Statistics and Its Interface Volume 8 (2015) 321–330

Application of structured low-rank approximation methods for imputing missing values in time series

Jonathan Gillard[*] and Anatoly Zhigljavsky

In this paper we consider an important statistical problem of imputing missing values into a time series data. We formulate this problem as a problem of structured low-rank approximation (SLRA), which is a problem of matrix analysis. One of the main difficulties in this SLRA problem is related to the fact that the norm which defines the quality of low-rank approximations is different from the Frobenius norm. We argue that the arising SLRA problem is a very difficult optimization problem and then consider and compare a number of algorithms for its solution.

Keywords and phrases: Time series, Missing data, Hankel structured low-rank approximation.

1. INTRODUCTION

In this paper, we consider the problem of imputing missing values into a time series data. This is a very important problem of statistics, which as a special case includes the problem of forecasting, when the missing values are located at the end of the series. We formulate the problem of imputing missing values in time series as a problem of structured low-rank approximation (SLRA), which is a problem of linear algebra and matrix analysis.

Let us introduce some notation and define the main problem. Let L, K and r be given positive integers such that $1 \leq r < L \leq K$, and set $N = L + K - 1$. By $\mathbb{R}^{L\times K}$ we denote the set of all real-valued $L \times K$ matrices $\mathbf{X} = (x_{l,k})_{l,k=1}^{L,K}$. Let $\mathcal{H}$ be the subset of $\mathbb{R}^{L\times K}$ containing matrices of a specified structure and $\mathcal{M}_r$ be the subset of $\mathbb{R}^{L\times K}$ containing all matrices of rank $\leq r$; that is, $\mathcal{M}_r = \{\mathbf{X} \in \mathbb{R}^{L\times K} \text{ s. t. } \mathrm{rank}(\mathbf{X}) \leq r\}$. We thus define $\mathcal{S}_r = \mathcal{M}_r \cap \mathcal{H}$ to be the set of structured $L \times K$ matrices of rank $\leq r$.

Assume we are given a matrix $\mathbf{X}_\star \in \mathcal{H}$. The general problem of structured low-rank approximation (SLRA) can be stated as the following constrained minimization problem:

$$\min_{\mathbf{X}\in\mathcal{S}_r} ||\mathbf{X} - \mathbf{X}_\star|| \,. \tag{1}$$

where $||\cdot||$ is some pre-defined norm. If $\mathcal{H} = \mathbb{R}^{L\times K}$, so that no structure is specified, then (1) defines the standard (unstructured) low-rank approximation problem abbreviated as LRA. We will only consider the case where $||\cdot||$ in (1) is a weighted Frobenius norm (or semi-norm) defined by

$$||\mathbf{X}||^2_{\mathbf{W}} = \sum_{l,k=1}^{L,K} w_{l,k} x^2_{l,k} \,, \tag{2}$$

where $\mathbf{W} = (w_{l,k})_{l,k=1}^{L,K}$ is a matrix of non-negative numbers (weights). If $w_{l,k} = 1$ for all l and k, then (2) defines the standard Frobenius norm.

Our main interest in this paper lies in matrices of Hankel structure and a special choice of the weight matrix $\mathbf{W}$ in (2). Recall that a matrix $\mathbf{X} = (x_{l,k}) \in \mathbb{R}^{L\times K}$ is Hankel if $x_{l,k}$=const for all pairs (l, k) with $l + k$ =const; this means that all elements on the anti-diagonals of $\mathbf{X}$ are equal. The weight matrix $\mathbf{W} = (w_{l,k})$ will be chosen so that $w_{l,k} = W_{l+k-1}$, where $\{W_1, \ldots , W_N\}$ is a set of 0-1 numbers: $W_i \in \{0, 1\}$ for $i = 1, \ldots , N$. Note that the weight matrix $\mathbf{W}$ is Hankel and that since some of the weights $w_{l,k}$ are equal to zero then (2) defines a semi-norm rather than a norm.

The following is our main problem of interest:

The main problem. *Given a Hankel matrix $\mathbf{X}_\star \in \mathcal{H}$, some integer r and a set of $0 - 1$ numbers $\{W_1, \ldots , W_N\}$, find an approximation to the solution of (1), where $||\cdot||$ is the weighted Frobenius semi-norm (2) with $w_{l,k} = W_{l+k-1}$.*

This problem belongs to the family of weighted SLRA problems and, as discussed in Section 2, can be considered as a problem of optimal imputation of missing values in time series.

SLRA is an important problem, which has applications in a number of different areas including signal processing, speech and audio processing, modal and spectral analysis, modelling dynamical systems, time series analysis, amongst others. For a list of references, see [14]. Behind many data modelling problems there is an equivalent low rank approximation problem [14], and thus the implication of developing methodology for (1) is wide-bearing.

Existing methods for the imputation of missing values in time series assume that the correct model for the observed data is known, or that the data closely follow a particular model. It is then desired to estimate the parameters of the model based on incomplete data. These model-based methods can be divided broadly into two groups; those that use

[*] Corresponding author.

the EM algorithm, and those which use so-called multiple imputation.

The EM algorithm is a well-known algorithm often used to maximise the likelihood function when only incomplete data is available. The method has been demonstrated to be useful for many applications, and there is much literature detailing these applications (see [15], and the references therein for example). For the likelihood function to be fully specified one needs to assume a fixed model. The EM algorithm then maximises the likelihood function based on incomplete data. Each iteration of the EM algorithm yields parameter estimates that incrementally maximise the likelihood function. It has been demonstrated however that the EM algorithm may converge to a spurious local maximizer, or may not converge at all [16]. Despite slow numerical convergence being observed in practise, the algorithm remains popular [20].

The general principle of multiple imputation can be described briefly as follows [17]. The missing data are initially imputed using some (often straightforward) method. Examples of methods include setting all missing values to a constant (such as the mean, or 0) or randomly imputing the missing data. The parameters of the assumed model are then estimated based on this data. This model is then used to re-impute the missing data, and this process is iterated until some pre-specified convergence criteria is met e.g. when there is little change in the estimated parameters and/or the imputed data.

Other possible methods described in the literature are based on considering the time series in the frequency domain. Again the EM algorithm provides a popular method of estimating the frequencies based on incomplete data, and other possible estimators and methods, such as those using the Fourier transform are described in [4]. Many of these methods however assume that the time series is stationary, to avoid complications with the spectral estimation.

The structure of the rest of the paper is as follows. Section 2 gives formal details as to how the Hankel SLRA problem is directly connected to the analysis of time series, and details of imputing missing values in time series is also provided. In Section 3 we describe some properties of problem (1) in the case when $\mathcal{H} = \mathbb{R}^{L\times K}$ and show how these properties can be used to develop algorithms to approximate a solution of (1). These algorithms are built upon to develop algorithms specifically for the case where $\mathcal{H}$ is the space of Hankel matrices and describe how these can be used for the imputation of missing data in time series. Examples and discussion of these algorithms are included in Section 4 before the paper is concluded in Section 5. The main contribution of this paper is to adapt algorithms of weighted low rank approximation to the case where the approximation is also required to be of some structure. We are thus able to discuss the potential of these adapted algorithms for the problem of imputing missing values in time series.

2. THE MAIN PROBLEM AS A PROBLEM OF TIME SERIES ANALYSIS

In the discussions below $\mathcal{H}$ will mostly be assumed to be the set of Hankel matrices. There is a one-to-one correspondence between $L \times K$ Hankel matrices and vectors of size $N = L + K - 1$. Indeed, for a vector $Y = (y_1, \dots, y_N)^T$, the matrix $\mathbf{X} = \mathbb{H}(Y) = (x_{l,k}) \in \mathbb{R}^{L\times K}$ with elements $x_{l,k} = y_{l+k-1}$ is Hankel and vise versa. In the statistical applications, which are dealt with in this paper, the vector $Y = (y_1, \dots, y_N)^T$ is a time series.

In the signal processing literature (see e.g. [1, 11, 19]) the SLRA problem with Hankel structure is often interpreted as a problem of estimating frequencies in sums of damped sinusoids. Indeed, solving the Hankel SLRA problem (1) is equivalent to finding a parametric representation of a vector (time series) whose elements can be represented as sums of exponentially damped sinusoids. This is motivated by the fact that if r is even and a vector $Y = (y_1, \dots, y_N)^T$ is such that $y_j = y_j(\theta)$, where

$$(3)\quad y_j(\theta) = \sum_{l=1}^{r/2} a_l \exp(d_l j) \sin(2\pi\omega_l j + \phi_j), \quad j = 1, \dots, N\,,$$

then the associated Hankel matrix $\mathbf{X}$ belongs to the set $\mathcal{S}_r$ (here θ denote the set of unknown parameters).

We assume that we are given a time series $Y = (y_1, \dots, y_N)^T$ so that some number m of the values y_i are missing and we need to impute these missing values. Let us insert arbitrary numbers in place of the missing values and denote the resulting series $Y_\star = (y_{1\star}, \dots, y_{N\star})$. Define the corresponding $L \times K$ matrix $\mathbf{X}_\star = \mathbb{H}(Y_\star)$.

As we assume that there is an observational error involved, we are allowed to make some small changes to the existing values $y_{i\star}$ and any changes to the inserted values $y_{i\star}$ to ensure that the resulting matrix $\tilde{\mathbf{X}} = (\tilde{x}_{i,j}) = \mathbb{H}(\tilde{Y})$ with $\tilde{x}_{i,j} = \tilde{y}_{i+j-1}$ has rank $\leq r$, where r is given. The series $\tilde{Y} = (\tilde{y}_1, \dots, \tilde{y}_N)$ is an approximation to the series $Y_\star$ and we assume that it satisfies 'the sums of damped sinusoids' model (3); that is, the resulting matrix $\tilde{\mathbf{X}}$ has to belong to the space $\mathcal{S}_r = \mathcal{M}_r \cap \mathcal{H}$, where $\mathcal{M}_r$ and $\mathcal{H}$ are respectively the sets of matrices of rank $\leq r$ and Hankel matrices.

Let $I = \{i_1, \dots, i_m\}$ be the set of indices such that the values y_i with $i \in I$ are missing. The values $y_i \in \bar{I} = \{1, 2, \dots, N\} \setminus I$ are assumed known but evaluated with an observation error. Let us also define the set of weights $\{W_1, \dots, W_N\}$ so that $W_i = 0$ for $i \in I$ and $W_i \geq 0$ for $i \notin I$. Values W_i for $i \notin I$ may be chosen to be inversely proportional to the measurement errors in the corresponding values of y_i; for example, $W_j = \infty$ would indicate that the value y_j is known exactly, with no observation error. For simplicity, we assume $W_i = 1$ for $i \notin I$ so that

$$(4)\qquad W_i = \begin{cases} 0, & \text{if } \ i \in I \\ 1, & \text{if } \ i \notin I\,. \end{cases}$$

Since we need to change values $y_{i\star}$ as little as possible, a very natural norm in (1) is the weighted Frobenius semi-norm (2) with $w_{i,j} = W_{i+j-1}$, where the values $W_1, \ldots, W_N$ are as above. We thus have arrived to the formulation of the problem of imputing missing values as stated in the introduction.

The underlying statistical model used for fitting missing values in this approach assumes that the observations follow the 'signal plus noise' scheme, where the signal has a structure described by the damped sinusoids model (3). Additional details about this model are given in Appendix A.

3. ALGORITHMS

In this section we first describe properties of the optimization problem (1) defined by the distance (2), where we define the space $\mathcal{H}$ to be $\mathcal{H} = \mathbb{R}^{L\times K}$; that is, we consider the weighted unstructured LRA problem. We discuss how one can use these properties to develop the algorithms of solving the weighted unstructured LRA problem and then formulate some algorithms. The second part of this section uses these algorithms of solving the weighted unstructured LRA problem as part of the methodologies which are designed to the the main problem formulated in Introduction; that is, the problem of imputing missing data in time series.

Once the algorithms have been run, there are two potential uses of the output, which will be a vector Y such that the matrix $\mathbb{H}(Y)$ has rank r:

1. The vector Y can be viewed as a 'model-fit' of rank r to $Y_\star$ where the algorithms to be proposed have simultaneously imputed the missing values and found a rank r fit.
2. Elements of Y corresponding to the missing elements of $Y_\star$ can be inserted into $Y_\star$, thus filling in the missing elements of the original series.

In this paper we focus on the former use.

3.1 Algorithms for solving (1) with $\mathcal{H} = \mathbb{R}^{L\times K}$

In this section we describe two properties of the optimization problem (1) defined by the distance (2) where we define the space $\mathcal{H}$ to be $\mathcal{H} = \mathbb{R}^{L\times K}$. These properties can be viewed as two approaches to representing low-rank matrices; which have been called the image and kernel representation respectively [14]. We then suggest how these representations can lead to elementary algorithms for solving (1) with $\mathcal{H} = \mathbb{R}^{L\times K}$.

3.1.1 Image representation and the alternating projections algorithm

If $\mathbf{X} \in \mathcal{M}_r$, then there exists matrices $\mathbf{U} = ||u_{i,k}||_{l,i=1}^{L,r} \in \mathbb{R}^{L\times r}$ and $\mathbf{V} = ||v_{k,j}||_{i,k=1}^{r,K} \in \mathbb{R}^{r\times K}$ such that $\mathbf{X} = \mathbf{UV}$. Hence it is possible to define the LRA problem (1) as an unconstrained optimization over $\mathbf{U}$ and $\mathbf{V}$:

$$(5) \qquad g(\mathbf{U}, \mathbf{V}) \to \min_{\mathbf{U}\in\mathbb{R}^{L\times r}, \mathbf{V}\in\mathbb{R}^{r\times K}} .$$

The equivalent weighted Frobenius norm to (2), can be written

(6)

$$g(\mathbf{U}, \mathbf{V}) = ||\mathbf{UV} - \mathbf{X}_*||_W^2 = \sum_{l=1}^{L}\sum_{k=1}^{K} w_{l,k}\left(\sum_{i=1}^{r} u_{l,i}v_{i,k} - x^*_{l,k}\right)^2$$

There are some disadvantages with the image representation rank r matrices. The decomposition $\mathbf{X} = \mathbf{UV}$ is not unique. Additionally, it is possible to find matrices $\mathbf{M} \in \mathbb{R}^{r\times r}$ such that $g(\mathbf{UM}, \mathbf{M}^{-1}\mathbf{V}) = g(\mathbf{U}, \mathbf{V})$. In this sense it can be claimed that the image representation leads to an overparameterized representation of low-rank matrices. Despite this, it is relatively straightforward to develop algorithms to approximate solutions of (1) based on this representation. The most common algorithm, known as alternating projections, is given below.

3.1.1.1. Algorithm: alternating projections (AP) AP uses the image representation of the rank constraint as follows. Start from initial $\mathbf{U}_0$, compute for $n = 0, 1, \ldots$:

$$\begin{aligned}
\mathbf{V}_n &= \arg\min_{\mathbf{V}\in\mathbb{R}^{r\times K}} ||\mathbf{U}_{n-1}\mathbf{V} - \mathbf{X}_*||_W^2 \\
\mathbf{U}_n &= \arg\min_{\mathbf{U}\in\mathbb{R}^{L\times r}} ||\mathbf{U}\mathbf{V}_n - \mathbf{X}_*||_W^2 .
\end{aligned}$$

This algorithm is run until some stopping criteria has been satisfied. At iteration n, the approximation to $\mathbf{X}_*$ is given by $\mathbf{X}_n = \mathbf{U}_n\mathbf{V}_n$.

3.1.2 Kernel representation and the steepest descent algorithm

Let $\mathbf{0}_{L\times(L-r)}$ be an $L \times (L-r)$ matrix with all entries set to zero. If $\mathbf{X} \in \mathcal{M}_r$, then there exists $\mathbf{R} \in \mathbb{R}^{K\times(L-r)}$ such that $\mathbf{XR} = \mathbf{0}_{L\times(L-r)}$. Hence it is possible to define the LRA problem (1) as the double minimization

$$(7) \qquad \min_{\mathbf{R}\in\mathbb{R}^{K\times(L-r)},\ \mathbf{R}^T\mathbf{R}=\mathbf{I}} \left(\min_{\mathbf{X}\in\ \mathbb{R}^{L\times K},\ \mathbf{XR}=\mathbf{0}} ||\mathbf{X} - \mathbf{X}_*||_W^2\right) .$$

The inner minimization has a unique, closed form solution [12]. Consequently (7) can be written

$$(8) \qquad \min_{\mathbf{R}\in\mathbb{R}^{K\times(K-r)},\ \mathbf{R}^T\mathbf{R}=\mathbf{I}} f(\mathbf{R}) ,$$

$f(\mathbf{R})$ is some function, its expression though is a little bit complicated and can be found in [12]. The kernel representation of low-rank matrices as described above yields a one-to-one parameterization, and so there is no problem of non-uniqueness as in the image representation in the previous section.

Note that it is also possible to set-up a Lagrangean function corresponding to the double minimization (7). Let Γ be an $L \times (L-r)$ matrix of Lagrange multipliers. The Lagrangean is given by

$$\Phi(\mathbf{X},\mathbf{R},\Gamma) = \frac{1}{2}||\mathbf{X}-\mathbf{X}_*||_W^2 + \mathrm{tr}\Gamma^{\mathrm{T}}\mathbf{XR}\,. \tag{9}$$

The derivatives of Φ can be computed as

$$\begin{aligned} \frac{\partial\Phi}{\partial\mathbf{X}} &= \mathbf{W}\odot(\mathbf{X}-\mathbf{X}_*)+\Gamma\mathbf{R}^T, \\ \frac{\partial\Phi}{\partial\mathbf{R}} &= \Gamma^T\mathbf{X}, \\ \frac{\partial\Phi}{\partial\Gamma} &= \mathbf{XR}, \end{aligned} \tag{10}$$

where $\odot$ is the element-wise or so-called Hadamard product. Setting the derivatives to matrices containing zeroes, one obtains conditions on the matrices $\mathbf{X}$ and $\mathbf{R}$ to be at a local minimizer of (7). Additionally, postmultipliying (10) by $\mathbf{X}^T$, one obtains the so-called orthogonality condition that any solution $\mathbf{X}$ of (7) needs to satisfy to be at a local minimizer of (7). This is given by

$$\{\mathbf{W}\odot(\mathbf{X}-\mathbf{X}_*)\}\mathbf{X}^T = \mathbf{0}_{L\times L}\,. \tag{11}$$

This orthogonality condition (11) was discussed in [2] and more recently, for the SLRA problem in [7]. Note that one may also write a Lagrangean function corresponding to the image representation as given in Section 3.1.1.

3.1.2.1. Algorithm: steepest descent (SD) One may optimise the function $f(\mathbf{R})$ directly using an appropriate numerical optimization procedure. The steepest descent algorithm will be used to minimize $f(\mathbf{R})$ in this paper. Technical details and more information as to its implementation can be found in Algorithm 11 as described in [12].

3.2 Algorithms for imputing missing data in time series

Before describing algorithms for solving the main problem as stated in the Introduction, we need to introduce two projections.

Projection to the rank space The celebrated Eckart-Young theorem [5] states that the closest rank r matrix to $\mathbf{X}\in\mathbb{R}^{L\times K}$, for the Frobenius norm, can be computed using the singular value decomposition (SVD) of $\mathbf{X}$ as follows. Let $\sigma_i = \sigma_i(\mathbf{X})$, the singular values of $\mathbf{X}$, be ordered such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_L$. Denote $\Sigma_L = diag(\sigma_1,\sigma_2,\ldots,\sigma_L)$ and $\Sigma_r = diag(\sigma_1,\sigma_2,\ldots,\sigma_r,0,\ldots,0)$. Then the SVD of $\mathbf{X}$ can be written as $\mathbf{X} = U\Sigma_L V^T$ and the matrix

$$\pi^{(r)}(\mathbf{X}) = U\Sigma_r V^T \tag{12}$$

belongs to $\mathcal{M}_r$ and minimizes the value $||\mathbf{X}-\mathbf{A}||_F^2$ over $\mathbf{A}\in\mathcal{M}_r$; that is, $\pi^{(r)}(\mathbf{X})$ is a projection of $\mathbf{X}$ onto $\mathcal{M}_r$. The distance $||\mathbf{X}-\pi^{(r)}(\mathbf{X})||_F^2$ is given by $\sum_{j=r+1}^L \sigma_j^2$.

Projection to the space of Hankel matrices The space $\mathcal{H}$ is a linear subspace of $\mathbb{R}^{L\times K}$ and the closest Hankel matrix to any given matrix is obtained by using the simple diagonal averaging procedure. We thus define $\pi_{\mathcal{H}}(\mathbf{X})$ to be projection of the matrix $\mathbf{X}\in\mathbb{R}^{L\times K}$ onto the space $\mathcal{H}$ as follows. The element $\tilde{x}_{ij}$ of $\pi_{\mathcal{H}}(\mathbf{X})$ is given by

$$\tilde{x}_{i,j} = s_{i+j-1}^{-1}\sum_{l+k=i+j} x_{l,k}\,,$$

where

$$s_n = \begin{cases} n & \text{for } n=1,\ldots,L-1\,, \\ L & \text{for } n=L,\ldots,K-1, \\ N-n+1 & \text{for } n=K,\ldots,N. \end{cases} \tag{13}$$

The value s_n is equal to the number of times the element y_n of the vector Y is repeated in the Hankel matrix $\mathbb{H}(Y)$.

General structure of algorithms The general structure of all the algorithms considered in the remainder of the paper is as follows

$$\mathbf{X}_0 = \mathbf{X}_\star,\;\; \mathbf{X}_{n+1} = \pi_{\mathcal{H}}\left[\mathcal{A}(\overline{\mathbf{X}}_n)\right] \quad \text{for } \; n=0,1,\ldots\,. \tag{14}$$

where $\mathcal{A}(\overline{\mathbf{X}}_n)$ denotes the result of performing some predefined algorithm starting at the matrix $\overline{\mathbf{X}}_n$. We consider the following two forms of $\overline{\mathbf{X}}_n$:

$$\overline{\mathbf{X}}_n = \mathbf{X}_n\,, \tag{15}$$

and

$$\overline{\mathbf{X}}_n = \mathbf{X}_\star\odot\mathbf{W} + \mathbf{X}_n\odot(\mathbf{1}-\mathbf{W})\,, \tag{16}$$

where $\odot$ is the element-wise or so-called Hadamard product and $\mathbf{1}$ is the matrix of ones. In (15), all elements of the matrix $\mathbf{X}_n$ are updated at each iteration of an algorithm. However, if (16) is used then only the missing elements of $\mathbf{X}_\star$ are replaced by the corresponding elements of $\mathbf{X}_n$. We shall refer to (15) and (16) as the updating rules and now introduce the following algorithms.

In this paper we do not discuss in detail selection of the parameters L and r. Typically in the literature on SLRA, one finds arguments for selecting $L=r+1$ [13] and one can also find arguments for selecting L so that the initial Hankel matrix $\mathbf{X}_\star$ is as 'square' as possible. A useful discussion as to the selection of L and r for a number of subspace-based methods is included in [9].

3.2.1 SVD-based algorithms

We define the following SVD-based algorithm

$$\mathbf{X}_0 = \mathbf{X}_\star,\;\; \mathbf{X}_{n+1} = \pi_{\mathcal{H}}\left[\pi^{(r)}(\overline{\mathbf{X}}_n)\right] \quad \text{for } \; n=0,1,\ldots \tag{17}$$

Algorithm (17) with $\overline{\mathbf{X}}_n$ defined by (15), that is $\overline{\mathbf{X}}_n = \mathbf{X}_n$, is analagous to the so-called Cadzow iterations, see [6]. Cadzow iterations are the repeated alternating projections of the matrices, starting at $\mathbf{X}_\star$, to the set of matrices of rank r (by performing a singular value decomposition) and to the set of Hankel matrices (by diagonal averaging). Despite the fact that Cadzow iterations guarantee convergence to a point in the intersection of the spaces of matrices of Hankel structure and those of rank r, they can easily be shown to be sub-optimal in many examples, see [7, 3]. However, they remain popular due to their simplicity. Note also that one iteration of Cadzow iterations for Hankel SLRA corresponds to the basic version of the technique of time series analysis known as singular spectrum analysis (SSA), see [10]; for further details regarding the link between Cadzow iterations and SSA; see, for example, [6].

Algorithm (17) with $\overline{\mathbf{X}}_n$ defined by (16), is analogous to the EM algorithm as introduced in [18]. In [18], the problem (1) is considered with $\mathcal{H} = \mathbb{R}^{L\times K}$; the corresponding algorithm was shown to be effective in many examples, but it was noticed that this algorithm often converges to a local minimum which is not global.

3.2.2 AP-based algorithm

Let $\mathcal{A}_{AP}(\overline{\mathbf{X}}_n)$ denote the result of performing the alternating projections (AP) algorithm defined in Section 3.1.1 for a pre-defined number of iterations, or until some convergence criteria is met, starting at the matrix $\overline{\mathbf{X}}_n$. We can then introduce the following algorithm for imputing missing data in $\mathbf{X}_\star$:

$$\mathbf{X}_0 = \mathbf{X}_\star, \;\; \mathbf{X}_{n+1} = \pi_{\mathcal{H}}\left[\mathcal{A}_{AP}(\overline{\mathbf{X}}_n)\right] . \tag{18}$$

3.2.3 SD-based algorithm

Let $\mathcal{A}_{SD}(\overline{\mathbf{X}}_n)$ denote the result of performing the steepest descent (SD) algorithm defined in Section 3.1.2 for a pre-defined number of iterations, or until some convergence criteria is met, starting at the matrix $\overline{\mathbf{X}}_n$. We can then introduce the following algorithm for imputing missing data in $\mathbf{X}_\star$:

$$\mathbf{X}_0 = \mathbf{X}_\star, \;\; \mathbf{X}_{n+1} = \pi_{\mathcal{H}}\left[\mathcal{A}_{SD}(\overline{\mathbf{X}}_n)\right] . \tag{19}$$

3.2.4 Algorithms based on the ‘sums of damped sinusoids’ representation

The optimization algorithms directed to solving the main problem formulated above work directly in the space of Hankel matrices. Alternative algorithms may be developed based on the parametric representation (3). In this representation the feasible domain are the parameters included in θ instead of the space of Hankel matrices.

In general optimization problems using the representation (3) have less variables to be optimized and these optimization problems may seem to be more straightforward than those using a feasible domain of Hankel matrices. However, the objective functions obtained using (3) are severely multimodal and the associated global optimization problems can be very complex with the objective function possessing a large number of local minimizers and large Lipschitz constant, which make conventional optimization methods unsuitable. This has been demonstrated by the authors in recent papers [7, 8]. A summary of this work and an example is included in Appendix A. Objective functions arising using the space of Hankel matrices as a feasible domain also suffer some multi-extremality issues although these are not as severe as those described earlier. We thus restrict our attention in this paper to algorithms which use the space of Hankel matrices as a feasible domain.

4. EXAMPLES

4.1 Example 1

This is a very simple example where the true solution to the optimization problem (1) gives zero distance to the target matrix $\mathbf{X}_\star$. All examples selected in this Section are taken from a wider selection; these specific examples were chosen for their simplicity.

4.1.1 One missing value ($m = 1$)

Set $L = 5$, $r = 2$ and assume that we are given the time series $Y = (0, 1, 0, -1, 0, \times, 0, -1, 0, 1, 0)$, where × denotes a missing value. Let us insert some number α in place of the missing observation × and denote the resulting series $Y_\star = (0, 1, 0, -1, 0, \alpha, 0, -1, 0, 1, 0)$. Define the matrix $\mathbf{X}_\star$ to be such that $\mathbf{X}_\star = \mathbb{H}(Y_\star)$. When $\alpha = 1$ then $\mathrm{rank}(\mathbf{X}_\star) = 2$.

Table 1 contains the value of the weighted norm (2) obtained using the algorithms described in Section 3.2; namely the SVD-based algorithm (17), the AP-based algorithm (18) and the SD-based algorithm (19) using the two updating rules (15) and (16). At each iteration of (18) and (19) the AP and SD components of the algorithm were run for 1000 iterations. Each algorithm was allowed to continue until it converged to a feasible solution. Each algorithm was initialized at different values of $Y_\star$ with α as given in the table.

The SVD-based algorithm (17) with updating rule (15) performs least favourably, apart from the case when $\alpha = 1$. Remember however that when $\alpha = 1$ then $\mathrm{rank}(\mathbf{X}_\star) = 2$ and so the starting value of the algorithm is already at a feasible solution. This algorithm performs more favourably for values of α close to 1.

The AP-based algorithm (18) with updating rule (15) performs well for values of $\alpha > -0.5$, and although it performs better than SVD-based algorithm (17) with updating rule (15) for $\alpha = -0.75$ and $\alpha = -1$, has noticeably poorer performance when $Y_\star$ is initially imputed with these values of α.

The SD-based algorithm (19) with updating rule (15) performs well for many values of α, but gives particularly poor solutions for $\alpha = -0.5$, $\alpha = 1.5$ and $\alpha = 2$. This implies that on occasion the SD-based algorithm with updating rule (15) may get stuck in regions close to a local minima. It can be seen that all of the algorithms perform better across the

Table 1. Weighted norms (2) obtained using the algorithms described in Section (3.2); namely the SVD-based algorithm (17), the AP-based algorithm (18) and the SD-based algorithm (19) using the two updating rules (15) and (16). Here (1) denotes that the algorithm has been run using updating rule (15); (2) denotes that the algorithm has been run using updating rule (16)

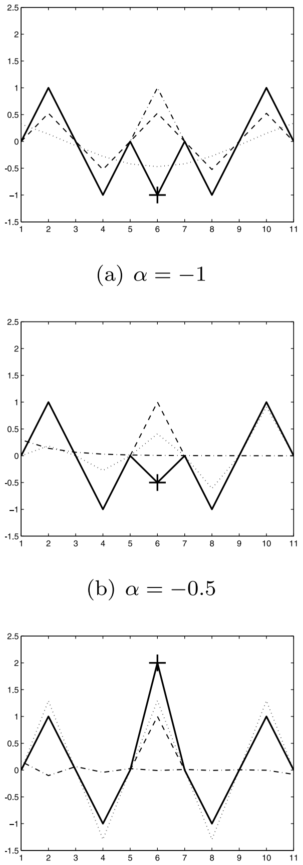
α	SVD(1)	AP(1)	SD(1)	SVD(2)	AP(2)	SD(2)
-1	9.124	2.710	0	0	0	0
-0.75	9.382	1.313	0	0	0	0
-0.5	4.064	0	11.841	0	0	0.655
-0.25	1.622	0	0	0	0	0
0	1.038	0	0	0	0	0
0.25	0.584	0	0	0	0	0
0.50	0.600	0	0	0	0	0
0.75	0.065	0	0	0	0	0
1	0	0	0	0	0	0
1.25	0.065	0	0	0	0	0
1.50	0.260	0	10.335	0	0	0
1.75	0.584	0	0	0	0	0
2	1.038	0	12.111	0	0	0

entire range of α using updating rule (16). However the SD-based algorithm with this updating rule has been trapped at a local minima when $\alpha = -0.5$.

Figure 1 contains plots of the original time series $Y_\star$ with three approximations for selected values of α. The three approximations have rank $r = 2$ and are obtained using the SVD-based algorithm (17), AP-based algorithm (18) and SD-based algorithm (19) using updating rule (15). The particular updating rule and values of α were selected in order to see differences between the algorithms more clearly. The missing value (imputed with initial guess α) is highlighted with a $+$. For $\alpha = -1$ it can be seen that the SVD-based algorithm (17) has converged to a rank 2 solution far away from the observed data $Y_\star$. As explained in Section 3.2, the SVD-based algorithm (17) with updating rule (15) is guaranteed to converge to the space of rank 2 Hankel matrices (assuming that it is run for a sufficient number of iterations) but it is not guaranteed to converge to the optimal solution. Recall that when $\alpha = 1$ then $\text{rank}(\mathbf{X}_\star) = 2$. The SD-based algorithm converged to this solution, whilst the AP-based algorithm converged to a different rank 2 solution. These observations for $\alpha = -1$ also hold for the cases $\alpha = -0.5$ and $\alpha = 2$.

4.1.2 Four missing values ($m = 4$)

We now reconsider the example above, but assume that there are additional missing values. Set $L = 5$, $r = 2$ and assume that we are given the time series $Y = (\times, 1, 0, -1, 0, \times, \times, -1, 0, 1, \times)$, where $\times$ denotes a missing value. We insert a number α in place of the missing observations $\times$ and denote the resulting series $Y_\star = (\alpha, 1, 0, -1, 0, \alpha, \alpha, -1, 0, 1, \alpha)$. Define the matrix $\mathbf{X}_\star$ to be such that $\mathbf{X}_\star = \mathbb{H}(Y_\star)$.



(a) $\alpha = -1$

(b) $\alpha = -0.5$

(c) $\alpha = 2$

Figure 1. Plots of the original time series $Y_\star$ (solid line) with $r = 2$ approximations obtained using the SVD-based algorithm (17) (dotted line), AP-based algorithm (18) (dashed line) and SD-based algorithm (19) using updating rule (15) (dot-dashed line). The missing value (imputed with initial guess α) is highlighted with a $+$.

Table 2 contains the value of the weighted norms (2) obtained using the algorithms described in Section (3.2); namely the SVD-based algorithm (17), the AP-based algorithm (18) and the SD-based algorithm (19) using the two updating rules (15) and (16). The AP and SD components of the algorithms were run for 1000 iterations. Each algorithm was allowed to continue until it converged to a feasible solution. Each algorithm initiated at different values of $Y_\star$ with α as given in the table.

As this is a more difficult example, with just under half of the observations missing, it is not surprising that the algorithms display a poorer performance for this example. Here it can be seen more clearly that updating rule (16) is preferable for all algorithms considered. This time it can be seen that the SD-based algorithms under both updating rules (15) and (16) have problems, on occasion, converging to the optimal solution. The SVD-based algorithm with updating

Table 2. Weighted norms (2) obtained using the algorithms described in Section (3.2); namely the SVD-based algorithm (17), the AP-based algorithm (18) and the SD-based algorithm (19) using the two updating rules (15) and (16). Here (1) denotes that the algorithm has been run using updating rule (15); (2) denotes that the algorithm has been run using updating rule (16)

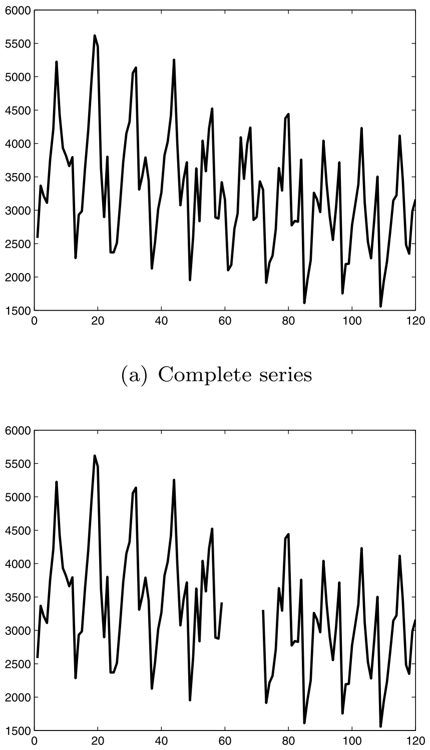
α	SVD(1)	AP(1)	SD(1)	SVD(2)	AP(2)	SD(2)
-1	9.463	6.204	0	4.672	2.537	0
-0.75	10.337	7.294	0	0	1.548	0
-0.5	9.986	11.336	12.000	0	0	1.653
-0.25	1.705	0.001	10.346	0	0	0
0	1.038	0	9.991	0	0	0
0.25	0.668	0.001	6.858	0	0	1.442
0.50	0.613	0.001	0	0	0	0
0.75	0.886	0.002	1.381	0	0	0
1	1.489	0.005	10.982	0	0	0
1.25	2.406	0.008	11.942	0	0	0
1.50	3.616	0.013	10.335	0	0	2.183
1.75	5.090	0.019	0	0	0	0
2	6.794	0.026	0	0	0	0

rule (16) performed poorly for $\alpha = -1$, as did the AP-based algorithm with the same updating rule (which also found a poor solution when $\alpha = -0.75$).

4.2 Example 2: fortified wine

To demonstrate the methods of filling in missing data practically, we now consider a 'real-life' time series. The time series to be considered is the monthly volumes of fortified wine sales observed in Australia from January 1980 until January 1990. A plot of the time series is given in Figure 2(a). In this example we removed 12 known values, starting at the 61st point, that is, we assume that the values for one year (January 1985 – December 1985) are unknown. The data with missing section is shown in Figure 2(b). This time series was also studied in [10] using the method of SSA as described following equation (17). In [10] they recommend the selection of the parameters $L = 60$ and $r = 11$, and hence these are the parameters that will be used in this example. To impute the missing data, in [10] they use the method of multiple imputation (described in Section 1) with the technique of SSA.

Due to the sometimes erratic behaviour observed with the SD-based algorithms under both updating rules (15) and (16), we will concentrate our attention on the SVD-based algorithm (17) and AP-based algorithm (18) under what appears to be the preferable updating rule (16). We initially set all missing values of the time series to 0 (note that, for this example, similar results were found if the missing values were set to the mean of the series). The AP components of the algorithm (18) were run for 1000 iterations. Each algorithm was allowed to continue until it converged to a feasible solution.



(a) Complete series

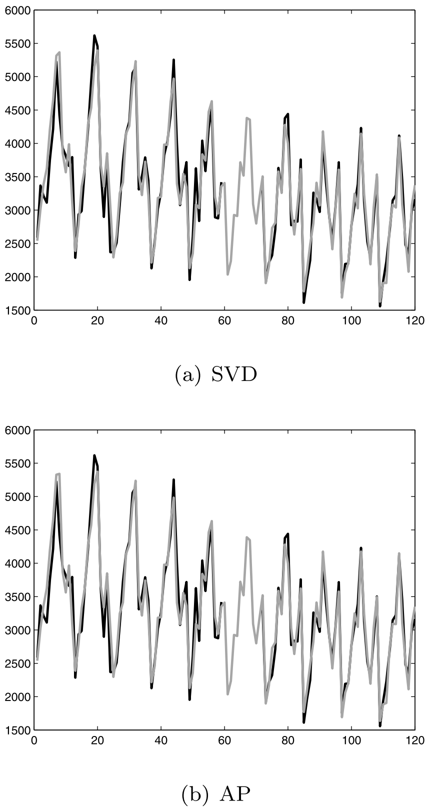
(b) Series with section missing

Figure 2. Monthly volumes of fortified wine sales in Australia from January 1980 until January 1990.

Figure 3 contains plots of the monthly volumes of fortified wine sales in Australia from January 1980 until January 1990 with rank 11 approximations obtained by the SVD-based algorithm (17) and the AP-based algorithm (18) using updating rule (16). The missing values have also been imputed by both of these algorithms. There appears to be little difference between the two approximations. The square root of the average mean square deviations (MSD) comparing the imputed missing values with their observed (but not used) values were computed to be 214.01 and 215.51 for the SVD-based algorithm (17) and the AP-based algorithm (18) using updating rule (16) respectively. These results are comparable to the best MSD given in [10], using SSA and multiple imputation, which was reported to be 216.2, see [10], page 103.

5. CONCLUSION

In this paper we have considered the application of the structured low-rank approximation methods for imputing missing values in time series. After introducing the general problem of structured low-rank approximation we restricted our attention to defining the closeness of our approximation by a weighted Frobenius norm with each weight taking a single value in the set $\{0, 1\}$. We related the main problem of



(a) SVD

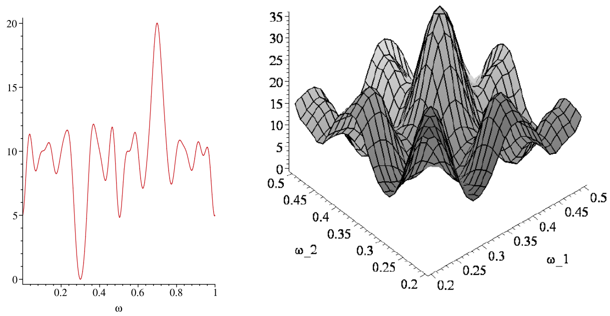
(b) AP

Figure 3. Monthly volumes of fortified wine sales in Australia from January 1980 until January 1990 (black) with rank 11 approximations (grey) obtained by the SVD-based algorithm (17) *and the AP-based algorithm* (18) *using updating rule* (16).

the paper with the problem of Hankel SLRA and described how this is equivalent to finding a parametric representation of a time series whose elements can be represented as sums of exponentially damped sinusoids. We then described algorithms to solve the unstructured low-rank approximation problem and developed these for the purpose of imputing missing data in time series. Algorithms based on the use of the SVD, or the so-called alternating projections (AP) seemed more reliable, and both algorithms gave satisfactory results in the examples considered.

APPENDIX A. REPRESENTATION BY THE SUMS OF DAMPED SINUSOIDS

In the signal processing literature on Hankel SLRA (see e.g. [19]), a common approach often used is to seek a solution in the parametric form (3). To do this, one assumes that the observations y_j are $y_j(\theta) + \varepsilon_j$ where $\varepsilon_1, \dots \varepsilon_N$ is noise and $y_j(\theta)$ follow the model (3), where parameters are $\theta = (a, d, \omega, \phi)$ with a, d, ω and ϕ vectors.



(a) Plot of $f(\omega_1, 0.45)$

(b) Plot of $f(\omega_1, \omega_2)$

Figure 4. Plots of the function $f(\omega_1, \omega_2)$ with $\sigma^2 = 0$ and $\omega^{(0)} = (0.3, 0.45)$.

The following simpler representation is also often used:

$$y_j(\theta) = \sum_{l=1}^{r/2} a_l \sin(2\pi\omega_l j + \phi_j), \;\; j = 1, \dots, N\,. \tag{20}$$

For complex-valued series, the analogue of the series (3) is a sum of complex damped exponentials:

$$y_j(\theta) = \sum_{l=1}^{r} a_l \exp(i\phi_l) \exp[(i2\pi\omega_l + d_l)j], \;\; j = 1, \dots, N\,. \tag{21}$$

The case of signal poles co-inciding is usually omitted, as this would induce additional polynomial terms in j into (21). For an explanation, see for example [11]. In this section, we shall use the form (20). Some of the parameters may be omitted if they are assumed known.

If we assume that there is a true signal represented in the form (3) or (20), such as in the standard ‘signal plus noise’ model, then we denote the true values of parameters by $a^{(0)}$, $d^{(0)}$, $\omega^{(0)}$ and $\phi^{(0)}$. The associated true signal values will be $y_j^{(0)}$, $j = 1, \dots, N$.

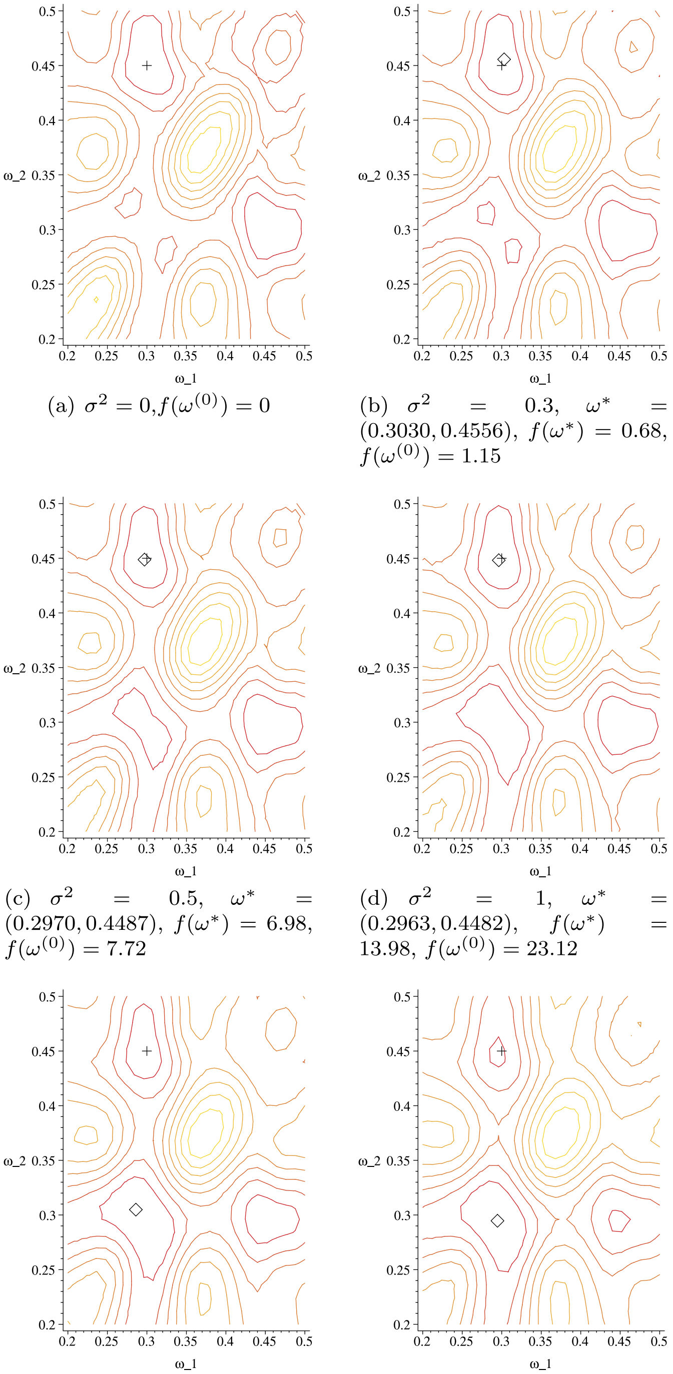
If the observations are noise-free, then the vector of observations $Y = (y_1, \dots, y_N)^T$ coincides with the signal vector $Y^{(0)} = (y_1^{(0)}, \dots, y_N^{(0)})^T$. Otherwise Y is different from $Y^{(0)}$. In the signal plus noise model, $y_j = y_j^{(0)} + n_j$, where $\{n_j, j = 1, \dots, N\}$ is the series of noise terms (not necessarily random).

If we use the weighted semi-norm (2) for defining the objective function with weights $w_{l,k} = W_{l+k-1}$ then, given an observed vector $Y = (y_1, \dots, y_N)^T$, we can write the objective function explicitly as

$$f(\theta) = \sum_{j=1}^{N} W_j \varepsilon_j^2(\theta) \tag{22}$$

where

$$\varepsilon_j(\theta) = y_j - \sum_{i=1}^{r/2} a_i \exp(d_i j) \sin(2\pi\omega_i j + \phi_i) \tag{23}$$



(a) $\sigma^2 = 0, f(\omega^{(0)}) = 0$

(b) $\sigma^2 = 0.3$, $\omega^* = (0.3030, 0.4556)$, $f(\omega^*) = 0.68$, $f(\omega^{(0)}) = 1.15$

(c) $\sigma^2 = 0.5$, $\omega^* = (0.2970, 0.4487)$, $f(\omega^*) = 6.98$, $f(\omega^{(0)}) = 7.72$

(d) $\sigma^2 = 1$, $\omega^* = (0.2963, 0.4482)$, $f(\omega^*) = 13.98$, $f(\omega^{(0)}) = 23.12$

(e) $\sigma^2 = 2$, $\omega^* = (0.2860, 0.3049)$, $f(\omega^*) = 26.27$, $f(\omega^{(0)}) = 38.12$

(f) $\sigma^2 = 4$, $\omega^* = (0.2947, 0.2947)$, $f(\omega^*) = 49.59$, $f(\omega^{(0)}) = 74.27$

Figure 5. Contour plots of $f(\omega_1, \omega_2)$ with $\omega^{(0)} = (0.3, 0.45)$ (marked $+$) and the global minimizer ω^* (marked $\diamond$), for different values of σ^2. Values of f at $\omega^{(0)}$ and ω^* are provided.

for the full sums of damped sinusoids representation (3) and

$$\varepsilon_j(\theta) = y_j - \sum_{i=1}^{r/2} a_i \sin(2\pi\omega_i j + \phi_i) \tag{24}$$

for case (20).

The fact that some of the weights W_j are equal to zero does not make much difference in what concerns properties of the objective function (22).

We are aware of only three papers, which are [7, 8, 11], which contain discussion about the behaviour of the objective function (22) (with weights $W_j = 1$) and its multiextremality. However, there were no algorithms proposed that would effectively deal with this phenomena. In [7, 8] it is demonstrated that the classical methods often do not even converge to a locally optimal matrix. To give an example demonstrating the multiextremality, we consider the following example (which is similar to an example considered in [7] in the case $W_j = 1$ for all j).

Consider the following objective function

(25)

$$f(\omega) = f(\omega_1, \omega_2) = \sum_{j=1}^{N} W_j \left(f_j - \sin(2\pi\omega_1 j) - \sin(2\pi\omega_2 j)\right)^2$$

and assume we have a series of $N = 10$ observations, with $\omega^{(0)} = (0.3, 0.45)$. In this example, we assume that y_j are not observed, but rather $y_j + n_j$ where $\{n_j, j = 1, \ldots, N\}$ is the series of uncorrelated normally distributed noise terms with variance σ^2. The value y_5 was assumed missing. Fig. 4 contains plots of the function f highlighting the multimodality of the function. Fig. 5 contains contourplots of $f(\omega_1, \omega_2)$ for varying values of σ^2. Values of f at $\omega^{(0)}$ and the global minimizer, which we denote ω^*, are provided.

In summary the optimization problem obtained by using the sums of damped sinusoids parameterization (3) is very difficult with the objective function possessing many local minima. The objective functions has very large Lipschitz constants which increase with N, the number of observations [7]. Moreover, the number of local minima in the neighbourhood of the global minimum increases linearly in N. Increasing the noise variance of the observed data increases the complexity of the objective function and moves the global minimizer away from the true value.

Received 4 February 2014

REFERENCES

[1] Bishop, W. B. and Djuric, P. M. (1996). Model order selection of damped sinusoids in noise by predictive densities. *IEEE Trans. Signal. Process.* **44** 611–619.

[2] De Moor, B. (1993). Structured total least squares and L2 approximation problems. *Linear Algebra and its Applications* **188–189** 163–205. MR1223460

[3] De Moor, B. (1994). Total least squares for affinely structured matrices and the noisy realization problem. *Signal Processing, IEEE Transactions on* **42** 3104–3113.

[4] Dunsmuir, W. and Robinson, P. M. (1981). Estimation of Time Series Models in the Presence of Missing Data. *Journal of the American Statistical Association* **76** 560–568.

[5] Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1** 211–218.

[6] Gillard, J. (2010). Cadzow's basic algorithm, alternating projections and singular spectrum analysis. *Statistics and Its Interface* **3** 335–343. MR2720137

[7] Gillard, J. and Zhigljavsky, A. A. Optimization challenges in the structured low rank approximation problem. *Journal of Global Optimization (to appear).* MR3119378

[8] Gillard, J. and Zhigljavsky, A. A. (2011). Analysis of Structured Low Rank Approximation as an Optimization Problem. *Informatica* **22** 489–505. MR2885683

[9] Golyandina, N. (2010). On the choice of parameters in Singular Spectrum Analysis and related subspace-based methods. *Statistics and Its Interface* **3** 259–279. MR2720132

[10] Golyandina, N. and Zhigljavsky, A. A. (2013). *Singular Spectrum Analysis for time series. Springer Briefs in Statistics.* Springer. MR3024734

[11] Lemmerling, P. and Van Huffel, S. (2001). Analysis of the structured total least squares problem for Hankel/Toeplitz matrices. *Numerical Algorithms* **27** 89–114. MR1847986

[12] Manton, J. H., Mahony, R. and Hua, Y. (2003). The geometry of weighted low-rank approximations. *Signal Processing, IEEE Transactions on* **51** 500–514. MR1956702

[13] Markovsky, I. (2010). Bibliography on total least squares and related methods. *Statistics and Its Interface* **3** 329–334. MR2720136

[14] Markovsky, I. (2012). *Low rank approximation: Algorithms, implementation, applications.* Springer. MR2867878 MR2867878

[15] McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions* **382**. John Wiley & Sons. MR2392878

[16] McLachlan, G. J. and Peel, D. (1999). Computing issues for the EM algorithm in mixture models. In *Interface'99* **3** 421–430. Interface Foundation of North America.

[17] Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91** 473–489.

[18] Srebro, N. and Jaakkola, T. (2003). Weighted low-rank approximations. In *ICML* **3** 720–727.

[19] Van Huffel, S. (1993). Enhanced resolution based on minimum variance estimation and exponential data modeling. *Signal Processing* **33** 333–355.

[20] Wu, C. F. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11** 95–103. MR0684867

Jonathan Gillard
School of Mathematics
Cardiff University
UK
E-mail address: GillardJW@Cardiff.ac.uk

Anatoly Zhigljavsky
School of Mathematics
Cardiff University
UK
E-mail address: ZhigljavskyAA@Cardiff.ac.uk