Statistics and Its Interface Volume 11 (2018) 615–630

A sequential majorization method for approximating weighted time series of finite rank*

Hou-Duo Qi†, Jian Shen, and Naihua Xiu

The low-rank Hankel matrix optimization has become one of the main approaches to the signal extraction from noisy time series of finite rank. The approach is particularly effective if different weights are enforced to the data points to reflect their relative importance. Two guiding principles for developing such an approach are (i) the Hankel matrix optimization should be computationally tractable, and (ii) the objective in the optimization should be a close approximation to the original weighted least-squares. In this paper, we introduce a sequential approximation that satisfies (i) and (ii) based on the technique of majorization. A new approximation is constructed as soon as a new iterate is computed from the previous approximation and it makes use of the latest gradient information of the objective, leading to more accurate an approximation to the objective. The resulting subproblem bears a similar structure to an existing scheme and hence can be efficiently solved. Convergence of the sequential majorization method (SMM) is guaranteed provided that the solution of the subproblem satisfies a sandwich inequality. We also compare SMM with two leading methods in literature on real-life problems. Significant improvement is observed in some cases.

AMS 2000 subject classifications: Primary 62M10, 62M15; secondary 62P99.
Keywords and phrases: Singular spectrum analysis, Time series of finite rank, Hankel matrix, Majorization method, Alternating projection.

1. INTRODUCTION

Singular spectrum analysis (SSA) has become one of the leading methods in approximating structured time series and has been widely used in many disciplines. We refer to the two comprehensive books [13, 15] for a systematic treatment of SSA; however, one drawback that SSA has often been criticized of is that it assigns Hankel-type weights for the time series concerned. In particular, the weights assigned to data at the both ends of the data series are smaller than those at the middle. This drawback could be serious when the newest data should bear higher weights than the old ones. At extreme, it is more reasonable that there should be no weights (e.g., zero weights) assigned to the missing data points. Therefore, developing an efficient scheme of handling weights is extremely important yet challenging and it has become a focal point of the two recent papers [11, 28], where tractable approximation schemes have been developed. In this paper, we introduce a new approach of sequential approximation via majorization. We will demonstrate that the new approach is very promising in handling arbitrarily given weights.

Let $\mathbf{x}^T = (x_1, x_2, \ldots, x_N) \in \mathbb{X}^N$ be a time series of length N. Suppose $L > 0$ is a given integer known as window length. We define the L-lagged vector

$$\mathbf{x}_i^T = [x_i, x_{i+1}, \ldots, x_{i+L-1}]$$

and the L-lagged data matrix X

$$X = \mathcal{T}(\mathbf{x}) := [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K], \tag{1}$$

where $K := N - L + 1$, $\mathcal{T} : \mathbb{X}^N \mapsto \mathcal{H}$ maps $\mathbf{x}$ to the Hankel matrix X defined in (1) and ":=" means "define". Let $r > 0$ be another given integer. We let $\mathbb{X}_r^N$ denote all the time series of length N whose Hankel matrices do not have rank greater than r:

$$\mathbb{X}_r^N := \left\{ \mathbf{x} \in \mathbb{X}^N \;\mid\; \operatorname{rank}(\mathcal{T}(\mathbf{x})) \le r \right\}.$$

Such a low-rank Hankel representation of time series has many important applications such as true signal separation from its noisy corruptions. In particular, it has a close link to time series that are governed by a linear recurrence relation. We refer the interested reader to [13, 15, 11, 28]. We say that a time series in $\mathbb{X}_r^N$ has a finite rank.

Now suppose we have a noise corrupted series $\mathbf{y} \in \mathbb{X}^N$. Our purpose is to find the nearest approximation to $\mathbf{y}$ from $\mathbb{X}_r^N$ in the following weighted norm:

$$\min \; \|\mathbf{x} - \mathbf{y}\|_{\mathbf{w}}^2 := \sum_{i=1}^N w_i (x_i - y_i)^2, \qquad \text{s.t.} \quad \mathbf{x} \in \mathbb{X}_r^N, \tag{2}$$

where $\mathbf{w}^T := (w_1, \ldots, w_N)$ with $w_i \ge 0$, $i = 1, \ldots, N$ being given weights. As pointed out in [11], the choice of w_i's can reflect the importance of known data. Practical scenarios include data missing and the newest data weighting more

*This research was supported by the National Science Foundation of China (11728101) and the 111 Project of China (B16002).
†Corresponding author.

than the old data (e.g., monotone weights $w_1 \le w_2 \le \cdots \le w_N$). Therefore, (2) provides more a convenient model in approximating observed time series; however, the price for such convenience is that the problem is difficult to solve. It is noted that the problem (2) is the same as that defined by Eq. (1) in [11] with the distance function $\rho(\cdot,\cdot)$ defined by Eq. (8) there.

The method of SSA is a leading method for (2) with a special choice of the weight vector $\mathbf{w}$. It tries to solve the following optimization problem in terms of the rank-deficient Hankel matrix:

$$(3) \qquad \min \ \|X - Y\|^2, \qquad \text{s.t.} \ \ X \in \mathcal{H} \cap \mathcal{M}_r,$$

where $Y := \mathcal{T}(\mathbf{y})$ and $\mathcal{M}$ is the set of matrices of size $L \times K$ and $\mathcal{M}_r \subset \mathcal{M}$ consists of those in $\mathcal{M}$ whose ranks are not greater than r. A widely used method for (3) is the Cadzow method [4] and the SSA is just one iteration of Cadzow's method. Both methods are to be described in the next section (see (7) and (8) below). Translating to the form of (2), the problem (3) corresponds to the following choice of $\mathbf{w}$ (assuming $L \le K$ without loss of any generality, see [28, Eq.(4)]):

$$w_i := \begin{cases} i & \text{for } i = 1, \ldots, L-1 \\ L & \text{for } i = L, \ldots, K \\ N - i + 1 & \text{for } i = K+1, \ldots, N. \end{cases}$$

It is obvious that the weights at the both ends of the series are smaller than that in the middle part and this weighting is against the common usage that the newest data should have weights not smaller than those for the old data.

Gillard and Zhigljavsky [11] proposed a scheme that approximates (2) by solving the computationally tractable (Q, R)-norm problem:

$$(4) \qquad \begin{array}{ll} \min & \|X - Y\|^2_{(Q,R)} := \mathrm{Tr}\Big(Q(X-Y)R(X-Y)^T\Big) \\ \text{s.t.} & X \in \mathcal{H} \cap \mathcal{M}_r \end{array}$$

where Q and R are positive diagonal matrices of size $L \times L$ and $K \times K$ respectively, and $\mathrm{Tr}(A)$ is the trace of A. The choice of (Q, R) matrices was accomplished in [11] through a convolution operator, which itself is a nonlinear optimization problem. In another development, Zvonarev and Golyandina [28] propose to solve

$$(5) \qquad \begin{array}{ll} \min & \|X - Y\|^2_C := \mathrm{Tr}\Big((X-Y)C(X-Y)^T\Big) \\ \text{s.t.} & X \in \mathcal{H} \cap \mathcal{M}_r, \end{array}$$

where C is a $K \times K$ positive semidefinite symmetric matrix. A computationally tractable choice is when C is positive diagonal, leading to a class of choices represented as `Cadzow`(α) with $\alpha > 0$ and a particular $\widehat{C}$ choice in [28]. Many interesting results can be found in the two papers about those two approximation schemes. In particular, the equivalence between (2) and (4) or (5) has been established in [11, 28] for some special choices of the weight vector $\mathbf{w}$.

In this paper, we propose a new approximation scheme via majorization, a technique that has been widely used in dealing with hard optimization problems such as multidimensional scaling [6], the low-rank correlation matrix problem [8], convex semidefinite programming [19, 22], and the Euclidean distance matrix problem with low-embedding dimensions [24]. Rather than approximating the hard problem (2) once as in [11, 28], the new scheme yields a sequential approximations, which are hoped to provide more and more an accurate approximation each step. Each approximation subproblem of our scheme enjoys the form of (4) or (5) and hence it is relatively easy to solve. We will demonstrate the advantage of the proposed scheme over the existing methods in [11, 28] on two popular test problems in literature.

The paper is organized as follows. In the next section, we describe some necessary material needed for later development. In particular, we will describe the weighted projections onto $\mathcal{H}$ and $\mathcal{M}_r$ respectively. Section 3 develops the sequential approximation scheme based on the majorization technique and its resulting method, named as `SMM` for *Sequential Majorization Method*. We will show that the subproblem of `SMM` can be reformulated as (Q, R)-norm approximation problem, see Prop. 3.1. Under a sandwich inequality condition, `SMM` is shown to be convergent in terms of the objective function value (Thm. 3.3). We will use linear programming to improve our approximation in Sect. 4, where we list 4 choices in constructing fast and practical approximations. In Sect. 5, we compare our method with the (Q, R)-norm method [11] and the `Cadzow`(α) method [28] using two popular test problems. Significant improvement is observed in some cases. We conclude our paper in Sect. 6.

2. PRELIMINARIES

This part describes most notation used in the paper with some results on the classical SSA and Cadzow's method [4]. Continuing with the notation already used in the introduction such as $\mathcal{M}$, $\mathcal{M}_r$ and $\mathcal{H}$, we emphasize that we use boldfaced lower case letters for column vectors. For example, $\mathbf{x}$ is a column vector in $\mathbb{R}^N$ and its elements are denoted as x_i, $i = 1, \ldots, N$. In particular, $\mathbf{1}_L$ is the vector of all ones in $\mathbb{R}^L$. We use $\mathbf{x} \ge 0$ to say that $\mathbf{x}$ is nonnegative and $\sqrt{\mathbf{x}}$ to be the componentwise square root vector of $\mathbf{x} \ge 0$ (i.e., the ith element of $\sqrt{\mathbf{x}}$ is $\sqrt{x_i}$). The upper case letter such as A is often for a matrix, whose elements are denoted as a_{ij}. The inner product in $\mathcal{M}$ is standard:

$$\langle A, \ B \rangle := \mathrm{Tr}(A^T B), \qquad \forall \ \ A, B \in \mathcal{M}.$$

Throughout the paper, the induced Frobenius norm is used and is denoted as $\|\cdot\|$ and Y is reserved for the Hankel matrix $\mathcal{T}(\mathbf{y})$ that defines (3). For two vectors $\mathbf{p}$ and $\mathbf{q}$ in $\mathbb{R}^N$, $(\mathbf{p} \circ \mathbf{q})$ denotes the vector of their Hadamard product defined by

$$\mathbf{p} \circ \mathbf{q} := \Big(p_i q_i\Big)_{i=1}^{N}.$$

The Hadamard product of two matrices of same size is defined similarly: $(A \circ B)_{ij} := a_{ij} b_{ij}$.

Suppose Ω is a closed set in $\mathcal{M}$. For a given $A \in \mathcal{M}$, we use $\Pi_\Omega(A)$ to denote its projection on Ω. That is, $\Pi_\Omega(A)$ is an optimal solution of the following problem:

$$\min \ \|X - A\|, \qquad \text{s.t.} \ \ X \in \Omega. \tag{6}$$

When $\Omega = \mathcal{H}$, then $\Pi_\mathcal{H}(A)$ has a closed-form solution, which is obtained by anti-diagonal averaging of A (see for example [13, 14]). When $\Omega = \mathcal{M}_r$, $\Pi_{\mathcal{M}_r}(A)$ also has a closed form solution, which can be obtained through the singular value decomposition (SVD) of A (see for example [26, 12, 4, 13, 14]). The two projections are the basic elements in the method of SSA and Cadzow's method [4], both of which aim to solve the problem (3). The SSA computes

$$\widehat{X} = \Pi_\mathcal{H}\Big(\Pi_{\mathcal{M}_r}(Y)\Big) \tag{7}$$

and the Cadzow method iteratively computes a sequence $\{X^k\}$ by starting with $X^0 = Y$

$$X^{k+1} = \Pi_\mathcal{H}\Big(\Pi_{\mathcal{M}_r}(X^k)\Big), \quad k = 0, 1, \ldots. \tag{8}$$

Cadzow's method outputs $\widehat{X}$ as the final iterate in (8) when convergence is observed. In fact, one can view (7) as just one iteration of (8). The rest of the SSA method and the Cadzow method is same and is to extract structural signals from the spectral information of $\widehat{X}$. An excellent empirical comparison of the two methods can be found in Gillard [9].

We note that the problem (3) corresponds to a special choice of the weight vector $\mathbf{w}$; however, our purpose is to develop methods for arbitrary given weights. We will need the following generalized projections. For two positive vectors $\mathbf{p} \in \mathbb{R}^L$ and $\mathbf{q} \in \mathbb{R}^K$, define

$$P := \text{diag}(\sqrt{\mathbf{p}}) \qquad \text{and} \qquad Q := \text{diag}(\sqrt{\mathbf{q}})$$

Note that the square root was used to define the diagonal matrices P and Q (this will simplify our future computation). We let $\Pi_\mathcal{H}^{(\mathbf{p},\mathbf{q})}(A)$ denote the optimal solution of the following problem

$$\min \ \ \|P(X - A)Q\|, \qquad \text{s.t.} \ \ X \in \mathcal{H}.$$

Due to the special structure of Hankel matrices, $\Pi_\mathcal{H}^{(\mathbf{p},\mathbf{q})}(A)$ has a closed-form solution, which is given by (see [28, Prop. 2])

$$\Pi_\mathcal{H}^{(\mathbf{p},\mathbf{q})}(A) = \mathcal{T}(\mathbf{a}), \tag{9}$$

where $\mathbf{a} \in \mathbb{R}^N$ is given by

$$a_i = \frac{\sum_{\ell+k=i+1} (p_\ell q_k) a_{ij}}{\sum_{\ell+k=i+1} (p_\ell q_k)}, \qquad i = 1, 2, \ldots, N.$$

It is easy to see that when $\mathbf{p} = \mathbf{1} \in \mathbb{R}^L$ and $\mathbf{q} = \mathbf{1} \in \mathbb{R}^K$, $\Pi_\mathcal{H}^{(\mathbf{p},\mathbf{q})}(A)$ reduces to $\Pi_\mathcal{H}(A)$ obtained via anti-diagonally averaging A. We further let $\Pi_{\mathcal{M}_r}^{(\mathbf{p},\mathbf{q})}(A)$ denote an optimal solution of the problem

$$\min \ \ \|P(X - A)Q\|, \qquad \text{s.t.} \ \ X \in \mathcal{M}_r.$$

It follows from [11, Theorem 2] that $\Pi_{\mathcal{M}_r}^{(\mathbf{p},\mathbf{q})}(A)$ can also be obtained through SVD of (PAQ) and is given by

$$\Pi_{\mathcal{M}_r}^{(\mathbf{p},\mathbf{q})}(A) = P^{-1}\left(\Pi_{\mathcal{M}_r}(PAQ)\right) Q^{-1}. \tag{10}$$

We finish this section with a brief description of the well-known `Cadzow`(α) method [28]. Suppose $h := N/L$ is an integer. For a given $\alpha \geq 0$, we define $\mathbf{q} := \mathbf{q}(\alpha) \in \mathbb{R}^K$ by

$$q_i(\alpha) = \begin{cases} 1 & \text{if } i = jL + 1 \text{ for } j = 0, \cdots, h-1 \\ \alpha & \text{otherwise.} \end{cases}$$

For this special case, it is known ([10, Lemma 1]) that the problem (2) with $w_i \equiv 1$ is equivalent to the problem (5) with $C = \text{diag}(\mathbf{q}(0))$. For a given $\alpha > 0$, the `Cadzow`(α) method [28, Alg. 5] is as follows. Start with $X^0 = Y$, $\mathbf{p} = \mathbf{1}_L$ and $\mathbf{q} = \mathbf{q}(\alpha)$, compute

$$X^{k+1} = \Pi_\mathcal{H}^{(\mathbf{p},\mathbf{q})}\Big(\Pi_{\mathcal{M}_r}^{(\mathbf{p},\mathbf{q})}(X^k)\Big), \quad k = 0, 1, \ldots.$$

This method with $\alpha = 0.2$ will be tested for the data of fortified wine sales ($N = 168, L = 84$) in Sect. 5.3.

3. SEQUENTIAL APPROXIMATIONS VIA MAJORIZATION

In this section, we will describe our new approximation scheme and draw connections whenever possible to that studied in [11, 28], which also handle arbitrarily given weights in (2).

3.1 Reformulation as Hankel matrix optimization

Our first step is to reformulate (2) as an optimization of Hankel matrix. For a given weight vector $\mathbf{w} \in \mathbb{R}^N$ with $w_i \geq 0$, we let $\sqrt{\mathbf{w}}$ denote the componentwise square-root vector, whose elements are $\sqrt{w_i}$, $i = 1, \ldots, N$. We also define the vector $\mathbf{v} \in \mathbb{R}^N$ by (assuming $L \leq K$)

$$v_i := \begin{cases} 1/i & \text{for } i = 1, \ldots, L-1 \\ 1/L & \text{for } i = L, \ldots, K \\ 1/(N-i+1) & \text{for } i = K+1, \ldots, N. \end{cases}$$

Let

$$M := \mathcal{T}(\sqrt{\mathbf{v}} \circ \sqrt{\mathbf{w}}). \tag{11}$$

It is straightforward to verify that the weighted problem (2) is equivalent to

$$\min\ f(X) := \frac{1}{2}\|M\circ(X-Y)\|^2, \qquad \text{s.t. } X \in \mathcal{H}\cap\mathcal{M}_r. \tag{12}$$

It is noted that the equivalence between (2) and (12) is also a direct consequence of [28, Propostion 4] and that there are other choices for $\mathbf{v}$ which serve the same purpose of establishing the equivalence between (2) and (12). We further note that the problem (12) is extremely difficult to solve unless the weight vector $\mathbf{w}$ has certain special properties. Unfortunately, even the most common choice of $\mathbf{w} = \mathbf{1}$ (the vector of ones) would render the problem computationally intractable to get its optimal solution. Hence, approximations appear to be an effective approach for (12).

Zvonarev and Golyandina [28] proposed to approximate $f(X)$ in (12) by

$$\begin{aligned} \min\ f_{\mathtt{ZG}}(X) &= \frac{1}{2}\mathrm{Tr}\Big((X-Y)C(X-Y)^T\Big) \\ &= \|X-Y\|_C^2, \qquad \text{s.t. } X \in \mathcal{H}\cap\mathcal{M}_r, \end{aligned}$$

where C is a $K\times K$ positive definite matrix. We refer to the norm $\|X\|_C$ as the C-norm of X. A particular case is when C is diagonal where $C := \mathrm{diag}(\widehat{\mathbf{c}})$ with $\widehat{\mathbf{c}} \in \mathbb{R}^K$ being defined as the column average of the matrix $(M\circ M)$:

$$\widehat{c}_j := \frac{1}{L}\sum_{i=1}^{L} m_{ij}^2, \qquad j = 1,\ldots,K. \tag{13}$$

With this particular choice of $\widehat{\mathbf{c}}$, the corresponding weighted Cadzow method of (8) is called the $\mathtt{Cadzow}$-$\widehat{C}$ method in [28]. Another choice is $C := \mathrm{diag}(\mathbf{c}(\alpha))$ with $\alpha > 0$ and $\mathbf{c}(\alpha) \in \mathbb{R}^K$ being defined in a special manner, see [28, Eq. 20]. This choice results in the $\mathtt{Cadzow}(\alpha)$ method in [28].

Gillard and Zhigljavsky [11] aim to provide a closer approximation to (12) by the (Q,R)-norm problem (4), where the diagonal matrices $Q := \mathrm{diag}(\mathbf{q})$ and $R := \mathrm{diag}(\mathbf{r})$ are selected through the following optimization:

$$\min\ \ \|\mathbf{w} - \mathbf{q}\star\mathbf{r}\|^2, \qquad \text{s.t. } \mathbf{q}\in\mathbb{R}^L,\ \ \mathbf{r}\in\mathbb{R}^K, \tag{14}$$

where $\mathbf{q}\star\mathbf{r}$ is their convolution, see [11, Theorem 1]. The optimization problem (14) is usually hard to solve due to the nonlinearity of the convolution operator. We refer to (4) as the (Q,R)-norm method.

3.2 Approximation via majorization

The purpose of this part is to develop a new approximation which is easier to solve by existing methods such as Cadzow's method. Suppose we have a pair of positive vectors $\mathbf{p}\in\mathbb{R}^L$ and $\mathbf{q}\in\mathbb{R}^K$ satisfying the property:

$$\|M\circ X\| \le \|(\sqrt{\mathbf{p}}\sqrt{\mathbf{q}}^T)\circ X\| \qquad \text{for all } \ X\in\mathcal{M}, \tag{15}$$

where the matrix M is defined by (11). There are many such pairs $(\mathbf{p},\mathbf{q})$. We will address the issue of choosing $\mathbf{p}$ and $\mathbf{q}$ in the next section. Both sides of (15) can be thought as weighted norms of X. The difference is that the weighted norm on the right-hand side is defined by a rank-one matrix $(\mathbf{p}\mathbf{q}^T)$ and the optimization problem with respect to this weighted norm is easier to solve than that given by the matrix M on the left-hand side. Moreover, the weighted norm on the right is always above that on the left. This will yield the important majorization problem below. Define the new function for all $X, Z\in\mathcal{M}$

$$\begin{aligned} f_m(X,Z) := \ & f(Z) + \langle \nabla f(Z),\ X-Z\rangle \\ & + \tfrac{1}{2}\|(\sqrt{\mathbf{p}}\sqrt{\mathbf{q}}^T)\circ(X-Z)\|^2. \end{aligned} \tag{16}$$

We have the following properties

Proposition 3.1. *Consider the problem (12). It holds that*

$$f_m(X,X) = f(X) \qquad \forall\ X\in\mathcal{M} \tag{17}$$

and

$$f(X) \le f_m(X,Z) \qquad \forall\ X,Z\in\mathcal{M}. \tag{18}$$

Moreover, we have for any given $Z\in\mathcal{M}$

$$\arg\min_{X\in\mathcal{M}}\ f_m(X,Z) = \arg\min_{X\in\mathcal{M}}\ \frac{1}{2}\|P(X-\Delta_Z)Q\|^2, \tag{19}$$

where $P := \mathrm{diag}(\sqrt{\mathbf{p}})$ *and* $Q := \mathrm{diag}(\sqrt{\mathbf{q}})$*, and the matrix* $\Delta_Z\in\mathcal{M}$ *is given by*

$$\Delta_Z := Z - P^{-2}(M\circ M\circ(Z-Y))Q^{-2}. \tag{20}$$

Proof. The equality (17) is obvious. As for the inequality (18), we have

$$\begin{aligned} f_m(X,Z) &= f(Z) + \langle\nabla f(Z),\ X-Z\rangle + \frac{1}{2}\|M\circ(X-Z)\|^2 \\ &\quad + \underbrace{\frac{1}{2}\|(\sqrt{\mathbf{p}}\sqrt{\mathbf{q}}^T)\circ(X-Z)\|^2 - \frac{1}{2}\|M\circ(X-Z)\|^2}_{\ge 0 \ \text{ because of (15)}} \\ &\ge f(Z) + \langle\nabla f(Z),\ X-Z\rangle + \frac{1}{2}\|M\circ(X-Z)\|^2 \\ &= f(X). \end{aligned}$$

The last equality holds because $f(\cdot)$ is quadratic and its second order Taylor expansion is exact.

We now prove (19). We note that

$$\|(\sqrt{\mathbf{p}}\sqrt{\mathbf{q}}^T)\circ X\| = \|PXQ\|, \qquad \forall\ X\in\mathcal{M},$$

Using this observation, we have

$$\begin{aligned}
& \arg\min \ f_m(X,Z) = \arg\min_{X\in\mathcal{M}} \langle \nabla f(Z),\ X-Z\rangle \\
& +\frac{1}{2}\|(\sqrt{\mathbf{p}}\sqrt{\mathbf{q}}^T)\circ(X-Z)\|^2 \\
= \ & \arg\min_{X\in\mathcal{M}} \langle M\circ M\circ(Z-Y),\ X-Z\rangle \\
& +\frac{1}{2}\|P(X-Z)Q\|^2 \\
= \ & \arg\min_{X\in\mathcal{M}} \ \frac{1}{2}\|P(X-Z)Q \\
& +P^{-1}(M\circ M\circ(Z-Y))Q^{-1}\|^2 \\
= \ & \arg\min_{X\in\mathcal{M}} \frac{1}{2}\|P(X-\Delta_Z)Q\|^2,
\end{aligned}$$

where

$$\Delta_Z := Z - P^{-2}(M\circ M\circ(Z-Y))Q^{-2}.$$

This proved the claim in (19). □

Because of the properties in Proposition 3.1, $f_m(Z,X)$ is known as a majorization of $f(X)$ at a given point. The computational implication is that we may minimize this new function instead of the original function provided that the new function is easier to minimize. This approximation procedure shall be repeated until convergence is observed. We formally describe this computational procedure below.

Let X^k be the current iterate. We try to find the next iterate by

$$(21)\quad X^{k+1} \in \arg\min \ f_m(X,X^k), \qquad \text{s.t.} \ \ X\in\mathcal{H}\cap\mathcal{M}_r.$$

We note that the problem may have multiple solutions. Suppose $X^k\in\mathcal{H}\cap\mathcal{M}_r$. We immediately have

$$(22)\quad f(X^{k+1}) \le f_m(X^{k+1},X^k) \le f_m(X^k,X^k) = f(X^k),$$

where the first inequality follows from (18); the second inequality is because of (21); and the last equality follows from (17). This means that the majorization procedure generates a sequence $\{X^k\}$ with decreasing objective function values in the original function $f(X)$. The inequality in (22) is known as the sandwich inequality [7]. We also note that $f_m(X,X^k)$ makes use of the latest gradient information $\nabla f(X^k)$ and hence provides a rather accurate (local) approximation of $f(X)$ at X^k. We make further comments below.

(**R1**) The first comment is that our subproblem (21) is the type of (Q,R)-norm problem (4) studied in [11]. It follows from (19) that the optimization problem (21) is equivalent to

$$(23)\qquad \min \frac{1}{2}\|P(X-\Delta^k)Q\|^2, \qquad \text{s.t.} \ \ X\in\mathcal{H}\cap\mathcal{M}_r,$$

where Δ^k is obtained from (20) by replacing Z with X^k:

$$\Delta^k := X^k - P^{-2}(M\circ M\circ(X^k-Y))Q^{-2}.$$

We further note that

$$\begin{aligned}
\|P(X-\Delta^k)Q\|^2 &= \mathrm{Tr}\Big(P(X-\Delta^k)QQ(X^k-\Delta^k)^TP\Big) \\
&= \mathrm{Tr}\Big(P^2(X-\Delta^k)Q^2(X-\Delta^k)^T\Big) \\
&= \|X-\Delta^k\|^2_{(P^2,Q^2)},
\end{aligned}$$

which is exactly the type of the (Q,R)-norm defined in [11, Eq. (9)] (see also (4)). Therefore, our subproblem is the type of the approximation problem considered in [11]. Consequently, the method of alternating projection proposed in [11]) (e.g., Cadzow's method) can be used to solve (23). The essential difference from [11] is how we have derived the sequential approximations by defining Δ^k.

(**R2**) The second comment is that our approximation also includes the problems (3) and (4) as special cases. This follows from choosing $M=E$ (the matrix of all ones) and $\mathbf{p}=\mathbf{1}\in\mathbb{R}^L$, $\mathbf{q}=\mathbf{1}\in\mathbb{R}^K$. This immediately suggests that we may use the diagonally weighted version of the SSA method or Cadzow's method to solve the subproblem (23), with the projections $\Pi_{\mathcal{H}}(\cdot)$ and $\Pi_{\mathcal{M}_r}(\cdot)$ in (7) and (8) being respectively replaced by $\Pi^{(\mathbf{p},\mathbf{q})}_{\mathcal{H}}(\cdot)$ in (9) and $\Pi^{(\mathbf{p},\mathbf{q})}_{\mathcal{M}_r}(\cdot)$ in (10). We will address the implementation issue in the numerical part. Moreover, if we choose $M=\mathbf{p}\mathbf{q}^T$, then the subproblem (23) reduces to

$$\min \ \frac{1}{2}\|P(X-Y)Q\|^2, \qquad \text{s.t.} \ \ X\in\mathcal{H}\cap\mathcal{M}_r,$$

which is exactly the approximation problem of (4) considered in [11].

In summary, we developed a computational scheme for the weighted problem (2). The scheme amounts to solving a sequence of approximation problems of (23). It includes both (3) and (4) as special cases and it is essentially different from that of [11, 28] in the way how Δ^k is being defined. We put the scheme in the following algorithmic framework. We call it Sequential Majorization Method (SMM).

Algorithm 3.2. *(Algorithm SMM)*

(S.1) Initialization: Given time series $\mathbf{x}\in\mathbb{R}^n$ *and the weight vector* $\mathbf{w}\in\mathbb{R}^N$. *Choose the window length* L. *Compute* $Y:=\mathcal{T}(\mathbf{x})$ *and matrix* M *by (11). Start with* $X^0=Y$ *and set* $k:=0$.

(S.2) Compute the vectors $(\mathbf{p},\mathbf{q})$ *to satisfy the inequality (15).*

(S.3) Compute the next iterate X^{k+1}, *which approximately solves (21):*

$$
\begin{aligned}
X^{k+1} &\approx \arg\min\{f_m(X, X^k),\ \ s.t.\ \ X \in \mathcal{H} \cap \mathcal{M}_r\} \\
&= \arg\min\Big\{\frac{1}{2}\|P(X-\Delta^k)Q\|^2, \\
&\qquad\qquad s.t.\ \ X \in \mathcal{H} \cap \mathcal{M}_r\}
\end{aligned}
$$

with Δ^k being defined by

$$\Delta^k := X^k - P^{-2}\Big(M \circ M \circ (X^k - Y)\Big)Q^{-2},$$

where $P := \mathrm{diag}(\sqrt{\mathbf{p}})$ and $Q := \mathrm{diag}(\sqrt{\mathbf{q}})$.

(S.4) Repeat (S.3) until certain stopping criterion is met.

We state a convergence result for Alg. 3.2.

Theorem 3.3. *(Convergence in terms of the objective function) Suppose Alg. 3.2 generates a sequence of $\{X^k\}$, $k = 1, 2, \ldots$. We also assume that the approximate solution X^{k+1} in (S.3) satisfies*

$$f_m(X^{k+1}, X^k) \le f_m(X^k, X^k). \tag{24}$$

Then the objective function sequence $\{f(X^k)\}$ is nonincreasing and converges.

Proof. The proof is simple. If the condition (24) is satisfied by X^k and X^{k+1}, then the sandwich inequality (22) automatically holds because of the reasons given below (22). This shows that the whole sequence of $\{f(X^k)\}$ is nonincreasing. Since the function $f(X)$ is bounded below by 0 and the sequence $\{f(X^k)\}$ is nonincreasing, it must converge. $\square$

We emphasize that in order for the sequence $\{f(X^k)\}$ to be nonincreasing, we do not have to require X^{k+1} be the optimal solution of (21), which would be too demanding because the problem is nonconvex. Fortunately, it is enough to just compute a point X^{k+1} satisfying (24). As long as this condition holds, the sandwich inequality will hold and the functional value sequence $\{f(X^k)\}$ will be nonincreasing. Note that this is true even we do not require X^k to be in the set of $\mathcal{H} \cap \mathcal{M}_r$. This observation is important to our practical computation and it provides a justification for using the Cadzow method (8), which does not enforce the constraint be strictly satisfied. We now address how to compute $(\mathbf{p}, \mathbf{q})$ in (S.2) in the next section. We will leave (S.4) to our numerical part.

4. IMPROVING THE APPROXIMATION

The quality of our approximation is governed by the inequality (15), which relies on the two vectors $\mathbf{p}$ and $\mathbf{q}$. A measurement of quality of the approximation is that the tighter the inequality is, the better the approximation would be. In this section, we propose two schemes for generating a pair of $(\mathbf{p}, \mathbf{q})$. The rationale behind the two schemes is that the computation should be very fast. As a matter of fact, both can be done by solving a linear programming problem.

4.1 A simple choice of $(\overline{\mathbf{p}}, \overline{\mathbf{q}})$ and its improvement

A particular choice of the pair $(\mathbf{p}, \mathbf{q})$ satisfying (15) is as follows. We denote the pair by $(\overline{\mathbf{p}}, \overline{\mathbf{q}})$:

$$
\left\{
\begin{array}{lcl}
\overline{p}_\ell & := & \max\{m_{\ell k} \mid k = 1, \ldots, K\},\ \ \ell = 1, \ldots, L \\
\overline{q}_k & := & \max\{m_{\ell k} \mid \ell = 1, \ldots, L\},\ \ k = 1, \ldots, K.
\end{array}
\right. \tag{25}
$$

It is easy to check that the choice of $(\overline{\mathbf{p}}, \overline{\mathbf{q}})$ satisfies (15) by referring to the M matrix (11).

Our purpose below is to reduce $(\overline{\mathbf{p}}, \overline{\mathbf{q}})$ as much as we can under the constraint that it still satisfies (15). We note that a necessary and sufficient condition for $(\mathbf{p}, \mathbf{q})$ to satisfy (15) is

$$\sum_{\ell=1}^{L} \sum_{k=1:\ \ell+k=i+1}^{K} (p_\ell q_k) \ge w_i, \quad i = 1, \ldots, N. \tag{26}$$

Suppose $(\mathbf{p}, \mathbf{q})$ takes the following form:

$$\mathbf{p} := \overline{\mathbf{p}} - \mathbf{s}, \quad \mathbf{q} := \overline{\mathbf{q}} - \mathbf{t}, \quad 0 \le \mathbf{s} \le \overline{\mathbf{p}}, \quad 0 \le \mathbf{t} \le \overline{\mathbf{q}}. \tag{27}$$

From all such representations, we would like to find a best pair that minimizes the following problem:

$$
\begin{array}{ll}
\min & \sum_{\ell=1}^{L}\sum_{k=1}^{K} p_\ell q_k \\
\text{s.t.} & \text{Constraints in (26).}
\end{array} \tag{28}
$$

It follows from (27) that

$$
\begin{aligned}
p_\ell q_k &= (\overline{p}_\ell - s_\ell)(\overline{q}_k - t_k) \\
&= \overline{p}_\ell \overline{q}_k - (\overline{p}_\ell t_k + \overline{q}_k s_\ell) + s_\ell t_k \\
&\ge \overline{p}_\ell \overline{q}_k - (\overline{p}_\ell t_k + \overline{q}_k s_\ell),
\end{aligned}
$$

which implies

$$\sum_{\ell+k=i+1} (p_\ell q_k) - w_i \ge \sum_{\ell+k=i+1} \overline{p}_\ell \overline{q}_k - \sum_{\ell+k=i+1} (\overline{p}_\ell t_k + \overline{q}_k s_\ell) - w_i.$$

Hence, the constraints in (28) are satisfied if

$$\sum_{\ell+k=i+1} (\overline{p}_\ell t_k + \overline{q}_k s_\ell) \le \sum_{\ell+k=i+1} \overline{p}_\ell \overline{q}_k - w_i,\ \ i = 1, \ldots, N. \tag{29}$$

On the other hand, we have for the objective that

$$
\begin{aligned}
\sum_{\ell=1}^{L}\sum_{k=1}^{K} p_\ell q_k &= \mathbf{1}_L^T(\mathbf{p}\mathbf{q}^T)\mathbf{1}_K \\
&= \mathbf{1}_L^T((\overline{\mathbf{p}} - \mathbf{s})(\overline{\mathbf{q}} - \mathbf{t})^T)\mathbf{1}_K \\
&= (\mathbf{1}_L^T\mathbf{s})(\mathbf{1}_K^T\mathbf{t}) - \big[(\mathbf{1}_L^T\mathbf{s})(\mathbf{1}_K^T\overline{\mathbf{q}}) + (\mathbf{1}_L^T\overline{\mathbf{p}})(\mathbf{1}_K^T\mathbf{t})\big] \\
&\quad + (\mathbf{1}_L^T\overline{\mathbf{p}})(\mathbf{1}_K^T\overline{\mathbf{q}}).
\end{aligned} \tag{30}
$$

Define

$$\alpha := \mathbf{1}_L^T \mathbf{s}, \quad \beta := \mathbf{1}_K^T \mathbf{t}, \quad \lambda := \mathbf{1}_L^T \overline{\mathbf{p}}, \quad \gamma := \mathbf{1}_L^K \overline{\mathbf{q}}.$$

The objective becomes

$$\sum_{\ell=1}^{L}\sum_{k=1}^{K} p_\ell q_k = \alpha\beta - (\alpha\gamma + \beta\lambda) + \lambda\gamma = (\lambda - \alpha)(\gamma - \beta),$$

and because of (27)

$$0 \le \alpha \le \lambda \qquad \text{and} \qquad 0 \le \beta \le \gamma.$$

It is easy to verify that

$$\text{(31)} \quad \alpha\beta - (\alpha\gamma + \beta\lambda) + \lambda\gamma \le \left(\sqrt{\lambda\gamma} - \frac{1}{2\sqrt{\lambda\gamma}}(\alpha\gamma + \beta\lambda)\right)^2.$$

We replace the objective function by the right-hand-side quantity of (31) and replace the constraints in (28) by (29) to derive a new optimization problem:

(32)

$$\begin{array}{ll} \min_{\alpha,\beta} & \left(\sqrt{\lambda\gamma} - \frac{1}{2\sqrt{\lambda\gamma}}(\alpha\gamma + \beta\lambda)\right)^2 \\ \text{s.t.} & \sum_{\ell+k=i+1}(\overline{p}_\ell t_k + \overline{q}_k s_\ell) \le \sum_{\ell+k=i+1} \overline{p}_\ell \overline{q}_k - w_i, \\ & i = 1, \ldots, N, \; 0 \le \mathbf{s} \le \overline{\mathbf{p}}, \quad 0 \le \mathbf{t} \le \overline{\mathbf{q}}. \end{array}$$

It follows from the inequality

$$\sqrt{\lambda\gamma} \ge \frac{1}{2\sqrt{\lambda\gamma}}(\alpha\gamma + \beta\lambda)$$

that the problem (32) is equivalent to

(33)

$$\begin{array}{ll} \max_{\mathbf{s},\mathbf{t}} & \alpha\gamma + \beta\gamma = (\mathbf{1}_L^T \mathbf{s})(\mathbf{1}_K^T \overline{\mathbf{q}}) + (\mathbf{1}_L^T \overline{\mathbf{p}})(\mathbf{1}_K^T \mathbf{t}) \\ \text{s.t.} & \sum_{\ell+k=i+1}(\overline{p}_\ell t_k + \overline{q}_k s_\ell) \le \sum_{\ell+k=i+1} \overline{p}_\ell \overline{q}_k - w_i, \\ & i = 1, \ldots, N, \; 0 \le \mathbf{s} \le \overline{\mathbf{p}}, \quad 0 \le \mathbf{t} \le \overline{\mathbf{q}}. \end{array}$$

If the optimal solution $\mathbf{s}$ (or $\mathbf{t}$) is too close to $\overline{\mathbf{p}}$ (or $\overline{\mathbf{q}}$), then the resulting $\mathbf{p}$ (or $\mathbf{q}$) in (27) would be very close to 0, which would cause numerical difficulties when we use the reciprocal of it in our algorithm (see (S.3) in Alg. 3.2). Therefore, we introduce a safeguarding parameter $0 < \rho < 1$ such that $\mathbf{s} \le \rho\overline{\mathbf{p}}$ and $\mathbf{t} \le \rho\overline{\mathbf{q}}$. The final problem that we aim to solve is as follows.

(34)

$$\begin{array}{ll} \max_{\mathbf{s},\mathbf{t}} & \alpha\gamma + \beta\gamma = (\mathbf{1}_L^T \mathbf{s})(\mathbf{1}_K^T \overline{\mathbf{q}}) + (\mathbf{1}_L^T \overline{\mathbf{p}})(\mathbf{1}_K^T \mathbf{t}) \\ \text{s.t.} & \sum_{\ell+k=i+1}(\overline{p}_\ell t_k + \overline{q}_k s_\ell) \le \sum_{\ell+k=i+1} \overline{p}_\ell \overline{q}_k - w_i, \\ & i = 1, \ldots, N, \; 0 \le \mathbf{s} \le \rho\overline{\mathbf{p}}, \quad 0 \le \mathbf{t} \le \rho\overline{\mathbf{q}}. \end{array}$$

The benefit of all those calculations is that the problem (34) is a linear programming problem and it can be efficiently solved by any standard linear programming solver. A side note is that the above technique leading to (33) is known as relaxation in optimization. We now formally state our algorithm for improving the pair $(\overline{\mathbf{p}}, \overline{\mathbf{q}})$.

Algorithm 4.1. *($LP(\overline{\mathbf{p}}, \overline{\mathbf{q}})$)*

(S.1) Input: a pair of positive vectors $(\overline{\mathbf{p}}, \overline{\mathbf{q}})$ satisfying (15); $0 < \rho < 1$, weight vector $\mathbf{w} \in \mathbb{R}^N$ and the window length L and $K = N - L + 1$.

(S.2) Use any standard Linear Programming solver to the problem (34) for the optimal $(\mathbf{s}, \mathbf{t})$.

(S.3) Output: Let

$$\mathbf{p} := \overline{\mathbf{p}} - \mathbf{s} \qquad \textit{and} \qquad \mathbf{q} := \overline{\mathbf{q}} - \mathbf{t}.$$

4.2 A choice of $\widehat{\text{c}}$ vector and its improvement

Another particular choice of $(\mathbf{p}, \mathbf{q})$ that gives rise to the $\widehat{C}$-norm proposed in [28] is when $\mathbf{p} := \mathbf{1}_L$ and $\mathbf{q} := \widehat{\mathbf{c}} \in \mathbb{R}^K$ that satisfies the inequality (15). For example, we can choose $\widehat{\mathbf{c}}$ by

$$\text{(35)} \quad \widehat{c}_k := \max\left\{ m_{\ell k}^2 \;\mid\; \ell = 1, \ldots, L \right\}, \qquad k = 1, \ldots, K.$$

Obviously, we have

$$\text{(36)} \quad \|M \circ X\|^2 \le \|X \mathrm{diag}(\sqrt{\widehat{\mathbf{c}}})\|^2 = \mathrm{Tr}(X\widehat{C}X^T) = \|X\|_{\widehat{C}}^2,$$

where $\widehat{C} := \mathrm{diag}(\widehat{\mathbf{c}})$. This is exactly the Cadzow $\widehat{C}$-norm used in [28]. The difference here is on the choice of $\widehat{C}$ for it to satisfy (36) (note that $\widehat{\mathbf{c}}$ in [28] was chosen by (13)).

Following the idea in Subsection 4.1, we can improve $\widehat{\mathbf{c}}$ so that the inequality (36) becomes tighter. Let

$$\mathbf{c} := \widehat{\mathbf{c}} - \mathbf{t}, \qquad \mathbf{0} \le \mathbf{t} \le \rho\widehat{\mathbf{c}},$$

where ρ is the proportion (i.e., $0 < \rho \le 1$) that $\widehat{\mathbf{c}}$ can be reduced. If ρ is close to 1, then some of the elements in the resulting $\mathbf{c}$ may be close to 0, causing numerical instability due to the same reasons as for the problem (34). Throughout our experiments in Sect. 5, we set $\rho = 0.6$.

Hence, we choose the best $\mathbf{c}$ via the linear programming problem:

$$\begin{array}{ll} \min_{\mathbf{t} \in \mathbb{R}^K} & \langle \mathbf{1}_K, \; \mathbf{c} \rangle = \sum_{k=1}^K \widehat{c}_k - \sum_{k=1}^K t_k \\ \text{s.t.} & \sum_{\ell=1}^L \sum_{k=1:\ell+k=i+1}^K c_k \ge w_i, \; i = 1, 2, \ldots, N \\ & \mathbf{0} \le \mathbf{t} \le \rho\widehat{\mathbf{c}}, \end{array}$$

which is equivalent to

(37)

$$\begin{array}{l} \min_{\mathbf{t} \in \mathbb{R}^K} \; -\sum_{k=1}^K t_k \\ \text{s.t.} \; \sum_{\ell=1}^L \sum_{k=1:\ell+k=i+1}^K t_k \le \sum_{\ell=1}^L \sum_{k=1:\ell+k=i+1}^K \widehat{c}_k - w_i \\ \qquad i = 1, 2, \ldots, N, \quad \mathbf{0} \le \mathbf{t} \le \rho\widehat{\mathbf{c}}. \end{array}$$

This problem is a simpler type of the linear programming problem (33). We also note that the problem (37) is derived without having to rely on any relaxations such as (30), (31) and (32). We formally call this improvement via (37) as $\mathrm{LP}(\widehat{\mathbf{c}})$, which is stated as follows.

Algorithm 4.2. *($LP(\widehat{\mathbf{c}})$)*

(S.1) Input: the weight vector $\mathbf{w} \in \mathbb{R}^N$, the proportion parameter $0 < \rho \le 1$, and the vector $\widehat{\mathbf{c}}$ by (13)
(S.2) Use any standard Linear Programming solver to the problem (37) for the optimal $\mathbf{t}$.
(S.3) Output: Let

$$\mathbf{c} := \widehat{\mathbf{c}} - \mathbf{t}.$$

4.3 Quality of the approximation of the weight vector

In this part, we demonstrate how good is the $(\mathbf{p}, \mathbf{q})$-approximation to a given weight vector $\mathbf{w} \in \mathbb{R}^N$. We consider the following type of the weight vector, which was extensively used by [11]:

$$\mathbf{w} = (w_1, w_2, \ldots, w_{N-m}, w_{N-m+1}, \ldots, w_N),$$

where $m \ge 0$ is a given integer and for a given $\beta \ge 1$,

$$w_i := \beta^i, \qquad i = 1, \ldots, N-m$$

and

$$w_i := w_{N-m} - \frac{w_{N-m}}{m+1}(i-(N-m)), \text{ for } i = N-m+1, \ldots, N.$$

Two particular choices of β that were used in [11] are $\beta = 1$ and $\beta = 1.01$. Following [11], we label the weight vector $\mathbf{w}$ from $\beta = 1$ by $\mathbf{w}_1$ and $\mathbf{w}_2$ for $\beta = 1.01$.

For a given pair $(\mathbf{p}, \mathbf{q})$, the corresponding weight vector $\widetilde{\mathbf{w}} \in \mathbb{R}^N$ is given by

$$\widetilde{w}_i := \sum_{\ell=1}^{L} \sum_{k=1:\ \ell+k=i+1}^{K} (p_\ell q_k), \qquad i = 1, \ldots, N.$$

Because of the majorization inequality (15), we must have

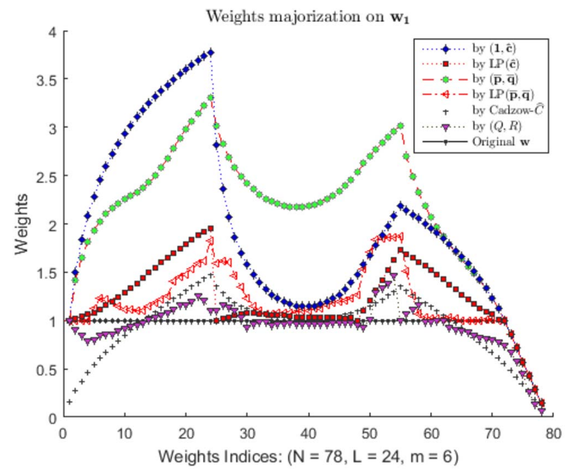
$$\widetilde{w}_i \ge w_i, \qquad i = 1, \ldots, N. \tag{38}$$

We consider the 4 pairs of $(\mathbf{p}, \mathbf{q})$ listed in Table 1. We also consider the choice $(\mathbf{1}, \widehat{\mathbf{c}})$ with $\widehat{\mathbf{c}}$ given by (13) used in [28]. The corresponding weights with the original ones are plotted in Fig. 1.

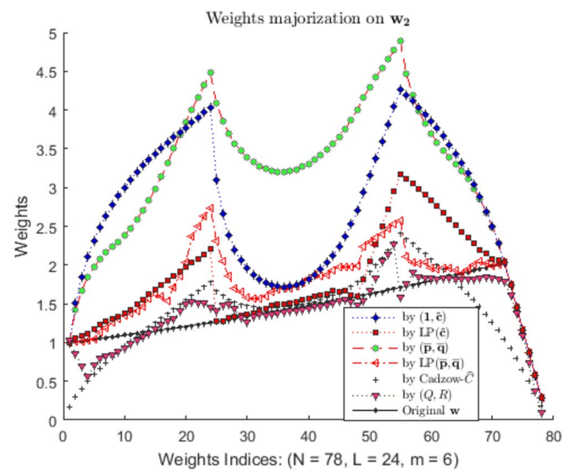
It is not surprising to see that all approximations from our 4 cases are above the original weights. In other words, the inequality (38) holds. However, for the choice (13) of [28] and the (Q, R) weights (14) of [11], the weights on both ends are below the original weights and the weights in the middle part are close to the original weights. Hence, the choice (13) or (14) does not give a majorized approximation. We would also like to point out that the approximation by $\mathrm{LP}(\overline{\mathbf{p}}, \overline{\mathbf{q}})$ closely follows the original weights. In theory, the closer the approximation is, the better the numerical performance should be provided that the subproblems can be solved globally. However, the subproblems are of a nonconvex nature. This echoes the need of global techniques in low-rank matrix approximation through optimization raised in a recent paper [5]. Therefore, different choice of $(\mathbf{p}, \mathbf{q})$ may have its own advantages depending on the actual applications. We will see the dependence in our numerical experiments below.

Table 1. Choices of $(\overline{\mathbf{p}}, \overline{\mathbf{q}})$

Case 1:	$(\overline{\mathbf{p}}, \overline{\mathbf{q}})$ by (25)
Case 2:	$(\mathbf{p}, \mathbf{q})$ by $\mathrm{LP}(\overline{\mathbf{p}}, \overline{\mathbf{q}})$
Case 3:	$(\mathbf{1}, \widehat{\mathbf{c}})$ with $\widehat{\mathbf{c}}$ given by (35)
Case 4:	$(\mathbf{1}, \mathbf{c})$ with $\mathbf{c}$ given by $\mathrm{LP}(\widehat{\mathbf{c}})$ with $\rho = 1$

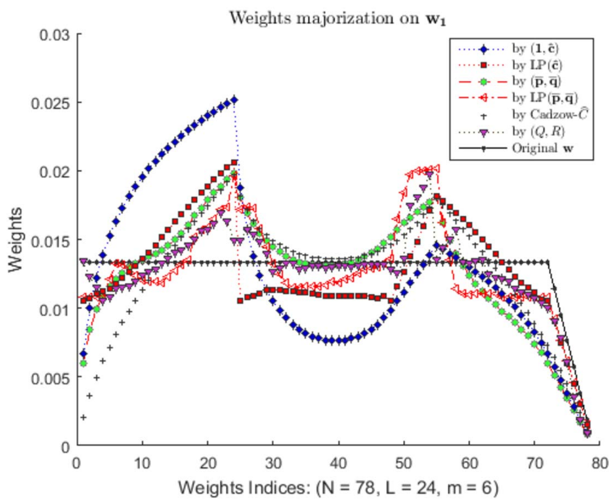


(a) Six weight vector approximations to $\mathbf{w_1}$

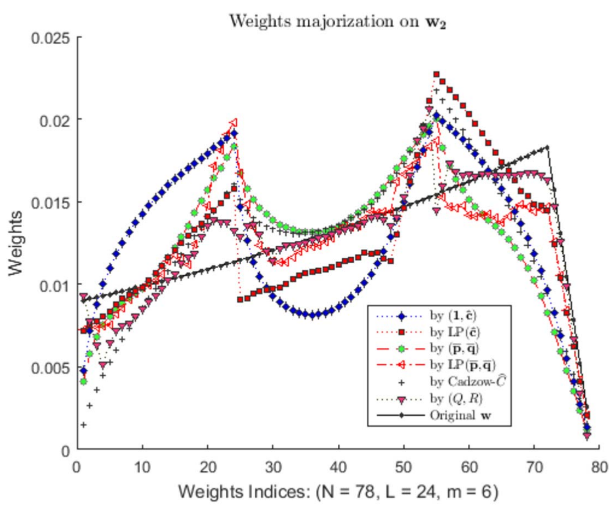


(b) Six weight vector approximations to $\mathbf{w_2}$

Figure 1. Six approximations of the weight vector $\mathbf{w}_1$ in Fig. 1(a) and of $\mathbf{w}_2$ in Fig. 1(b).



(a) Six weight vector approximations to normalized $\mathbf{w}_1$



(b) Six weight vector approximations to normalized $\mathbf{w}_2$

Figure 2. Six approximations of the normalized weight vector for $\mathbf{w}_1$ in Fig. 2(a) and for $\mathbf{w}_2$ in Fig. 2(b).

As in [28], we also plot the normalized weights (i.e., $\widetilde{w}_i / \sum \widetilde{w}_i$) for all the six choices and the normalized original weights in Fig. 2. As can be seen from Fig. 2, the normalized weights $\widetilde{w}_i / \sum \widetilde{w}_i$ are not necessarily above the corresponding normalized weights $w_i / \sum w_i$. The closer the normalized weights are to the original line, the less fluctuation the weight sequences behave (compare Fig. 1 with Fig. 2). In particular, the normalized (Q, R)-weights by (14) closely follow the original weights. This is not surprising because it is obtained by the overall least-square approximation without having to force any constraints (e.g., majorization constraint). It is an interesting question whether we can improve the majorized weights so that their normalized counterparts follow closely the line of the normalized original weights.

We finish this section by pointing out that there are other ways to generate $(\mathbf{p}, \mathbf{q})$. For example, the inequality (15) that $(\mathbf{p}, \mathbf{q})$ has to satisfy can be cast as a rank-one nonnegative matrix factorization such that

$$M \approx \sqrt{\mathbf{p}}\sqrt{\mathbf{q}}^T$$

and $(\mathbf{p}, \mathbf{q})$ satisfies the inequalities in (26). Nonnegative matrix factorization has many applications and hence has many algorithms (see [20, 21] and the references therein).

5. NUMERICAL EXPERIMENTS

In this part, we report our preliminary numerical results on two popular test problems of real life: the USA death data [13] and the fortified wine sales data [18] with comparison to the two major methods (Q, R)-norm approximation [11] and `Cadzow`(α) method in [28]. The main reason for choosing the two test problems is that they each was well studied in [11] and [28] respectively and hence they provide a good base for our comprehensive comparison with the two main methods. We first describe the implementation issues of our Alg. 3.2.

5.1 Solving the subproblem

The major computational part in Alg. 3.2 is on solving its subproblem (23) in (S.3). We propose to use the Cadzow method for the subproblem. For easy reference, we name the resulting method as `SMM-Cadzow` method, which runs as follows: Start with X^0, $k := 0$, compute the next iterate X^{k+1} as the final iterate of the following iterative procedure (Cadzow's method applied to the subproblem (23)):

$$(39) \quad \widetilde{X}^{j+1} = \Pi^{(\mathbf{p},\mathbf{q})}_{\mathcal{H}} \Big(\Pi^{(\mathbf{p},\mathbf{q})}_{\mathcal{M}_r} (\widetilde{X}^j) \Big), \quad \widetilde{X}^0 := \Delta^k, \quad j = 0, 1,$$

We terminate (39) if the following conditions are met

$$\frac{|f^k(\widetilde{X}^{j+1}) - f^k(\widetilde{X}^j)|}{\max\{1, f^k(\widetilde{X}^j\}} \le \mathtt{tol} \quad \text{or} \quad \frac{\|\widetilde{X}^{j+1} - \widetilde{X}^j\|}{\|\widetilde{X}^j\|} \le \mathtt{tol},$$

where `tol` is the tolerance level set by the user. In this part, we used $\mathtt{tol} = 10^{-3}$. We terminate Alg. 3.2 if

$$\frac{|f(X^{k+1} - f(X^k)|}{\max\{1, f(X^k)\}} \le \mathtt{tol} \quad \text{or} \quad \frac{\|X^{k+1} - X^k\|}{\|X^k\|} \le \mathtt{tol}.$$

The convergence of Alg. 3.2 relies on the sandwich inequality (22), which essentially requires computing $X^{k+1} \in \mathcal{H}$ that satisfies (24). However, the subproblem (23) is non-convex and Cadzow's method is not of global method, the sufficient condition (24) is not theoretically guaranteed. Surprisingly, when Cadzow's method is used to the subproblem (23), the inequality (24) is often observed. We will demonstrate this feature when we come to reporting the numerical results.

5.2 Example: USA death time series

This is a widely test data [13] and comprehensive results based on the two weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ in Subsect. 4.3 were reported in [11], which provides a basis for our comparison. The data set can be easily obtained online and contains the monthly accidental deaths in the USA between 1973 and 1978. The time series contains a total of $N = 78$ observations. Our task is to use the first 72 data points to forecast the remaining 6 observations (hence $m = 6$ in defining $\mathbf{w}_1$ and $\mathbf{w}_2$ in Subsect. 4.3). We will use the same parameters as those given in [17, 11] and follow the suggestion in [11] that there are several forecasting available to start with for further improvement. Those forecasts (for the last 6 data points) as well as their corresponding methods are included in Table 2, which is copied from [11] and includes 5 forecasting methods. In particular, Model I and Model II are examples of SARIMA models as described in [2]. HWS represents the model as fitted by the Holt-Winter seasonal algorithm. ARAR represents the model as fitted by transforming the data prior to fitting an autoregressive model. The forecasted values by `SSA` are taken from [17]. More details about those 5 methods can be found in [11, Sect. 7.3].

In our experiments below, we will use them as our initial guess for the last 6 data points. In other words, we have a data series $\mathbf{y} \in \mathbb{R}^N$ with the first 72 data being the first 72 data in the USA death time series, which is denoted as $\mathbf{x}^*$ and the last 6 points in $\mathbf{y}$ being one of the forecast values in Table 2. We use $\widehat{\mathbf{x}}$ to denote the obtained time series by Alg. 3.2. The Root of Mean Squared Error (`RMSE`) is then defined as

$$\texttt{RMSE} := \sqrt{\sum_{i=1}^{m} (\widehat{x}_{72+i} - x^*_{72+i})^2 / m},$$

which is often used to quantify how good the estimated values are to the original data. Obviously, the smaller the `RMSE` is, the better the forecasting is.

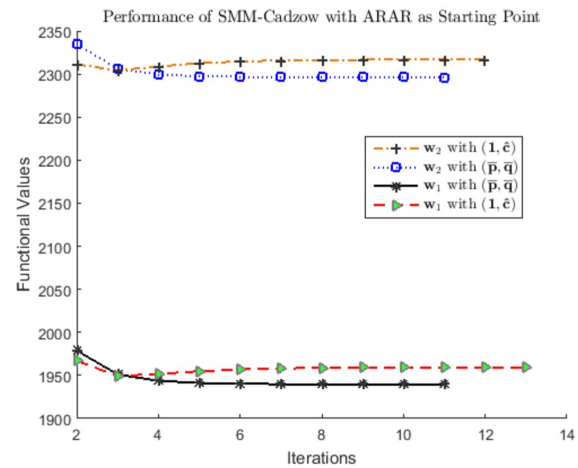
(a) Demonstration of convergence. We implemented Alg. 3.2 in `Matlab` and run it in `Matlab` 2015b. We would like to take this opportunity to demonstrate the convergence of Alg. 3.2 in terms of the objective values. For this experiment, we set `tol` $= 10^{-5}$ and the maximum number of iterations allowed for the subproblem is 100. The choice of this higher accuracy allows us to observe the trend of the objective values in many steps (lower accuracy would require a less number of iterations).

The functional sequence generated by our method are plotted in Fig. 3 from two starting points (ARAR and SSA respectively). `SMM-Cadzow` solved the subproblem iteratively by (39) and it can be observed in both figures in Fig. 3 that the functional sequence is decreasing and converges. This is because that the sufficient condition (24) is more often met than otherwise in `SMM-Cadzow`. This behaviour of convergence appears consistence for other test cases. Hence, we will not repeat this demonstration for the fortified wine sales data.

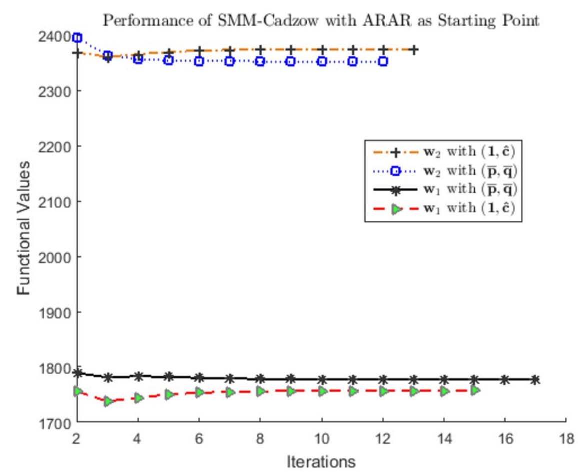
(b) RMSE comparison. We report the `RMSE` obtained by `SMM-Cadzow` with the 5 starting points from Table 2. The approximation weights $(\mathbf{p}, \mathbf{q})$ used are from Table 1. We will compare the obtained `RMSE` with those by the 5 start-

Table 2. Forecasts from five different models

Models	**1**	**2**	**3**	**4**	**5**	**6**
Original data	7798	7406	8363	8460	9217	9316
Model I	8441	7704	8549	8885	9843	10279
Model II	8345	7619	8356	8742	9795	10179
HWS	8039	7077	7750	7941	8824	9329
ARAR	8168	7196	7982	8284	9144	9465
SSA	7782	7428	7804	8081	9302	9333



(a) Convergence in the objective of `SMM-Cadzow`



(b) Convergence in the objective of `SMM-Cadzow`

Figure 3. Similar convergence was observed in both Fig. 3(a) and Fig. 3(b), in terms of functional values respectively starting from ARAR and SSA point in Table 2.

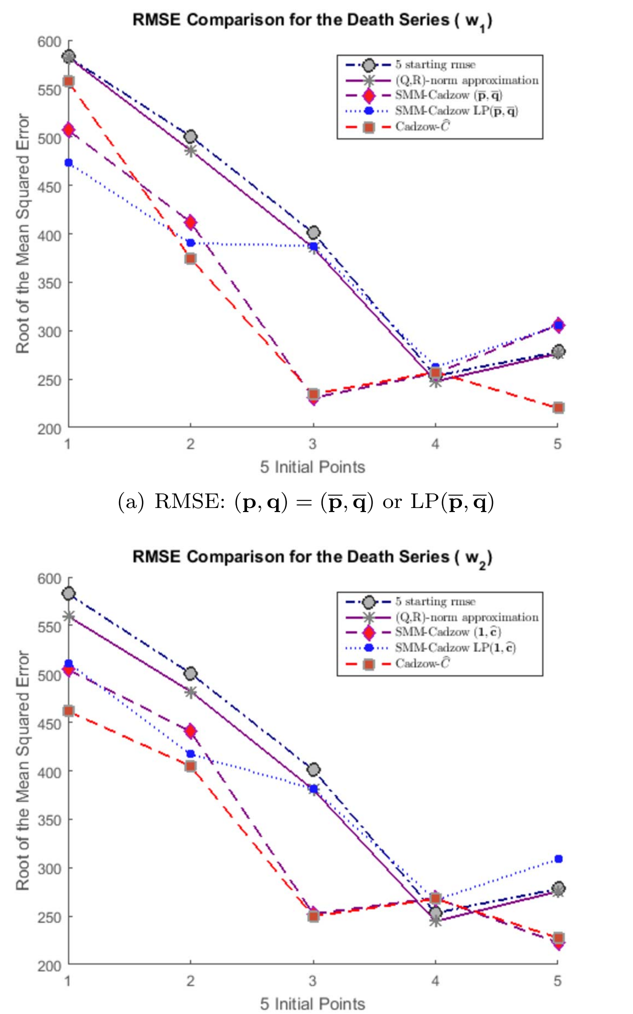
Table 3. RSME comparison between the method `SMM-Cadzow` and (Q,R)-norm approximation [11], Cadzow-$\widehat{C}$ method in [28]

β	$(\mathbf{p},\mathbf{q})$	5 starting points				
		Model I	Model II	HWS	ARAR	SSA
Initial	RMSE	582.63	500.5	401.26	253.20	278.20
$\beta = 1$ ($\mathbf{w}_1$)	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	507.11	411.98	**230.52**	256.18	305.78
	LP$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	**473.29**	390.22	387.43	262.53	305.41
	$(\mathbf{1},\hat{\mathbf{c}})$	490.79	395.64	425.15	266.32	329.73
	LP$(\hat{\mathbf{c}})$	488.37	389.35	410.13	281.85	324.14
	(Q,R)	582.20	486.03	385.81	**247.56**	276.28
	Cadzow-$\widehat{C}$	557.92	**374.32**	234.63	257.09	**220.11**
$\beta = 1.01$ ($\mathbf{w}_2$)	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	498.59	433.49	251.30	261.37	**218.13**
	LP$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	504.57	405.41	374.67	253.89	290.88
	$(\mathbf{1},\hat{\mathbf{c}})$	504.53	440.65	252.44	268.47	222.42
	LP$(\hat{\mathbf{c}})$	511.31	416.99	381.63	266.64	308.74
	(Q,R)	559.55	481.91	380.79	**244.61**	275.68
	Cadzow-$\widehat{C}$	**461.48**	**404.78**	**249.75**	267.71	227.56

ing points, the (Q,R)-norm approximation (its results are copied from [11]) and the Cadzow-$\widehat{C}$ method in [28] (based on our own implementation).

The RMSE results are reported in Table 3 for the two weight vectors $\mathbf{w}_1$ ($\beta = 1$) and $\mathbf{w}_2$ ($\beta = 1.01$). The first row (Initial RMSE) of the table includes the RMSE from the 5 initial points in Table 2. We first note that for many cases, there have been significant reductions in RMSE from each of the starting point. The numbers in bold indicate they are the best RMSE obtained from a given starting point by those methods. It can be observed that for both weighting schemes $\mathbf{w}_1$ and $\mathbf{w}_2$, the `Cadzow-`$\widehat{C}$ method of [28] worked very well. In particular, the `Cadzow-`$\widehat{C}$ achieved the 2 best RMSE for $\mathbf{w}_1$ and 3 best RMSE for $\mathbf{w}_2$. The `SMM-Cadzow` with $(\overline{\mathbf{p}},\overline{\mathbf{q}})$ closely followed the `Cadzow-`$\widehat{C}$ and achieved the overall best RMSE (218.13). This trend can be clearly seen in the Fig. 4, where we only plotted `SMM-Cadzow` with $(\overline{\mathbf{p}},\overline{\mathbf{q}})$ and LP$(\overline{\mathbf{p}},\overline{\mathbf{q}})$ for the case $\mathbf{w}_1$, and `SMM-Cadzow` with $(\mathbf{1},\hat{\mathbf{c}})$ and LP$(\mathbf{1},\hat{\mathbf{c}})$ for the case $\mathbf{w}_2$ for a better visualization. It can also be seen that the (Q,R)-norm approximation method of [11] did well for the ARAR starting point. For both $\mathbf{w}_1$ and $\mathbf{w}_2$, it achieved the best RMSE (247.56 and 244.61 respectively). Therefore, our purpose in the next part is to improve `SMM-Cadzow` in order to outperform both (Q,R)-norm method and the Cadzow-$\widehat{C}$ method.

The number of iterations for the above experiments are reported in Table 4, where `It` is the number of subproblems solved in `SMM` and `iter` is the total number of Cadzow iterations. For example, the first pair 4(77) in Table 4 means that for $\mathbf{w}_1$ and $(\overline{\mathbf{p}},\overline{\mathbf{q}})$, `SMM-Cadzow` solved 4 subproblems by a total of 77 iterations of (39). In our experiments, we set the maximum number of subproblems to be solved to 20. As we can see, there are a few cases where the maximum number (20) was reached and the numbers for `Iter` in some cases are large. Below we propose a strategy that would improve `SMM-Cadzow` both in quality of RMSE and in the number of total `iter`.



(a) RMSE: $(\mathbf{p},\mathbf{q}) = (\overline{\mathbf{p}},\overline{\mathbf{q}})$ or LP$(\overline{\mathbf{p}},\overline{\mathbf{q}})$

(b) RMSE: $(\mathbf{p},\mathbf{q}) = (\mathbf{1},\hat{\mathbf{c}})$ or LP$(\hat{\mathbf{c}})$

Figure 4. RMSE comparison in Fig. 4(a) for $\mathbf{w}_1$ and Fig. 4(b) for $\mathbf{w}_2$. Data were obtained by `SMM-Cadzow` with the 5 initial points given in Table 2.

Table 4. Number of iterations used by `SMM-Cadzow` starting with the 5 points from Table 2

β	$(\mathbf{p},\mathbf{q})$	5 starting points				
		Model I It (iter)	Model II It (iter)	HWS It (iter)	ARAR It (iter)	SSA It (iter)
	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	4 (77)	5 (97)	5 (89)	5 (85)	4 (80)
	LP$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	6 (117)	20 (360)	4 (72)	5 (91)	4 (80)
$\beta = 1$	$(\mathbf{1},\hat{\mathbf{c}})$	3 (59)	5 (97)	4 (79)	6 (98)	4 (80)
$(\mathbf{w}_1)$	LP$(\hat{\mathbf{c}})$	5 (95)	5 (95)	4 (80)	5 (92)	3 (60)
	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	4 (80)	5 (95)	5 (94)	5 (93)	5 (95)
	LP$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	20 (329)	20 (339)	5 (94)	20 (334)	4 (80)
$\beta = 1.01$	$(\mathbf{1},\hat{\mathbf{c}})$	5 (76)	5 (76)	5 (88)	5 (80)	6 (101)
$(\mathbf{w}_2)$	LP$(\hat{\mathbf{c}})$	20 (320)	20 (330)	13 (252)	20 (310)	20 (330)

Table 5. RSME comparison between the method `SMM-Cadzow` with the warmstart and (Q,R)-norm method [11], Cadzow-$\widehat{C}$ method in [28]

β	$(\mathbf{p},\mathbf{q})$	5 starting points				
		Model I	Model II	HWS	ARAR	SSA
Initial	RMSE	582.63	500.5	401.26	253.20	278.20
	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	507.11	411.98	**230.52**	256.18	305.78
	LP$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	**473.29**	390.22	387.43	262.53	305.41
$\beta = 1$	$(\mathbf{1},\hat{\mathbf{c}})$	490.79	395.64	425.15	266.32	329.73
$(\mathbf{w}_1)$	LP$(\hat{\mathbf{c}})$	488.37	389.35	410.13	281.85	324.14
	(Q,R)	582.20	486.03	385.81	**247.56**	276.28
	Cadzow-$\widehat{C}$	557.92	**374.32**	234.63	257.09	**220.11**
	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	498.59	433.49	251.30	261.37	**218.13**
	LP$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	504.57	405.41	374.67	253.89	290.88
$\beta = 1.01$	$(\mathbf{1},\hat{\mathbf{c}})$	504.53	440.65	252.44	268.47	222.42
$(\mathbf{w}_2)$	LP$(\hat{\mathbf{c}})$	511.31	416.99	381.63	266.64	308.74
	(Q,R)	559.55	481.91	380.79	**244.61**	275.68
	Cadzow-$\widehat{C}$	**461.48**	**404.78**	**249.75**	267.71	227.56

(c) Improving `SMM-Cadzow` via warmstart. It follows from both Table 3 and Table 4 that the `SMM-Cadozw` method has already shown its potential in finding the best RMSE (218.13) among all the methods tested. In this part, we will show that its quality and efficiency can be further improved by incorporating a warm start strategy, which is often used in a sequential optimization setting. For example, it has been successfully used in the sequential matrix optimization for computing a low-rank correlation matrix [8, 23]. We describe this simple strategy below.

At the kth iteration with X^k obtained, we compute

$$\widetilde{\mathbf{x}}^k := \mathcal{T}^{-1}(X^k)$$

and we replace the first 72 data points of $\widetilde{\mathbf{x}}^k$ by the original ones in $\mathbf{y}$ to get $\widehat{\mathbf{x}}^k$:

$$\widehat{\mathbf{x}}^k := \widetilde{\mathbf{x}}^k, \qquad \widehat{\mathbf{x}}^k(1:72) = \mathbf{y}(1:72).$$

Finally, we replace X^k by $\widehat{X}^k$:

$$\widehat{X}^k := \mathcal{T}(\widehat{\mathbf{x}}^k).$$

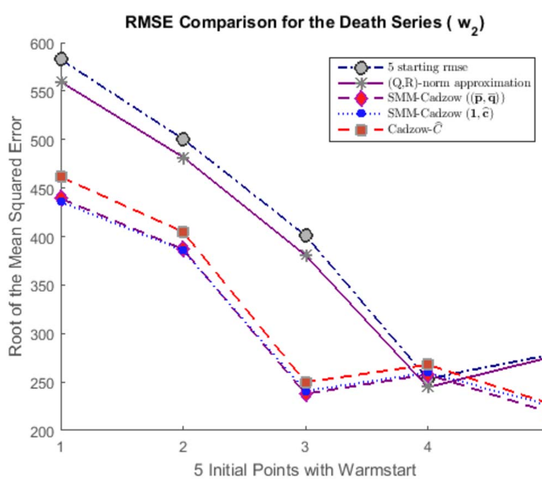
This warm start strategy uses the original 72 data (known) and the latest prediction for the last 6 missing values in $\mathbf{y}$ to define $\widehat{X}^k$. The subproblem is still solved by the Cadzow method (39). Hence, it is still convergent with this choice.

Under our stopping criterion (see Subsect. 5.1), only 2 subproblems were needed for each case. The results on RMSE and the iteration information are reported in Table 5 and Table 6, where `It` = 2 for all cases. It can be observed that not only the number of iterations (`iter`) has been reduced, the RMSE has also seen significant reduction. Moreover, 4 out of 5 best cases were obtained by our method for both $\mathbf{w}_1$ and $\mathbf{w}_2$. The remaining one has its RMSE (247.30 for ARAR starting point) not far from the best RMSE (244.61 for ARAR starting point by the (Q,R)-norm method). This is clearly demonstrated in Fig. 5(a), where the both lines for `SMM-Cadzow` with $(\overline{\mathbf{p}},\overline{\mathbf{q}})$ and $(\mathbf{1},\widehat{\mathbf{c}})$ are below the others.

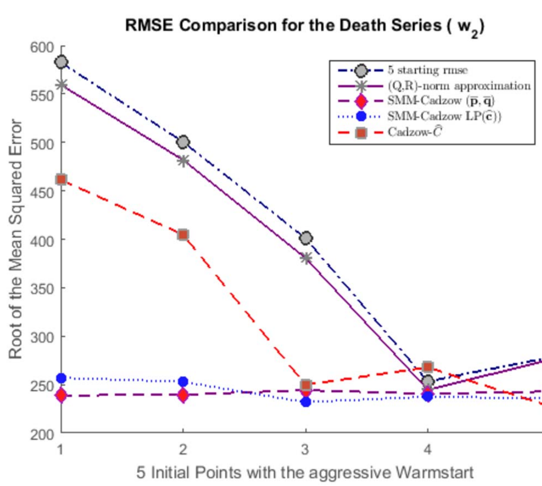
If we were even more "aggressive" in the sense that we apply the warm start strategy at every iteration. That is, we replace $\widetilde{X}^{j+1}$ in (39) by $\widehat{\widetilde{X}}^{j+1}$, we may achieve even more reduction in RMSE. Fig. 5(b) plotted two cases of `SMM-Cadzow` with $(\overline{\mathbf{p}},\overline{\mathbf{q}})$ and LP$(\widehat{\mathbf{c}})$. It is observed that both

Table 6. Number of iterations used by `SMM-Cadzow` starting with the 5 points from Table 2

β	$(\mathbf{p},\mathbf{q})$	5 starting points				
		Model I It (iter)	Model II It (iter)	HWS It (iter)	ARAR It (iter)	SSA It (iter)
	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	2 (34)	2 (34)	2 (34)	2 (34)	2 (40)
	$\mathrm{LP}(\overline{\mathbf{p}},\overline{\mathbf{q}})$	2 (34)	2 (32)	2 (30)	2 (36)	2 (40)
$\beta = 1$	$(\mathbf{1},\hat{\mathbf{c}})$	2 (38)	2 (34)	2 (38)	2 (36)	2 (40)
$(\mathbf{w}_1)$	$\mathrm{LP}(\hat{\mathbf{c}})$	2 (34)	2 (34)	2 (40)	2 (34)	2 (40)
	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	2 (30)	2 (30)	2 (34)	2 (32)	2 (34)
	$\mathrm{LP}(\overline{\mathbf{p}},\overline{\mathbf{q}})$	2 (30)	2 (32)	2 (40)	2 (36)	2 (40)
$\beta = 1.01$	$(\mathbf{1},\hat{\mathbf{c}})$	2 (32)	2 (32)	2 (34)	2 (32)	2 (32)
$(\mathbf{w}_2)$	$\mathrm{LP}(\hat{\mathbf{c}})$	2 (32)	2 (32)	2 (38)	2 (34)	2 (36)



(a) RMSE: $(\mathbf{p},\mathbf{q}) = (\overline{\mathbf{p}},\overline{\mathbf{q}})$ or $(\mathbf{1},\widehat{\mathbf{c}})$



(b) RMSE: $(\overline{\mathbf{p}},\overline{\mathbf{q}})$ or $\mathrm{LP}(\mathbf{1},\widehat{\mathbf{c}})$

Figure 5. RMSE comparison in Fig. 5(a) for $\mathbf{w}_2$ with the warm-start once and Fig. 5(b) for $\mathbf{w}_1$ with the more "aggressive" warm-start. Data were obtained by `SMM-Cadzow` with the 5 initial points given in Table 2.

lines for $(\overline{\mathbf{p}},\overline{\mathbf{q}})$ and $\mathrm{LP}(\widehat{\mathbf{c}})$ are below the other lines (except one value 227.56 from `Cadzow-`$\widehat{C}$). An interesting property is that the two lines are almost horizontal, suggesting that the solutions are less dependent on the starting point used. While we note that it may yield better RMSE in some cases, however, the big question for this "aggressive" use of the warm start strategy is that the resulting algorithm may suffer non-convergence. Hence, we choose not to report any further result for this choice and leave it to our future research.

5.3 Example: fortified wine sales

This data [18], denoted as $\mathbb{X}^*$, consists of the monthly sales of the fortified wine in Australia from January 1980 to December 1993 and it has been well studied via several methods including Basic SSA [13], ESPRIT [25, 15] and the Cadzow(α) algorithms in [28]. In particular, the original series $\mathbb{X}^*$ ($N = 168$) can be regarded as a realization of the noisy signal $\mathbb{X} = \mathbb{S} + \widetilde{\mathbb{E}}$, where the true signal $\mathbb{S} = (s_1, \ldots, s_N)$ and the noise $\widetilde{\mathbb{E}} = (\widetilde{\epsilon}_1, \ldots, \widetilde{\epsilon}_N)$ are respectively given for $k = 1, \ldots, N$ by

$$\begin{aligned} s_k = {} & 3997.74(0.9967)^k + 1174.75(0.9942)^k \sin(\frac{2\pi k}{12} - 2.249) \\ & + 425.75(1.0001)^k \sin(\frac{2\pi k}{4} + 2.333) \\ & + 211.55(1.004)^k \sin(\frac{2\pi k}{6} + 1.677) \\ & + 169.33(1.0007)^k \sin(\frac{2\pi k}{2.4} + 1.533) \\ & + 361.07(0.9884)^k \sin(\frac{2\pi k}{3} - 2.901) \end{aligned}$$

and $\widetilde{\epsilon}_k = 353.17(0.9967)^k \varepsilon_k$ with ε_k being Gaussian white noise with mean of zero and variance of one. Please refer to [28, Sect. 6] for why $\mathbb{X}^*$ can be approximated this way.

We follow the computational setting used in [28], where $L = 84$, $r = 11$ and 1000 random realizations of $\mathbb{X}$ have been used to test the reliability of the underlying algorithms to extract the true signal $\mathbb{S}$ from $\mathbb{X}$. It was demonstrated there that among a class of `Cadzow`(α) algorithms, `Cadzow`(0.2)

Table 7. RSME and computational cost comparison between the method SMM-Cadzow and Cadzow(0.2) in [28]. The results are average over 1000 realizations of $\mathbb{X}$

Method	$(\mathbf{p},\mathbf{q})$	RMSE-$\mathbb{S}$	Mdist-$\mathbb{X}$	Mdist-$\mathbb{X}^*$	`It`	`iter`
`SMM-Cadzow`	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	111.69	255.31	306.45	4.43	41.54
	$\mathrm{LP}(\overline{\mathbf{p}},\overline{\mathbf{q}})$	114.49	255.80	307.94	4.75	49.85
	$(\mathbf{1},\hat{\mathbf{c}})$	114.53	255.51	307.81	5.90	58.19
	$\mathrm{LP}(\hat{\mathbf{c}})$	115.53	255.55	308.31	5.80	59.15
`Cadzow`(α)	$\alpha = 0.2$	123.55	261.87	310.61	-	10.74

Table 8. RSME and computational cost comparison between the method SMM-Cadzow and Cadzow(0.2) in [28]. The results are average over 1000 realizations of $\mathbb{X}$. The stopping criterion for SMM-Cadzow is `tol` $= 10^{-2}$ and the maximum number of iterations for each subproblem is set at 5

Method	$(\mathbf{p},\mathbf{q})$	RMSE-$\mathbb{S}$	Mdist-$\mathbb{X}$	Mdist-$\mathbb{X}^*$	`It`	`iter`
`SMM-Cadzow`	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	113.70	255.35	307.40	2.07	10.56
	$\mathrm{LP}(\overline{\mathbf{p}},\overline{\mathbf{q}})$	113.28	251.91	307.57	2.35	11.34
	$(\mathbf{1},\hat{\mathbf{c}})$	118.17	256.49	308.96	2.12	11.20
	$\mathrm{LP}(\hat{\mathbf{c}})$	118.56	255.87	309.28	2.27	11.97
`Cadzow`(α)	$\alpha = 0.2$	123.36	260.80	310.55	-	5

performed the best in terms of the RMSE from $\mathbb{S}$. Therefore, we will compare our `SMM-Cadzow` method with `Cadzow`(0.2) (our own implementation). The stopping criterion used for `Cadzow`(0.2) is same as that in [28]. The average results of RMSE over 1000 realizations are reported in Table 7, where RMSE-$\mathbb{S}$ is the average RMSE between the obtained series and the true signal $\mathbb{S}$. Since both $\mathbb{X}$ and $\mathbb{X}^*$ contain noise, RMSE is not an appropriate term any more to represent the mean-distance (Mdist for short) between the obtained series $\widehat{X}$ and $\mathbb{X}$ or $\mathbb{X}^*$. We hence use Mdist-$\mathbb{X}$ and Mdist-$\mathbb{X}^*$ instead of RMSE-$\mathbb{X}$ and RMSE-$\mathbb{X}^*$ respectively. The column `It` includes the average number of the subproblems solved and `iter` is the average number of the projections used. In our experiment, the Gaussian white noise was generated by setting the `Matlab` random generator `rng = 'default'` so that the 1000 realizations can be reproduced.

It can be clearly observed that `SMM-Cadzow` yielded better RMSEs for all cases than `Cadzow`(0.2). In particular, the choice $(\overline{\mathbf{p}},\overline{\mathbf{q}})$ led to the lowest RMSE-$\mathbb{S}$ (111.69), but the cost of the number of iterations is about 5 times as expensive as `Cadzow`(0.2) (41.54 vs 10.74 under the `iter` column in Table 7). If we restrict the maximum number of projections used for each subproblem to be 5 and set the `tol` $= 10^{-2}$ (rather than 10^{-3} previously), we still obtained better results than `Cadzow`(0.2). This time, the cost is roughly the same order of magnitude. The results are reported in Table 8.

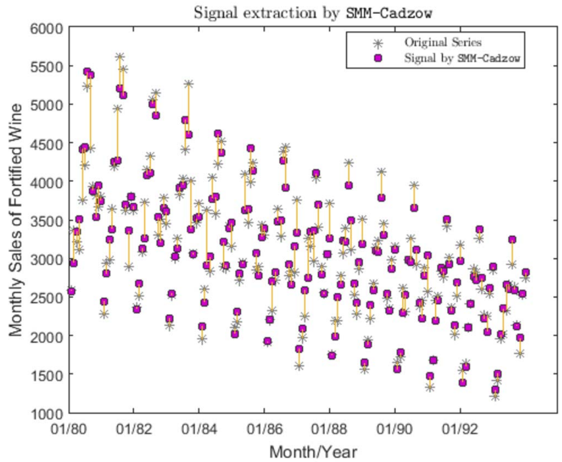
When the true sales data $\mathbb{X}^*$ is used, the extracted signals by `SMM-Cadzow` and `Cadzow`(0.2) are plotted in Fig. 6 and the respective RMSE and iteration information are included in Table 9. Once again, `SMM-Cadzow` obtained significantly better results than that by `Cadzow`(0.2). In particular, if we

Table 9. RSME and computational cost comparison between the method SMM-Cadzow and Cadzow(0.2) in [28] when the sales data $\mathbb{X}^*$ is used. The stopping criterion for SMM-Cadzow is set `tol` $= 10^{-3}$

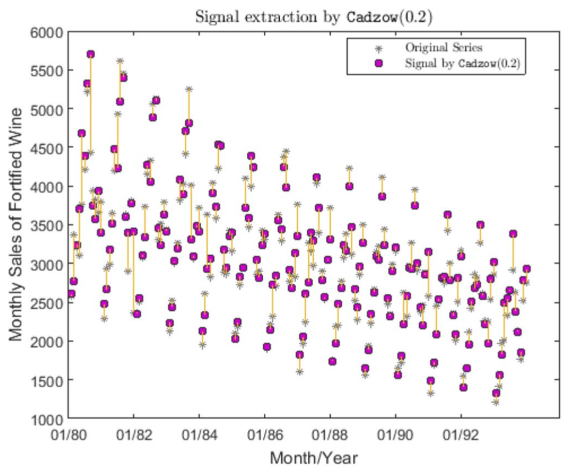
Method	$(\mathbf{p},\mathbf{q})$	RMSE-$\mathbb{S}$	Mdist-$\mathbb{X}^*$	`It`	`iter`
`SMM-Cadzow`	$(\overline{\mathbf{p}},\overline{\mathbf{q}})$	99.18	266.52	5	44
	$\mathrm{LP}(\overline{\mathbf{p}},\overline{\mathbf{q}})$	98.98	266.83	3	30
	$(\mathbf{1},\hat{\mathbf{c}})$	99.44	265.93	5	59
	$\mathrm{LP}(\hat{\mathbf{c}})$	97.82	266.22	5	55
`Cadzow`(α)	$\alpha = 0.2$	121.99	279.55	–	11

look closely at the plots in Fig. 6, the difference between the extracted signals by the two methods are sizeable and visible at certain points. For example, at the 5th and 8th month, the real sales data are $(3756, 4426)$. The extracted data for the two months by `SMM-Cadzow` are $(4418, 5373)$ and $(4689, 5684)$ by `Cadzow`(0.2). Hence, a significant improvement was made by `SMM-Cadzow` at those two points. We also like to point out that the performance of the two methods is quite similar at many other points (see Fig. 6).

Our final remark is about the question whether our final iterate is able to return a rank r solution. We take it for example of the final iterate $\widehat{X}$ by our `SMM-Cadzow` method with $(\mathbf{p},\mathbf{q}) = \mathrm{LP}(\overline{\mathbf{p}},\overline{\mathbf{q}})$. The sum of the first 11 largest singular values of $\widehat{X}$ explains 99.94% of the sum of all singular values. The ratio between the 12th largest singular value and the 11th largest singular value is 1.45×10^{-5}. Hence, it is safe to assume that $\widehat{X}$ has the finite rank r. The magnitude of all the singular values can be seen from its log scale plot in Fig. 7.



(a) `SMM-Cadzow`: $(\mathbf{p}, \mathbf{q}) = \mathrm{LP}(\overline{\mathbf{p}}, \overline{\mathbf{q}})$



(b) `Cadzow`(α): $\alpha = 0.2$

Figure 6. Fig. 6(a) plots the signal extracted by `SMM-Cadzow` with the choice LP$(\overline{\mathbf{p}}, \overline{\mathbf{q}})$) and RMSE $= 98.98$. Fig. 6(b) plots the signal extracted `Cadzow`(α) with $\alpha = 0.2$ in [28] and RMSE $= 121.99$. The extracted series and their corresponding original series in $\mathbb{X}^*$ are linked by a line.

6. CONCLUSION

In this paper, we studied the problem of extracting a time series of finite rank from its noisy observation under arbitrary weights. This problem is known computationally very challenging and two major methods are already available. One is the method based on the (Q, R)-norm approximation by Gillard and Zhigljavsky [11] and the other is the class of methods represented as `Cadzow`(α) by Zvonarev and Golyandina [28]. The purpose of this paper is to provide a third choice based on the technique of majorization.

The key difference is that we approximate the original problem by a sequence of well-defined subproblems, which are easier to solve. We demonstrated the advantages of the proposed method, named as `SMM-Cadzow` due to the fact that the subproblem is solved by the Cadzow method [4]. For example, the latest gradient information was used to construct the new approximation once a new iterate was obtained. The approximation can be improved through a smaller $(\mathbf{p}, \mathbf{q})$ weights, which can be refined by linear programming (cheap computational cost). The method was guaranteed to converge if the sandwich inequality is satisfied at each iterate. Moreover, its numerical performance was demonstrated against two popular test problems and a thorough comparison with the two existing methods were conducted to show its improvement.

This paper mainly focuses on introducing a new approximation framework and how its quality can be improved. Several important questions remain to be investigated. For example, instead of using the Cadzow method for the subproblems, a more powerful approach would be to use global methods as already explored in [16, 5]. It is also an interesting question whether we can define a new objective function that is ensured to be decreasing to lead to convergence. The research in nonconvex optimization [1] has strongly indicated that it is promising in tackling nonconvex structural problems as considered in this paper. We leave those topics for our future research.

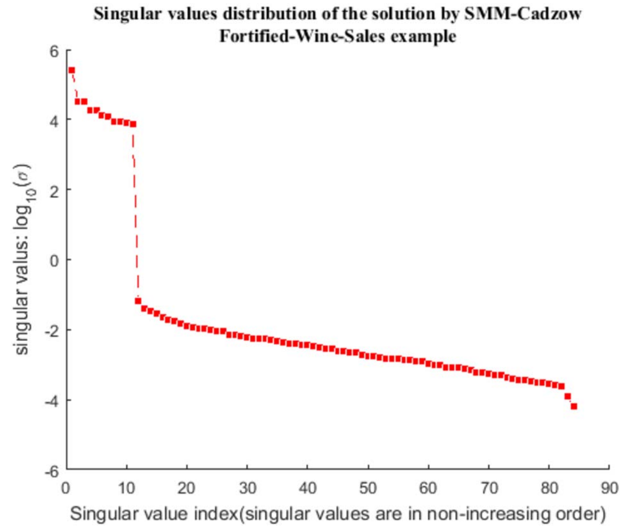


Figure 7. Singular values of $\widehat{X}$ in their $\log_{10}$ scale.

ACKNOWLEDGEMENTS

We would like to thank Prof Paul Harper of Cardiff University for drawing our attention to the topic and its application to health sciences [27] during his February 2017 CORMSIS seminar at the University of Southampton. The authors are also very grateful to the two referees for their constructive comments, which have led to significant improvement of the paper.

Received 15 May 2017

REFERENCES

[1] Attouch, H., Bolte, J., Redont, P. and Soubeyran, A. (2010). Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Lojasiewicz inequality. *Math. Oper. Res.* **35, 2** 438–457. MR2674728

[2] Box G.E., Jenkins G.M.and Reinsel, G.C. (2013). *Time Series Analysis: Forecasting and Control*. Wiley, New York.

[3] Butcher, H., Bolte, J. and Gillard, J. (2017). Simple nuclear norm based algorithms for imputing missing data and forecasting in time series. *Statistics and its Interface* **10** 19–25. MR3562108

[4] Cadzow, J.A. (1988). Signal enhancement-a composite property mapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **36, 1** 49–62.

[5] Chu M.T., Lin, M.M. and Wang, L. (2014). A study of singular spectrum analysis with global optimization techniques. *J. Glob. Optim.* **60** 551–574.

[6] de Leeuw, J. and Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software* **31, 3** 1–30.

[7] de Leeuw, J. (1993). *Fitting distances by least squares.* Technical report, University of California, Los Angeles.

[8] Gao, Y. and Sun, D.F. (2010). *A majorized penalty approach for calibrating rank constrained correlation matrix problems.* Technical report, Department of Mathematics, National University of Singapore.

[9] Gillard, J.W. (2010). Cadzow's basic algorithm, alternating projections and singular spectrum analysis. *Journal of Statistical Software* **3, 3** 335–343.

[10] Gillard, J.W. and Zhigljavsky, A.A. (2015). Stochastic algorithms for solving structured low-rank matrix approximation problems. *Commu. Nonlinear Sci. Numer. Simu.* **21** 70–88.

[11] Gillard, J.W. and Zhigljavsky, A.A. (2016). Weighted norms in subspace-based methods for time series analysis. *Numer. Linear Alg. Appl.* **23** 947–967.

[12] Golub, G.H. and Van Loan, C.F. (1983). *Matrix Computations*. Baltimore, MD: Johns Hopkins University Press.

[13] Golyandina, N., Nekrutkin, V. and Zhigljavsky, A.A. (2001). *Analysis of Time Series Structure: SSA and related techniques*. CRC Press.

[14] Golyandina, N. and Korobeynikov, A. (2014). Basic singular spectrum analysis and forecasting with R. *Computational Statistics and Data Analysis* **71** 934–954. MR3132018

[15] Golyandina, N. and Zhigljavsky, A.A. (2013). *Singular Spectrum Analysis for Time Series*. Springer: Berlin, Heidelberg.

[16] Golyandina, N. and Zhigljavsky, A.A. (2013). Optimization challenges in the structured low rank approximation problem. *J. Global Optim.* **57, 3** 733–751.

[17] Hassani, H. (2007). Singular spectrum analysis: methodology and comparison. *J. Data Science.* **5** 239–257.

[18] Hyndman, R.J. (2015). Time Series Data Library.

[19] Jiang, K., Sun, D. and Toh, K.-C. (2012). An inexact accelerated proximal gradient method for large scale linearly constrained convex SDP. *SIAM J. Optim.* **22, 3** 1042–1064.

[20] Lee, D.D. and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401, 6755** 788–791.

[21] Lee, D.D. and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems* 556–562.

[22] Li, M., Sun, D. and Toh, K.-C. (2016). A majorized ADMM with indefinite proximal terms for linearly constrained convex composite optimization. *SIAM J. Optim.* **26, 2** 922–950.

[23] Li, Q. and Qi, H.-D. (2011). A sequential semismooth Newton method for the nearest low-rank correlation matrix problem. *SIAM J. Optim.* **16, 4** 1641–1666. MR2869511

[24] Qi, H-D. and Yuan, X. (2014). Computing the nearest Euclidean distance matrix with low embedding dimensions. *Math. Prog.* **147, 1–2** 351–389.

[25] Roy, R. and Kailath, T. (1989). ESPRIT: estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoust.* **37** 984–995.

[26] Stewart, G.W. (1973). Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Review* **15, 4** 727–764.

[27] Vile, J. L., Vile, J. L., Gillard, J. W., Harper, P. R. and Knight, V. A. (2016). Time-dependent stochastic methods for managing and scheduling Emergency Medical Services. *Operations Research for Health Care* **8** 42–52.

[28] Zvonarev, N. and Golyandina, N. (2017). Iterative algorithms for weighted and unweighted finite-rank time-series approximations. *Statistics and Its Interface* **10** 5–18.

Hou-Duo Qi
School of Mathematics
The University of Southampton
Highfield, Southampton SO17 1BJ
UK
E-mail address: hdqi@soton.ac.uk

Jian Shen
School of Mathematics
The University of Southampton
Highfield, Southampton SO17 1BJ
UK
E-mail address: J.Shen@soton.ac.uk

Naihua Xiu
Department of Applied Mathematics
Beijing Jiaotong University
Beijing
China
E-mail address: nhxiu@bjtu.edu.cn