\mathbf{a}_m(\omega_k)} - \widehat{\alpha}(\omega_k) \right|^2, \quad (11)$$

where $\mathbf{R}_m$ is similarly defined to $\mathbf{R}_g$ in (6) but with P_k replaced by $|\widehat{\alpha}(\omega_k)|^2$ and $\mathbf{a}_g(\omega_k)$ replaced by $\mathbf{a}_m(\omega_k), k = 1, \cdots, K$. Let

$$\mathbf{h}_k^H = \frac{\mathbf{a}_m^H(\omega_k)\mathbf{R}_m^{-1}}{\mathbf{a}_m^H(\omega_k)\mathbf{R}_m^{-1}\mathbf{a}_m(\omega_k)}. \quad (12)$$

The solution to (11) is straightforward:

$$\widehat{\mathbf{y}}_m = \left(\sum_{k=1}^K \mathbf{h}_k\mathbf{h}_k^H\right)^{-1} \left[\sum_{k=1}^K \widehat{\alpha}(\omega_k)\mathbf{h}_k\right]. \quad (13)$$

3.2.2. MIAA-2

MIAA-1 fits the spectral estimate that would be obtained from the $\bar{N}$ missing data samples, if they were available, to the IAA spectral estimate obtained from the N available data samples. However, without increasing the computational burden too much, we could fit the spectral estimate associated with both missing and available data to the spectrum estimated from the available data, which should presumably improve the missing data estimation accuracy, especially when $\bar{N} \ll N$. To pursue this idea, define the covariance matrix estimate for both the missing and available data as:

$$\mathbf{R} = \sum_{k=1}^K \widehat{P}_k \mathbf{a}(\omega_k)\mathbf{a}^H(\omega_k), \quad (14)$$

where $\widehat{P}_k$ is obtained from the IAA estimate in (10). Then $\mathbf{y}_m$ is estimated as the solution to the following spectral fitting LS problem:

$$\min_{\mathbf{y}_m} \sum_{k=1}^{K} \left| \frac{\mathbf{a}^H(\omega_k)\mathbf{R}^{-1}\mathbf{y}}{\mathbf{a}^H(\omega_k)\mathbf{R}^{-1}\mathbf{a}(\omega_k)} - \widehat{\alpha}(\omega_k) \right|^2. \qquad (15)$$

Let

$$\frac{\mathbf{a}^H(\omega_k)\mathbf{R}^{-1}}{\mathbf{a}^H(\omega_k)\mathbf{R}^{-1}\mathbf{a}(\omega_k)} \triangleq [\mathbf{h}_{g_k}^H \quad \mathbf{h}_{m_k}^H], \qquad (16)$$

where the dimensions of $\mathbf{h}_{g_k}$ and $\mathbf{h}_{m_k}$ conform with those of $\mathbf{y}_g$ and $\mathbf{y}_m$. Using this notation, the objective function in (15) can be rewritten as:

$$\min_{\mathbf{y}_m} \sum_{k=1}^{K} \left| \mathbf{h}_{m_k}^H \mathbf{y}_m + [\mathbf{h}_{g_k}^H \mathbf{y}_g - \widehat{\alpha}(\omega_k)] \right|^2. \qquad (17)$$

The minimization of (17) with respect to $\mathbf{y}_m$ gives:

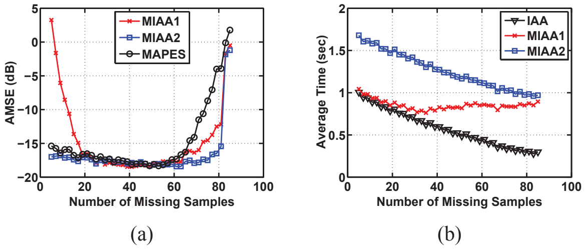
$$\widehat{\mathbf{y}}_m = \left(\sum_{k=1}^{K} \mathbf{h}_{m_k} \mathbf{h}_{m_k}^H \right)^{-1} \sum_{k=1}^{K} \mathbf{h}_{m_k} [\widehat{\alpha}(\omega_k) - \mathbf{h}_{g_k}^H \mathbf{y}_g]. \quad (18)$$

4. NUMERICAL EXAMPLES

This section illustrates the effectiveness of MIAA as a missing data recovery algorithm and compares the performance of MIAA and MAPES (more exactly the MAPES-EM1 algorithm in [7] [8]) for arbitrary missing data patterns.

The signal we consider consists of 4 complex-valued sinusoids located at $f_1 = 0.05$ Hz, $f_2 = 0.065$ Hz, $f_3 = 0.27$ Hz, $f_4 = 0.28$ Hz with complex amplitudes $\alpha_1 = \alpha_2 = \alpha_3 = 1$, $\alpha_4 = 0.5$. The data is corrupted by a zero-mean circularly symmetric complex Gaussian white noise with variance $\sigma_n^2 = 0.01$. The complete data sequence (uniformly sampled at a rate of 1 sample per sec) has 100 samples (i.e., $N + \bar{N} = 100$ and therefore the percentage of missing samples is $\bar{N}\%$). IAA, implemented by setting $K = 1000$ and the iteration number to 15, is used to obtain a spectrum estimate for MIAA. The locations of the missing samples are generated randomly in the interval $[1, 100]$, consequently, our numerical examples involve both interpolation and extrapolation problems.

We let $\bar{N}$ vary from 5 to 85. A total of 50 Monte-Carlo trials are performed. The noise, initial sinusoidal phases and data missing patterns vary independently from one trial to another. Define the average mean-squared error (AMSE) of the missing data estimates as $\text{AMSE}(\bar{N}) = \frac{1}{\bar{N}} E\left(||\mathbf{y}_m - \hat{\mathbf{y}}_m||^2\right)$, where $||\cdot||$ denotes the Euclidean norm and $E(\cdot)$ represents the average over the Monte-Carlo runs. The AMSE results are presented in Fig. 1(a). When $\bar{N}$ is small, such as $\bar{N} < 20$, MIAA-2 outperforms MIAA-1 significantly. This observation is in line with the remark made in the previous section on the way in which the fitting criteria of MIAA-1 and of MIAA-2 are defined. Both MIAA methods work well for $20 \leq \bar{N} \leq 60$. For $\bar{N}$ between 60 and 80, MIAA-1 degrades more quickly than MIAA-2. The performance of both methods degrades significantly for $\bar{N} > 80$, as in such a case $N = 100 - \bar{N} < 20$ is too small for IAA to provide an accurate spectrum estimate. Regarding the comparison with MAPES, the performance of MAPES is worse than that of MIAA-2, especially for $\bar{N} \geq 70$.



(a) (b)

Fig. 1. The performance comparison between MIAA-1, MIAA-2 and MAPES. (a) AMSE of estimated missing data vs. $\bar{N}$ (note that the noise floor is -20 dB), and (b) average computation time.

Fig. 1(b) compares the average computer time required by the two MIAA methods. Our simulations show that the average time required by MAPES varies from 250 to 320 sec, which is more than 100 times larger than the time required by the MIAA methods. Consequently, the average time for MAPES is not plotted in Fig. 1(b). We have observed that the time needed by the missing data recovery step of MIAA-1 (excluding the time needed by IAA) increases as $\bar{N}$ increases, while that needed by MIAA-2 is almost constant. This observation, together with the fact that the computation time needed by IAA decreases with $\bar{N}$, explains why the overall difference between the times needed by the two MIAA methods diminishes as $\bar{N}$ increases, as shown in Fig. 1(b). Note that MIAA-2 is slightly more expensive computationally than MIAA-1, but the (small) extra computation time required by MIAA-2 is completely justified by its better performance.

The missing samples and the available data, the estimated amplitude spectra, and the estimated missing samples for one Monte-Carlo trial are presented in Figs. 2-4 for $\bar{N} = 80$. This example shows both interpolation and extrapolation of data sequence. Note, once again, that the MIAA methods, especially MIAA-2, outperform MAPES significantly.

5. CONCLUSIONS

We have presented a new algorithm, referred to as MIAA, for missing data recovery. MIAA first uses IAA for spectral estimation from the available data samples and then estimates the missing data samples by employing the IAA spectral estimate in a least squares spectral fitting criterion. Numerical examples have been presented to show that MIAA is an effective approach that can be used for arbitrary missing data

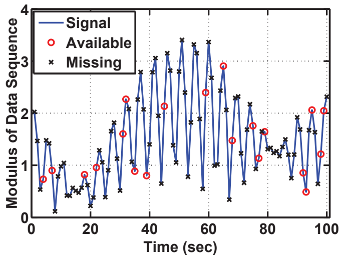
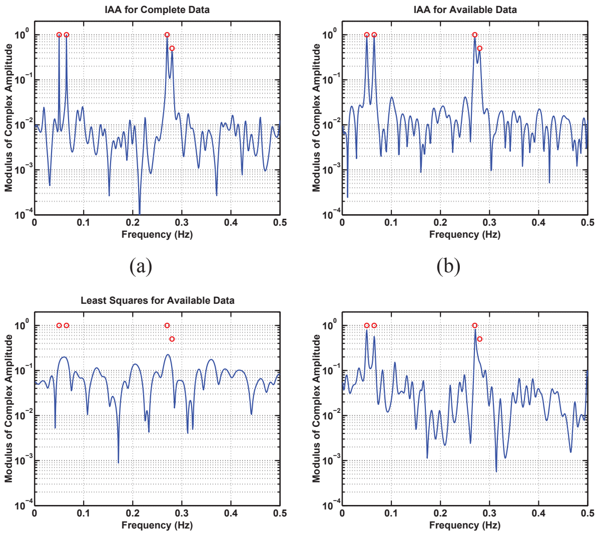


Fig. 2. The modulus of the original data sequence, and the missing and available data samples.



(a) (b)

(c) (d)

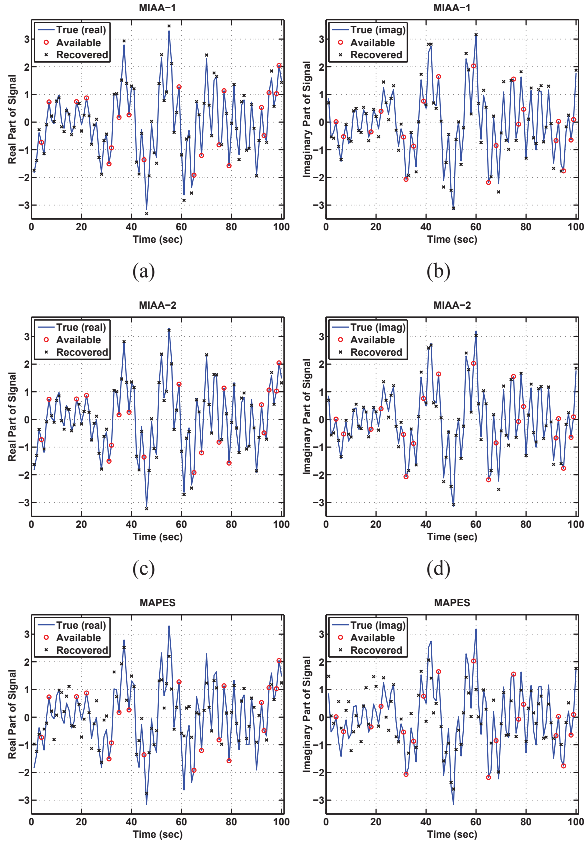
Fig. 3. Spectrum estimates obtained via IAA and the least squares approach. The circles represent the true line spectrum peaks. (a) IAA for the complete data, (b) IAA for the available data, (c) the least squares estimate for the available data. (d) IAA based on both available data and missing data estimated by MAPES.

patterns. Of the two versions of MIAA that we introduced in this paper, MIAA-2 appears to be preferable in most cases as it offers a better recovery performance than MIAA-1 at a reasonable additional computational cost. Finally, we remark on the fact that MIAA has the desirable feature of not altering the available samples during the missing data recovery process. This is in stark contrast with what other data recovery methods do, which also modify the given samples.

6. REFERENCES

[1] H. Ofir, D. Malah, and I. Cohen, "Audio packet loss concealment in a combined MDCT-MDST domain," *IEEE Signal Processing Letters*, vol. 14, pp. 1032–1035, December 2007.

[2] H. Ofir and D. Malah, "Packet loss concealment for audio streaming based on the GAPES and MAPES algorithms," *Proc. 24th IEEE Conv. Electrical and Electronic Engineers,* Eilat, Israel, pp. 280–284, November 2006.



(a) (b)

(c) (d)

(e) (f)

Fig. 4. Estimated missing data. The real parts are shown in (a)(c)(e) and the imaginary parts are shown in (b)(d)(f). (a)(b) MIAA-1. (c)(d) MIAA-2. (e)(f) MAPES.

[3] P. Stoica, E. G. Larsson, and J. Li, "Adaptive filterbank approach to restoration and spectral analysis of gapped data," *The Astronomical Journal*, vol. 120, pp. 2163–2173, October 2000.

[4] A. Afifi and R. Elashoff, "Missing observations in multivariate statistics I: Review of the literature," *J. Amer. Statist. Assoc.*, pp. 595–604, 1966.

[5] T. Orchard and M. A. Woodbury, "A missing information principle: theory and applications," *Prox. Sixth Berkeley Symp. on Math. Statist. and Prob.*, pp. 697–715, 1972.

[6] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychological Methods*, pp. 147–177, 2002.

[7] Y. Wang, J. Li, and P. Stoica, *Spectral Analysis of Signals, The Missing Data Case*. USA: Morgan & Claypool Publishers, 2005.

[8] Y. Wang, P. Stoica, J. Li, and T. L. Marzetta, "Nonparametric spectral analysis with missing data via the EM algorithm," *Digital Signal Processing*, vol. 15, pp. 192–206, March 2005.

[9] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NJ: Prentice-Hall, 2005.

[10] T. Yardibi, J. Li, P. Stoica, M. Xue, and A. B. Baggeroer, "Source localization and sensing: A nonparametric iterative adaptive approach based on weighted least squares," submitted to *IEEE Transactions on Aerospace and Electronic Systems*, 2007.