Mathematical Programming Computation (2022) 14:417–450
https://doi.org/10.1007/s12532-022-00217-1

FULL LENGTH PAPER


# A penalized method of alternating projections for weighted low-rank hankel matrix optimization

**Jian Shen[1] · Jein-Shan Chen[2] · Hou-Duo Qi[1] · Naihua Xiu[3]**

Received: 1 August 2020 / Accepted: 18 December 2021 / Published online: 3 February 2022


**Abstract**
Weighted low-rank Hankel matrix optimization has long been used to reconstruct contaminated signal or forecast missing values for time series of a wide class. The Method of Alternating Projections (MAP) (i.e., alternatively projecting to a low-rank matrix manifold and the Hankel matrix subspace) is a leading method. Despite its wide use, MAP has long been criticized of lacking convergence and of ignoring the weights used to reflect importance of the observed data. The most of known results are in a local sense. In particular, the latest research shows that MAP may converge at a linear rate provided that the initial point is close enough to a true solution and a transversality condition is satisfied. In this paper, we propose a globalized variant of MAP through a penalty approach. The proposed method inherits the favourable local properties of MAP and has the same computational complexity. Moreover, it is capable of handling a general weight matrix, is globally convergent, and enjoys local linear convergence rate provided that the cutting off singular values are significantly smaller than the kept ones. Furthermore, the new method also applies to complex data. Extensive numerical experiments demonstrate the efficiency of the proposed method against several popular variants of MAP.

✉ Jian Shen
j.shen@soton.ac.uk

Jein-Shan Chen
jschen@math.ntnu.edu.tw

Hou-Duo Qi
hdqi@soton.ac.uk

Naihua Xiu
nhxiu@bjtu.edu.cn

1 School of Mathematical Sciences, University of Southampton, Highfield, Southampton SO17 1BJ, UK

2 Department of Mathematics, National Taiwan Normal University, Taipei 11677, Taiwan

3 Department of Applied Mathematics, Beijing Jiaotong University, Beijing, China

**Keywords** Hankel matrix · Alternating projections · Global convergence · Linear convergence · Time series

**Mathematics Subject Classification** 47B35 · 62M10 · 65F55 · 90C26 · 90C30

## 1 Introduction

In this paper, we are mainly interested in the numerical methods for the weighted low-rank Hankel matrix optimization:

$$\min \ f(X) := \frac{1}{2}\|W \circ (X - A)\|, \ \text{ s.t. } \ X \in \mathbb{M}_r \cap \mathbb{H}^{k\times\ell}, \tag{1}$$

where $\mathbb{H}^{k\times\ell}$ is the space of all $k \times \ell$ Hankel matrices in the real/complex field with the standard trace inner product, $\|\cdot\|$ is the Frobenius norm, $\mathbb{M}_r$ is the set of matrices whose ranks are not greater than a given rank $r$, $A$ is given, and $W$ is a given weight matrix $W_{ij} \geq 0$. Here $(A \circ B)$ represents elementwise multiplication (e.g., Hadamard product) between $A$ and $B$. We note that the size of all matrices involved are of $k \times \ell$. The difficulties in solving (1) are with the low-rank constraint and how to effectively handle a general weight matrix $W$, the latter of which is often overlooked in existing literature. Our main purpose is to develop a novel, globally convergent algorithm for (1) and its efficiency will be benchmarked against several state-of-the-art algorithms.

In what follows, we first explain an important application of (1) to time series data, which will be tested in our numerical experiment part. We then review the latest advances on algorithms relating to the alternating projection method. We finish this section by explaining our approach and main contributions.

### 1.1 Applications in time series

Problem (1) arises from a large number of applications including signal processing, system identification and finding the greatest common divisor between polynomials [23]. To motivate our investigation on (1), let us consider a complex-valued time series $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ of finite rank [17, Chp. 5]:

$$a_t = \sum_{s=1}^{m} P_s(t)\lambda_s^t, \quad t = 1, 2, \ldots, n \tag{2}$$

where $P_s(t)$ are a complex polynomial of degree $(\nu_s - 1)$ ($\nu_s$ are positive integers) and $\lambda_s \in \mathbb{C} \setminus \{0\}$ are distinct. Define $r := \nu_1 + \ldots + \nu_m$ (":=" means "define"). Then it is known [29, Prop. 2.1] that the rank of the Hankel matrix $A$ generated by $\mathbf{a}$:

$$A = \mathcal{H}(\mathbf{a}) := \begin{bmatrix} a_1 & a_2 & \cdots & a_\ell \\ a_2 & a_3 & \cdots & a_{\ell+1} \\ \vdots & \vdots & \vdots & \vdots \\ a_k & a_{k+1} & \cdots & a_n \end{bmatrix}$$

must be $r$, where the choice of $(k, \ell)$ satisfies $n = k + \ell - 1$ and $r \le k \le n - r + 1$.

Suppose now that the time series $\mathbf{a}$ is contaminated and/or has missing values. To reconstruct $\mathbf{a}$, a natural approach is to computing its nearest time series $\mathbf{x}$ by the least squares:

$$\min \sum_{i=1}^{n} w_i |a_i - x_i|^2, \quad \text{s.t.} \quad \operatorname{rank}(X) \le r, \quad X = \mathcal{H}(\mathbf{x}), \tag{3}$$

where $\mathbf{w} = (w_1, \ldots, w_n) \ge 0$ are the corresponding weight vector emphasizing the importance of each elements of $\mathbf{a}$. The equivalent reformulation of (3) as (1) is obtained by setting

$$W := \mathcal{H}(\sqrt{\mathbf{v}} \circ \sqrt{\mathbf{w}}) \quad \text{and} \quad v_i = \begin{cases} 1/i & \text{for } i = 1, \ldots, k-1 \\ 1/k & \text{for } i = k, \ldots, n-k+1 \\ 1/(n-i+1) & \text{for } i = n-k+2, \ldots, n, \end{cases}$$

where $\mathbf{v}$ is known as the averaging vector of Hankel matrix of size $k \times \ell$ ($k \le \ell$) and $\sqrt{\mathbf{w}}$ is the elementwise square root of $\mathbf{w}$. We note that the widely studied (1) with $W_{ij} \equiv 1$ corresponds to $w_i = 1/v_i$, which is known as the trapezoid weighting. Another popular choice for financial time series is the exponential weights $w_i = \exp(\alpha i)$ for some $\alpha > 0$. We refer to [15, Sect. 2.4] for more comments on the choice of weights.

A special type of the time series of (2) arises from the spectral compressed sensing, which has attracted considerable attention lately [4]. In its one dimensional case, $a_t$ is often a superposition of a few complex sinusoids:

$$a_t = \sum_{s=1}^{r} d_s \exp\{(2\pi j \omega_s - \tau_s)t\}, \tag{4}$$

where $j = \sqrt{-1}$, $r$ is the model order, $\omega_s$ is the frequency of each sinusoid, and $d_s \ne 0$ is the weight of each sinusoid, and $\tau_s \ge 0$ is a damping factor. We note that (4) is a special case of (2) with $P_s(t) = d_s$ (hence $\nu_s = 1$) and $\lambda_s = \exp(2\pi j \omega_s - \tau_s)$. If $a_t$ is sampled at all integer values from 1 to $n$, we get a sample vector $\mathbf{a} \in \mathbb{C}^n$. Consequently, the rank of $\mathcal{H}(\mathbf{a})$ must be $r$. However, in practice, only a subset $\Omega$ of the sampling points $\{1, \ldots, n\}$ may be observed (possibly contaminated), leading to the question of how to best reconstruct $a(t)$ based on its partial observation $a_i$ on $\Omega$. This has led to the Hankel matrix completion/approximation problem of (1), see [4, Sect. II.A] and [2, Sect. 2.1]. A popular choice of $W$ in the spectral compressed sensing is $W_{ij} = 1$ for all $(i, j)$, resulting in the distance between $X$ and $A$ in (1) being measured by the standard Frobenius norm. In this paper, we assume

**Assumption 1** $W$ is Hankel and non-negative (i.e., $W_{ij} \ge 0$ for all $(i, j)$).

### 1.2 On alternating projection methods

Low-rank matrix optimization is an active research area. Our short review is only able to focus on a small group of those papers that motivated our research. We note that there are four basic features about the problem (1): (i) $X$ has to be low rank; (ii) $X$ has Hankel structure; (iii) the objective is weighted; and (iv) $X$ may be complex valued. The first feature is the most difficult one to handle because it causes the nonconvexity of the problem. Many algorithms have been developed proposing different ways to handle this low rank constraint. One of the most popular choices is to use the truncated singular value decomposition to project $X$ to be its nearest rank $r$ matrix and we denote the projection by $\Pi_{\mathbb{M}_r}(X)$. This has given rise to the basic Method of Alternating Projections (MAP) (also known as the Cadzow method [1]): Given $X^0 \in \mathbb{H}$, update

$$X^{\nu+1} = \Pi_{\mathbb{H}}\Big(\Pi_{\mathbb{M}_r}(X^{\nu})\Big), \quad \nu = 0, 1, 2, \ldots \tag{5}$$

where $\Pi_{\mathbb{H}}(\cdot)$ is the orthogonal projection operator to the Hankel subspace $\mathbb{H}^{k\times\ell}$. Despite its popularity in engineering sciences, Cadzow's method can not guarantee the convergence to an optimal solution. Even convergence occurs, not much is known about where it converges to. It has also been criticized for completely ignoring the objective function, see [5,6,8,14]. In particular, the weight matrix $W$ does not enter Cadzow's method at all because the truncated SVD does not admit a closed-form solution under a weighted norm. Gillard and Zhigljavsky [16] proposed to replace $\Pi_{\mathbb{M}_r}$ by its diagonally weighted variants and studied how to best approximate the weight matrix $W$ by a diagonal weight matrix. Qi et. al. [25] proposed to use a sequential diagonal weight matrices aiming to get a better approximation to the original weight matrix. Despite the improved numerical convergence, those methods in [16,25] still inherit the essential problem of convergence of Cadzow's method. Recently, Lai and Varghese [20] considered a similar method for a matrix completion problem and established the linear convergence of their method under a kind of "transversality" condition provided that the initial point is close enough to a true rank-$r$ completion. We refer to [7] for a more general transversality condition that ensures a local linear convergence rate of MAP onto nonconvex sets.

Alternating projections of $\Pi_{\mathbb{M}_r}(\cdot)$ and $\Pi_{\mathbb{H}}(\cdot)$ also play an important role in the class of iterative hard thresholding (IHT) algorithms for spectral compressed sensing. For example, Cai et. al. [3] established the convergence of IHT in the statistical sense (i.e., with high probability) under a coherence assumption on the initial observation matrix $A$. Although local convergence results (be in the sense of monotonically decreasing [20] or in the statistical sense [3]) may be established for MAP under some conditions, we are not aware of any existing global convergence results mainly due to the nonconvexity of the rank constraint. For the general weighted (1), it appears to be a difficult task to develop a variant of MAP that enjoys both global and local linear convergence properties. We will achieve this through a penalty approach.

Penalty approaches have long been used to develop globally convergent algorithms for problems with rank constraints, see [10,12,13,21,22,27,32,33]. For example, Gao [12] proposed the penalty function $p(X)$ based on the following observation:

$$\operatorname{rank}(X) \le r \quad \Longleftrightarrow \quad p(X) := \|X\|_* - \sum_{i=1}^{r} \sigma_i(X) = 0,$$

where $\|X\|_*$ is the nuclear norm of $X$ and $\sigma_1(X) \ge \cdots \sigma_n(X)$ are the singular values of $X$ in nonincreasing order. However, the resulting method, as well as those in [21,22,27], has nothing to do with MAP any more and its implementation is not trivial.

### 1.3 Our approach and main contributions

In this paper, we propose a new penalty function and develop a penalized method whose main step is the alternating projections. We call it the penalized MAP (pMAP). The new penalty function is the Euclidean distance function $d_{\mathbb{M}_r}(X)$ from $X$ to $\mathbb{M}_r$:

$$d_{\mathbb{M}_r}(X) := \min \{\|X - Z\| \mid Z \in \mathbb{M}_r\} \text{ and define } g_r(X) := \frac{1}{2} d^2_{\mathbb{M}_r}(X). \tag{6}$$

Obviously, the original problem (1) is equivalent to

$$\min \ f(X), \quad \text{s.t. } d_{\mathbb{M}_r}(X) = 0, \quad X \in \mathbb{H}.$$

We propose to solve the quadratic penalty problem with $\rho > 0$ being a penalty parameter:

$$\min \ F_\rho(X) := f(X) + \rho g_r(X), \quad \text{s.t. } X \in \mathbb{H}. \tag{7}$$

By following the standard argument [24, Thm. 17.1] for the classical quadratic penalty method, one can establish that the global solution of (7) converges to that of (1) as $\rho$ approaches infinity provided the convergence happens. However, in practice, it is probably as difficult to find a global solution of (7) as for the original problems. It is hence important to establish the correspondence between the first-order stationary points of (7) and that of (1). This is done in Theorem 1 under a generalized linear independence condition.

The remaining task is to efficiently compute a stationary point of (7) for a given $\rho > 0$. The key observation is that $g_r(X)$ can be represented as the difference of two convex functions, which can be easily majorized (later on its meaning) to get a majorization function $g_r^{(m)}(X, X^\nu)$ of $g_r(X)$ at the current iterate $X^\nu$. We then solve the majorized subproblem:

$$X^{\nu+1} = \arg\min \ f(X) + \rho g_r^{(m)}(X, X^\nu), \quad \text{s.t.} \quad X \in \mathbb{H}. \tag{8}$$

We will show that the update takes the following form:

$$X^{\nu+1} = \frac{W^{(2)}}{\rho + W^{(2)}} \circ A + \frac{\rho}{\rho + W^{(2)}} \circ \Pi_{\mathbb{H}}(\Pi_{\mathbb{M}_r}(X^\nu)), \tag{9}$$

where $W^{(2)} := W \circ W$ and the division $W^{(2)}/(\rho + W^{(2)})$ is taken componentwise. Compared with (5), this update is just a convex combination of the observation matrix $A$ and the MAP iterate in (5). In the special case that $W \equiv 0$ (which completely ignores the objective in (1)) or $\rho = \infty$, (9) reduces to MAP. We will analyze the convergence behaviour of pMAP (9). In particular, we will establish the following among others.

(i) The objective function sequence $\{F(X^\nu, \rho)\}$ will converge and $\|X^{\nu+1} - X^\nu\|$ converges to 0. Moreover, any limiting point of $\{X^\nu\}$ is an approximate KKT point of the original problem (1) provided that the penalty parameter is above certain threshold (see Theorem 2).
(ii) If $\widehat{X}$ is an accumulation point of the iterate sequence $\{X^\nu\}$, then the whole sequence converges to $\widehat{X}$ at a linear rate provided that $\sigma_r(\widehat{X}) \gg \sigma_{r+1}(\widehat{X})$ (see Theorem 3).

Our results in (i) and (ii) provide satisfactory justification of pMAP. It is not only globally convergent, but also enjoys a linear convergence rate under reasonable conditions. Furthermore, we can assess the quality of the solution as an approximate KKT of (1) if we are willing to increase the penalty parameter. Of course, balancing the fidelity term $f(X)$ and the penalty term is an important issue that is beyond the current paper. The result in (ii) is practically important too. Existing empirical results show that MAP often terminates at a point whose cut-off singular values $(\sigma_i, i \geq r + 1)$ are significantly smaller than the kept singular values $(\sigma_i, i \leq r)$. Such points are often said to have a numerical rank $r$, but the theoretical rank is higher than $r$. This is exactly the situation that was addressed in (ii). Those results are stated and proved for real-valued matrices. We will extend them to the complex case, thanks to a technical result (Proposion 2) that the subdifferential of $g_r(X)$ in complex domain can also be computed in a similar fashion as in the real domain. To our best knowledge, this is the first variant of MAP that can handle general weights and enjoys both global convergence and locally linear convergence rate under a reasonable condition (i.e., $\sigma_r \gg \sigma_{r+1}$).

The paper is organized as follows. In the next section, we will first set up our standard notation and establish the convergence result for the quadratic penalty approach (7) when $\rho \to \infty$. Sect. 3 includes our method of pMAP and its convergence results when $\rho$ is fixed. In Sect. 4, we will address the issue of extension to the complex-valued matrices, which arise from (2) and (4). The key concept used in this section is the Wirtinger calculus, which allows us to extend our analysis from the real case to the complex case. We report extensive numerical experiments in Sect. 5 and conclude the paper in Sect. 6.

## 2 Quadratic penalty approach

The main purpose of this section is to establish the convergence of the stationary points of the penalized problems (7) to that of the original problem (1) as the penalty parameter $\rho$ goes to $\infty$. For the simplicity of our analysis, we focus on the real case. We will extend our results to the complex case in Sect. 4. We first introduce the notation used in this paper.

### 2.1 Notation

For a nonnegative matrix such as the weight matrix $W$, $\sqrt{W}$ is its componentwise square root matrix $(\sqrt{W_{ij}})$. For a given matrix $X \in \mathbb{C}^{k\times\ell}$, we often use its singular value decomposition (assume $k \le \ell$)

$$X = U\,\mathrm{diag}(\sigma_1(X), \ldots, \sigma_k(X))V^T, \tag{10}$$

where $\sigma_1(X) \ge \cdots \ge \sigma_n(X)$ are the singular values of $X$ and $U \in \mathbb{C}^{k\times k}$, $V \in \mathbb{C}^{\ell\times\ell}$ are the left and right singular vectors of $X$. For a given closed subset $\mathcal{C} \subset \mathbb{C}^{k\times\ell}$, we define the set of all projections from $X$ to $\mathcal{C}$ by

$$\mathcal{P}_{\mathcal{C}}(X) := \arg\min\{\|X - Z\| : \ Z \in \mathcal{C}\}.$$

If $\mathcal{C}$ is also convex, then $\mathcal{P}_{\mathcal{C}}(X)$ is unique. When $\mathcal{C} = \mathbb{M}_r$, $\mathcal{P}_{\mathbb{M}_r}(X)$ may have multiple elements. We define a particular element in $\mathcal{P}_{\mathbb{M}_r}(X)$ that is based on the SVD (10):

$$\Pi_{\mathbb{M}_r}(X) = U_r\,\mathrm{diag}(\sigma_1(X), \ldots, \sigma_r(X))V_r^T,$$

where $U_r$ and $V_r$ consist of the first $r$ columns of $U$ and $V$ respectively.

Related to the function $g_r(X)$ defined in (6), the function

$$h_r(X) := \frac{1}{2}\|X\|_F^2 - g_r(X), \tag{11}$$

has the following properties by the classical result of Eckart and Young [9]:

$$\begin{aligned}\mathrm{dist}^2(X, \mathbb{M}_r) &= \|X - \Pi_{\mathbb{M}_r}\|^2 = \sigma_{r+1}^2(X) + \cdots + \sigma_n^2(X),\\ h_r(X) &= \frac{1}{2}\Big(\sigma_1^2(X) + \cdots + \sigma_r^2(X)\Big) = \frac{1}{2}\|\Pi_{\mathbb{M}_r}(X)\|^2.\end{aligned}$$

It follows from [12, Prop. 2.16] that $h_r(X)$ is convex and the subdifferentials of $h_r(X)$ and $g_r(X)$ in the sense of [26, Def. 8.3] are respectively given by

$$\partial h_r(X) = \mathrm{conv}(\mathcal{P}_{\mathbb{M}_r}(X)) \quad \text{and} \quad \partial g_r(X) = X - \partial h_r(X), \tag{12}$$

where $\mathrm{conv}(\Omega)$ denotes the convex hull of the set $\Omega$. Finally, we let $\mathcal{B}_\epsilon(X)$ denote the $\epsilon$-neighbourhood centred at $X$.

### 2.2 Convergence of quadratic penalty approach

The classical quadratic penalty methods try to solve a sequence of penalty problems:

$$X^\nu = \arg\min \ F_{\rho_\nu}(X), \quad \text{s.t.} \ \ X \in \mathbb{H}, \tag{13}$$

where the sequence $\rho_\nu > 0$ is increasing and goes to $\infty$. By following the standard argument (e.g., [24, Thm. 17.1]), one can establish that every limit of $\{X^\nu\}$ is also a global solution of (1). However, in practice, it is probably as difficult to find a global solution for (13) as for the original problem (1). Therefore, only an approximate solution of (13) is possible. To quantify the approximation, we recall the optimality conditions relating to both the original and penalized problems.

Following the optimality theorem [26, Thm. 8.15], we define the first-order optimality condition of problem (1) and (7):

**Definition 1** (*First-order optimality condition*) $\widehat{X} \in \mathbb{H}$ satisfies the first-order optimality condition of (1) if

$$0 \in \nabla f(\widehat{X}) + \widehat{\lambda}\partial d_{\mathbb{M}_r}(\widehat{X}) + \mathbb{H}^\perp, \tag{14}$$

where $\widehat{\lambda}$ is the Lagrangian multiplier. Similarly, we say $X^\nu \in \mathbb{H}$ satisfies the first-order optimality condition of the penalty problem (7) if

$$0 \in \nabla f(X^\nu) + \rho_\nu\ \partial g_r(X^\nu) + \mathbb{H}^\perp. \tag{15}$$

We generate $X^\nu \in \mathbb{H}$ such that the condition (15) is approximately satisfied:

$$\|\mathcal{P}_{\mathbb{H}}(\nabla f(X^\nu) + \rho_\nu(X^\nu - \Pi_{\mathbb{M}_r}(X^\nu)))\| \le \epsilon_\nu, \tag{16}$$

where $\epsilon_\nu \downarrow 0$. We can establish the following convergence result.

**Theorem 1** *We assume the sequence* $\{\rho_\nu\}$ *goes to* $\infty$ *and* $\{\epsilon_\nu\}$ *decreases to* 0. *Suppose each approximate solution* $X^\nu$ *is generated to satisfy (16).*

*Let* $\widehat{X}$ *be an accumulation point of* $\{X^\nu\}$ *and we assume*

$$\partial d_{\mathbb{M}_r}(\widehat{X}) \cap \mathbb{H}^\perp = \{0\}. \tag{17}$$

*Then* $\widehat{X}$ *satisfies the first-order optimality condition (14).*

***Proof*** Suppose $\widehat{X}$ is the limiting point of the subsequence $\{X^\nu\}_{\mathcal{K}}$. We consider the following two cases.

**Case 1** There exists an infinite subsequence $\mathcal{K}_1$ of $\mathcal{K}$ such that $\text{rank}(X^\nu) \le r$ for $\nu \in \mathcal{K}_1$. This would imply $\partial g_r(X^\nu) = \{0\}$, which with (16) implies $\|\mathcal{P}_{\mathbb{H}}(\nabla f(X^\nu))\| \to 0$. Hence (14) holds at $\widehat{X}$ with the choice $\widehat{\lambda} = 0$.

**Case 2** There exists an index $\nu_0$ such that $X^\nu \notin \mathbb{M}_r$ for all $\nu_0 \le \nu \in \mathcal{K}$. In this case, we assume that there exists an infinite subsequence $\mathcal{K}_2$ of $\mathcal{K}$ such that $\left\{(X^\nu - \Pi_{\mathbb{M}_r}(X^\nu))/d_{\mathbb{M}_r}(X^\nu)\right\}$ has the limit $\mathbf{v}$. We note that $(X^\nu - \Pi_{\mathbb{M}_r}(X^\nu))/d_{\mathbb{M}_r}(X^\nu) \in \partial d_{\mathbb{M}_r}(X^\nu)$ for $\nu \ge \nu_0$ by [26, (8.53)]. Therefore, its limit $\mathbf{v} \in \partial d_{\mathbb{M}_r}(\widehat{X})$ by the upper semicontinuity. By the assumption (17), we have $\mathbf{v} \notin \mathbb{H}^\perp$ because $\mathbf{v}$ has the unit length. Since $\mathbb{H}$ is a subspace, $\mathcal{P}_{\mathbb{H}}(\cdot)$ is a linear operator. It follows from (16) that

$$\rho_\nu\|\mathcal{P}_{\mathbb{H}}(X^\nu - \Pi_{\mathbb{M}_r}(X^\nu))\| - \|\mathcal{P}_{\mathbb{H}}(\nabla f(X^\nu))\| \le \|\mathcal{P}_{\mathbb{H}}(\nabla f(X^\nu) + \rho_\nu(X^\nu - \Pi_{\mathbb{M}_r}(X^\nu))\| \le \epsilon_\nu.$$

Hence

$$\|\mathcal{P}_{\mathbb{H}}(X^\nu - \Pi_{\mathbb{M}_r}(X^\nu))\| \le \frac{1}{\rho_\nu}\Big(\epsilon_\nu + \|\mathcal{P}_{\mathbb{H}}(\nabla f(X^\nu))\|\Big),$$

which, for $\nu \ge \nu_0$, is equivalent to

$$d_{\mathbb{M}_r}(X^\nu)\|\mathcal{P}_{\mathbb{H}}(X^\nu - \Pi_{\mathbb{M}_r}(X^\nu))/d_{\mathbb{M}_r}(X^\nu)\| \le \frac{1}{\rho_\nu}\Big(\epsilon_\nu + \|\mathcal{P}_{\mathbb{H}}(\nabla f(X^\nu))\|\Big).$$

Taking limits on $\{X^\nu\}_{\nu\in\mathcal{K}_2}$ and using the fact $\rho_\nu \to \infty$ leads to $d_{\mathbb{M}_r}(\widehat{X})\|P_{\mathbb{H}}(\mathbf{v})\| = 0$. Since $\mathbf{v} \notin \mathbb{H}^\perp$, we have $\|\mathcal{P}_{\mathbb{H}}(\mathbf{v})\| > 0$, which implies $d_{\mathbb{M}_r}(\widehat{X}) = 0$. That is, $\widehat{X}$ is a feasible point of (1). Now let $\lambda_\nu := \rho_\nu d_{\mathbb{M}_r}(X^\nu)$, we then have

$$\lambda_\nu \frac{X^\nu - \Pi_{\mathbb{M}_r}(X^\nu)}{d_{\mathbb{M}_r}(X^\nu)} = -\nabla f(X^\nu) + \xi^\nu, \quad \xi^\nu := \nabla f(X^\nu) + \rho_\nu(X^\nu - \Pi_{\mathbb{M}_r}(X^\nu)).$$

Projecting on both sides to $\mathbb{H}$ yields

$$\lambda_\nu \mathcal{P}_{\mathbb{H}}\left(\frac{X^\nu - \Pi_{\mathbb{M}_r}(X^\nu)}{d_{\mathbb{M}_r}(X^\nu)}\right) = \mathcal{P}_{\mathbb{H}}(-\nabla f(X^\nu)) + \mathcal{P}_{\mathbb{H}}(\xi^\nu). \tag{18}$$

Computing the inner product on both sides with $\mathcal{P}_{\mathbb{H}}((X^\nu - \Pi_{\mathbb{M}_r}(X^\nu))/d_{\mathbb{M}_r}(X^\nu))$, taking limits on the sequence indexed by $\mathcal{K}_2$, and using the fact $\mathcal{P}_{\mathbb{H}}(\xi^\nu) \to 0$ due to (16), we obtain

$$\lim_{\nu\in\mathcal{K}_2} \lambda_\nu\|\mathbf{v}\|^2 = \langle \mathbf{v},\ \mathcal{P}_{\mathbb{H}}(\nabla f(\widehat{X}))\rangle.$$

We then have

$$\widehat{\lambda} = \lim_{\nu\in\mathcal{K}_2} \lambda_\nu = \frac{1}{\|\mathbf{v}\|^2}\langle \mathbf{v},\ \mathcal{P}_{\mathbb{H}}(\nabla f(\widehat{X}))\rangle.$$

Taking limits on both sides of (18) yields

$$\mathcal{P}_{\mathbb{H}}(\nabla f(\widehat{X}) + \widehat{\lambda}\mathbf{v}) = 0,$$

which is sufficient for

$$0 \in \nabla f(\widehat{X}) + \widehat{\lambda}\partial d_{\mathbb{M}_r}(\widehat{X}) + \mathbb{H}^\perp.$$

This completes our result. □

***Remark 1*** Condition (17) can be interpreted as that any $0 \neq \mathbf{v} \in \partial d_{\mathbb{M}_r}(\widehat{X})$ is linearly independent of any set of basis of $\mathbb{H}^{\perp}$. Therefore, (17) can be seen as a generalization of the linear independence assumption required in the classical quadratic penalty method for a similar convergence result with all the functions involved being assumed continuously differentiable, see [24, Thm. 17.2]. In fact, what we really needed in our proof is that there exists a subsequence $\{(X^{\nu} - \Pi_{\mathbb{M}_r}(X^{\nu}))/d_{\mathbb{M}_r}(X^{\nu})\}$ in Case (ii) such that its limit $\mathbf{v}$ does not belong to $\mathbb{H}^{\perp}$. That could be much weaker than the sufficient condition (17).

Theorem 1 establishes the global convergence of quadratic penalty method when the penalty parameter approaches infinity, which drives $g_r(X^{\nu})$ smaller and smaller. In practice, however, we often fix $\rho$ and solve for $X^{\nu}$. We are interested in how far $X^{\nu}$ is from being a first-order optimal point of the original problem. For this purpose, we introduce the approximate KKT point, which keeps the first-order optimality condition (15) with an additional requirement that $g_r(X)$ is small enough.

**Definition 2** (*$\epsilon$-approximate KKT point*) Consider the penalty problem (7) and $\epsilon > 0$ is given. We say a point $\widehat{X} \in \mathbb{H}$ is an $\epsilon$-approximate KKT point of (1) if

$$0 \in \nabla f(\widehat{X}) + \rho\, \partial g_r(\widehat{X}) + \mathbb{H}^{\perp} \quad \text{and} \quad g_r(\widehat{X}) \leq \epsilon.$$

## 3 The method of pMAP

This section develops a new algorithm that solves the penalty problem (7), in particular for finding an approximate KKT point of (1). The first part is devoted to the construction of a majorization function for the distance function $\mathrm{dist}(X, \mathbb{M}_r)$. We then describe pMAP based on the majorization introduced and establish its global and local convergence.

### 3.1 Majorization and DC interpretation

We first recall the essential properties that a majorization function should have. Let $\theta(\cdot)$ be a real-valued function defined in a finite-dimensional space $\mathcal{X}$. For a given $\mathbf{y} \in \mathcal{X}$, we say a function $\theta^{(m)}(\cdot, \mathbf{y}) : \mathcal{X} \mapsto \mathbb{R}$ is a majorization of $\theta(\cdot)$ at $\mathbf{y}$ if

$$\theta^{(m)}(\mathbf{x}, \mathbf{y}) \geq \theta(\mathbf{x}), \quad \forall\, \mathbf{x} \in \mathcal{X} \quad \text{and} \quad \theta^{(m)}(\mathbf{y}, \mathbf{y}) = \theta(\mathbf{y}). \tag{19}$$

The motivation for employing the majorization is that the squared distance function $g_r(X)$ is hard to minimize when coupled with $f(X)$ under the Hankel matrix structure. It is noted that

$$g_r(X) = \frac{1}{2}\|X - \Pi_{\mathbb{M}_r}(X)\|^2 \leq \frac{1}{2}\|X - \Pi_{\mathbb{M}_r}(Z)\|^2 := g_r^{(m)}(X, Z), \quad \forall\, X, Z \in \mathbb{C}^{k\times \ell}$$

where the inequality used the fact that $\Pi_{\mathbb{M}_r}(X)$ is a nearest point in $\mathbb{M}_r$ to $X$. It is easy to verify that $g_r^{(m)}(X, Z)$ is a majorization function of $g_r(X)$ at $Z$.

The following way in deriving the majorization is crucial to our convergence analysis. We recall

$$\begin{aligned} h_r(X) &= \frac{1}{2}\|X\|^2 - g_r(X) = \frac{1}{2}\|X\|^2 - \frac{1}{2}\min\left\{\|X-Z\|^2 :\ Z \in \mathbb{M}_r\right\} \\ &= \max\left\{\langle X, Z\rangle - \frac{1}{2}\|Z\|^2 :\ Z \in \mathbb{M}_r\right\}. \end{aligned}$$

Being the pointwise maximum of linear functions when $Z \in \mathbb{M}_r$ is given, $h_r(X)$ is convex. The convexity of $h_r(X)$ and (12) yields

$$h_r(X) \geq h_r(Z) + \langle M,\ X - Z\rangle,\ \ \forall\, X, Z \in \mathbb{R}^{k\times\ell},\ \ M \in \mathcal{P}_{\mathbb{M}_r}(Z) \tag{20}$$

which further implies

$$\begin{aligned} g_r(X) &= \frac{1}{2}\|X\|^2 - h_r(X) \\ &\leq \frac{1}{2}\|X\|^2 - h_r(Z) - \langle \Pi_{\mathbb{M}_r}(Z),\ X - Z\rangle \\ &= \frac{1}{2}\|X - \Pi_{\mathbb{M}_r}(Z)\|^2 \\ &\quad -\frac{1}{2}\Big(\underbrace{\|\Pi_{\mathbb{M}_r}(Z)\|^2 - \|Z\|^2 + \|Z - \Pi_{\mathbb{M}_r}(Z)\|^2 - 2\langle \Pi_{\mathbb{M}_r}(Z),\ Z\rangle}_{=0}\Big) \\ &= \frac{1}{2}\|X - \Pi_{\mathbb{M}_r}(Z)\|^2 = g_r^{(m)}(X, Z). \end{aligned}$$

In other words, $g_r(X)$ can be seen as Difference of Convex functions, the so-called DC function. Using a subgradient is a common way to majorize DC functions, see [12].

### 3.2 The pMAP algorithm

We recall that our main problem is (7). Our first step is to construct a majorized function of $F_\rho(X)$ at the current iterate $X^\nu$:

$$\begin{aligned} & F_\rho^{(m)}(X, X^\nu) = \frac{1}{2}\|W \circ (X - A)\|^2 + \rho g_r^{(m)}(X, X^\nu) \\ &= \frac{1}{2}\|W \circ (X - A)\|^2 + \frac{\rho}{2}\|X - \Pi_{\mathbb{M}_r}(X^\nu)\|^2 \\ &= \frac{1}{2}\|W \circ X\|^2 + \frac{\rho}{2}\|X\|^2 - \langle W^{(2)} \circ A + \rho\Pi_{\mathbb{M}_r}(X^\nu),\ X\rangle + \frac{1}{2}\|W \circ A\|^2 \\ &\quad + \frac{\rho}{2}\|\Pi_{\mathbb{M}_r}(X^\nu)\|^2 \\ &= \frac{1}{2}\|\sqrt{\rho + W^{(2)}} \circ (X - X_\rho^\nu)\|^2 + \frac{1}{2}\|W \circ A\|^2 + \frac{\rho}{2}\|\Pi_{\mathbb{M}_r}(X^\nu)\|^2 - \frac{\rho + W^{(2)}}{2}\|X_\rho^\nu\|^2, \end{aligned}$$

where

$$X_\rho^\nu := \frac{\rho \Pi_{\mathbb{M}_r}(X^\nu) + W^{(2)} \circ A}{\rho + W^{(2)}}. \tag{21}$$

Note that the division is in the sense of componentwise. The subproblem to be solved at iteration $\nu$ is

$$\begin{aligned} X^{\nu+1} &= \arg\min \; F_\rho^{(m)}(X, X^\nu) \quad \text{s.t.} \;\; X \in \mathbb{H} \\ &= \arg\min \; \frac{1}{2}\|\sqrt{\rho + W^{(2)}} \circ (X - X_\rho^\nu)\|^2 \quad \text{s.t.} \;\; X \in \mathbb{H} \\ &= \Pi_{\mathbb{H}}(X_\rho^\nu), \end{aligned} \tag{22}$$

where $X_\rho^\nu$ is defined in (21). The last equation in (22) is due to $W_\rho := \sqrt{\rho + W^{(2)}}$ being Hankel and computing $X^{\nu+1}$ in (22) is equivalent to averaging $X_\rho^\nu$ along its all anti-diagonals. Since $A$, $\rho/(\rho + W^{(2)})$, $W^{(2)}/(\rho + W^{(2)})$ are all Hankel matrices (due to Assumption 1), $X^{\nu+1}$ can be calculated through (9).

**Algorithm 1** (pMAP)

1: **Input data:** Matrix $A$, weight matrix $W$, penalty parameter $\rho$, rank $r$, and the initial $X^0$. Set $\nu := 0$.
2: **Update** $X$: Compute $X^{\nu+1}$ by (9).
3: **Convergence check:** Terminate if some stopping criterion is satisfied.

***Remark 2*** Being a direct consequence of employing the majorization technique, the following decreasing property holds:

$$\begin{aligned} F_\rho(X^{\nu+1}) &\le F_\rho^{(m)}(X^{\nu+1}, X^\nu) \quad \text{(property of majorization (19))} \\ &\le F_\rho^{(m)}(X^\nu, X^\nu) \qquad \text{(because of (22))} \\ &= F_\rho(X^\nu) \qquad\qquad \text{(property of majorization (19)).} \end{aligned}$$

If $F_\rho$ is coercive (i.e., $F_\rho(X) \to \infty$ if $\|X\| \to \infty$, which would be the case if we require $W > 0$), the sequence $\{X^\nu\}$ will be bounded.

A widely used stopping criterion is $\|X^{\nu+1} - X^\nu\| \le \epsilon$ for some small tolerance $\epsilon > 0$. We will see below that $\|X^{\nu+1} - X^\nu\|$ approaches zero and hence such convergence check will eventually be satisfied. For our theoretical analysis, we assume that pMAP generates an infinite sequence (e.g., let $\epsilon = 0$).

### 3.3 Convergence of pMAP

We have more results on the convergence of pMAP.

**Theorem 2** *Let $\{X^\nu\}$ be the sequence generated by pMAP. The following holds.*

*(i) We have*

$$F_\rho(X^{\nu+1}) - F_\rho(X^\nu) \le -\frac{\rho}{2}\|X^{\nu+1} - X^\nu\|^2, \quad k = 1, 2, \ldots, .$$

*Furthermore, $\|X^{\nu+1} - X^\nu\| \to 0$.*

*(ii) Let $\widehat{X}$ be an accumulation point of $\{X^\nu\}$. We then have*

$$\nabla f(\widehat{X}) + \rho(\widehat{X} - \Pi_{\mathbb{M}_r}(\widehat{X})) \in \mathbb{H}^\perp.$$

*Moreover, for a given $\epsilon > 0$, if $X^0 \in \mathbb{M}_r \cap \mathbb{H}$ and*

$$\rho \ge \rho_\epsilon := \frac{f(X^0)}{\epsilon},$$

*then $\widehat{X}$ is an $\epsilon$-approximate KKT point of (1).*

***Proof*** We will use a number of facts to establish (i). The first fact is due to the convexity of $f(X)$:

$$f(X^\nu) - f(X^{\nu+1}) \ge \langle \nabla f(X^{\nu+1}), X^\nu - X^{\nu+1} \rangle \tag{23}$$

The second fact is the identity

$$\|X^{\nu+1}\|^2 - \|X^\nu\|^2 = 2\langle X^{\nu+1} - X^\nu, X^{\nu+1} \rangle - \|X^{\nu+1} - X^\nu\|^2 \tag{24}$$

The third fact is due to the convexity of $h_r(X)$ defined in (11) and $\Pi_{\mathbb{M}_r}(X) \in \partial h_r(X)$:

$$h_r(X^{\nu+1}) - h_r(X^\nu) \ge \langle \Pi_{\mathbb{M}_r}(X^\nu), X^{\nu+1} - X^\nu \rangle \tag{25}$$

The last fact is the optimality condition of problem (22):

$$\nabla f(X^{\nu+1}) + \rho(X^{\nu+1} - \Pi_{\mathbb{M}_r}(X^\nu)) \in \mathbb{H}^\perp. \tag{26}$$

Combining all facts above leads to a sufficient decrease in $F_\rho(X^k)$:

$$\begin{aligned}
&F_\rho(X^{\nu+1}) - F_\rho(X^\nu) \\
&= f(X^{\nu+1}) - f(X^\nu) + \rho g_r(X^{\nu+1}) - \rho g_r(X^\nu) \\
&\overset{(23)}{\leq} \langle \nabla f(X^{\nu+1}), X^{\nu+1} - X^\nu \rangle + \rho g_r(X^{\nu+1}) - \rho g_r(X^\nu) \\
&= \langle \nabla f(X^{\nu+1}), X^{\nu+1} - X^\nu \rangle + \frac{\rho}{2}(\|X^{\nu+1}\|^2 - \|X^\nu\|^2) - \rho(h_r(X^{\nu+1}) - h_r(X^\nu)) \\
&\overset{(24)}{=} \langle \nabla f(X^{\nu+1}) + \rho X^{\nu+1}, X^{\nu+1} - X^\nu \rangle - \frac{\rho}{2}(\|X^{\nu+1} - X^\nu\|^2) \\
&\quad - \rho(h_r(X^{\nu+1}) - h_r(X^\nu)) \\
&\overset{(25)}{\leq} \langle \nabla f(X^{\nu+1}) + \rho X^{\nu+1} - \rho \Pi_{\mathbb{M}_r}(X^\nu), X^{\nu+1} - X^\nu \rangle - \frac{\rho}{2}\|X^{\nu+1} - X^\nu\|^2 \\
&\overset{(26)}{\leq} -\frac{\rho}{2}\|X^{\nu+1} - X^\nu\|^2
\end{aligned} \tag{27}$$

In the above we also used the fact that $X^{\nu+1} - X^\nu \in \mathbb{H}$. Since the sequence $\{F_\rho(X^\nu)\}$ is non-increasing and is bounded from below by 0, we have $\|X^{\nu+1} - X^\nu\|^2 \to 0$.

(ii) Suppose $\widehat{X}$ is the limit of the subsequence $\{X^\nu\}_{\nu\in\mathcal{K}}$. It follows from $\|X^{\nu+1} - X^\nu\| \to 0$ that $\widehat{X}$ is also the limit of $\{X^{\nu+1}\}_{\nu\in\mathcal{K}}$. Taking limits on both sides of (26) and using the upper semi-continuity of the projections $\mathcal{P}_{\mathbb{M}_r}(\cdot)$ yields

$$\nabla f(\widehat{X}) + \rho(\widehat{X} - \Pi_{\mathbb{M}_r}(\widehat{X})) \in \mathbb{H}^\perp.$$

we use the fact that $\{F_\rho(X^\nu)\}$ is non-increasing to get

$$\begin{aligned}
f(X^0) = f(X^0) + \rho g_r(X^0) = F_\rho(X^0) &\geq \lim F_\rho(X^\nu) \\
&= F_\rho(\widehat{X}) = f(\widehat{X}) + \rho g_r(\widehat{X}) \geq \rho g_r(\widehat{X}).
\end{aligned}$$

The first equality holds because $g_r(X^0) = 0$ when $X^0 \in \mathbb{M}_r$. As a result,

$$g_r(\widehat{X}) \leq \frac{f(X^0)}{\rho} \leq \frac{f(X^0)}{\rho_\epsilon} = \epsilon. \tag{28}$$

Therefore, $\widehat{X}$ is an $\epsilon$-approximate KKT point of (1). □

We note that the first result (i) in Theorem 2 is standard in a majorization-minimization scheme and can be proved in different ways, see, e.g., [32, Thm. 3.7].

### 3.4 Final rank and linear convergence

This part reports two results. One is on the final rank of the output of pMAP and the rank is always bigger than the desired rank $r$ unless $A$ is already an optimal solution

of (1). The other is on the conditions that ensure a linear convergence rate of pMAP. For this purpose, we need the following result.

**Proposition 1** *[11, Thm. 25] Given the integer $r > 0$ and consider $\widehat{X} \in \mathbb{R}^{k\times\ell}$ of rank $(r + p)$ with $p \geq 0$. Suppose the SVD of $\widehat{X}$ is represented as $\widehat{X} = \sum_{i=1}^{r+p} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, where $\sigma_1(\widehat{X}) \geq \sigma_2(\widehat{X}) \geq \cdots \geq \sigma_{r+p}(\widehat{X})$ are the singular values of $\widehat{X}$ and $\mathbf{u}_i, \mathbf{v}_i$, $i = 1, \ldots, r + p$ are the left and right (normalized) eigenvectors. We assume $\sigma_r(\widehat{X}) > \sigma_{r+1}(\widehat{X})$ so that the projection operator $\Pi_{\mathbb{M}_r}(X)$ is uniquely defined in a neighbourhood of $\widehat{X}$. Then $\Pi_{\mathbb{M}_r}(X)$ is differentiable at $\widehat{X}$ and the directional derivative along the direction $Y$ is given by*

$$\nabla\ \Pi_{\mathbb{M}_r}(\widehat{X})(Y) = \Pi_{\mathcal{T}_{\mathbb{M}_r}(\widehat{X})}(Y) + \sum_{\substack{1\leq i\leq r\\ 1\leq j\leq p}} \left[\frac{\sigma_{r+j}}{\sigma_i - \sigma_{r+j}}\langle Y,\ \Phi^+_{i,r+j}\rangle \Phi^+_{i,r+j} - \frac{\sigma_{r+j}}{\sigma_i + \sigma_{r+j}}\langle Y,\ \Phi^-_{i,r+j}\rangle \Phi^-_{i,r+j}\right]$$

*where $\mathcal{T}_{\mathbb{M}_r}(\widehat{X})$ is the tangent subspace of $\mathbb{M}_r$ at $\widehat{X}$ and*

$$\Phi^{\pm}_{i,r+j} = \frac{1}{\sqrt{2}}\left(\mathbf{u}_{r+j}\mathbf{v}_i^T \pm \mathbf{u}_i\mathbf{v}_{r+j}^T\right).$$

**Theorem 3** *Assume that $W > 0$ and $\widehat{X}$ be an accumulation point of $\{X^\nu\}$. The following hold.*

*(i) $\mathrm{rank}(\widehat{X}) > r$ unless $A$ is already the optimal solution of (1).*

*(ii) Suppose $\widehat{X}$ has rank $(r + p)$ with $p > 0$. Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$ be the singular values of $\widehat{X}$. Define*

$$w_0 := \min\{W_{ij}\} > 0, \quad \epsilon_0 := \frac{w_0^2}{\rho}, \quad \epsilon_1 := \frac{\epsilon_0}{4 + 3\epsilon_0}, \quad c := \frac{1}{1 + \epsilon_1} < 1.$$

*Under the condition*

$$\frac{\sigma_r}{\sigma_{r+1}} \geq \frac{8pr}{\epsilon_0} + 1,$$

*it holds*

$$\|X^{\nu+1} - \widehat{X}\| \leq c\|X^\nu - \widehat{X}\| \quad \textit{for } \nu \textit{ sufficiently large}.$$

*Consequently, the whole sequence $\{X^\nu\}$ converges linearly to $\widehat{X}$.*

***Proof*** (i) Suppose $\widehat{X}$ is the limit of the subsequence $\{X^\nu\}_{k\in\mathcal{K}}$. We assume $\mathrm{rank}(\widehat{X}) \leq r$. It follows from Theorem 2 that

$$\{X^{\nu+1}\}_{k\in\mathcal{K}} \to \widehat{X} \quad \text{and} \quad \lim_{k\in\mathcal{K}} \Pi_{\mathbb{M}_r}(X^\nu) = \Pi_{\mathbb{M}_r}(\widehat{X}) = \widehat{X}.$$

Taking limits on both sides of (9) and using the fact that $\widehat{X}$ is Hankel, we get

$$\widehat{X} = \frac{W^{(2)}}{\rho + W^{(2)}} \circ A + \frac{\rho}{\rho + W^{(2)}} \circ \Pi_{\mathbb{H}}(\Pi_{\mathbb{M}_r}(\widehat{X})) = \frac{W^{(2)}}{\rho + W^{(2)}} \circ A + \frac{\rho}{\rho + W^{(2)}} \circ \widehat{X}.$$

Under the assumption $W > 0$, we have $\widehat{X} = A$. Consequently, $\text{rank}(A) \leq r$, implying that $A$ is the optimal solution of (1). Therefore, we must have $\text{rank}(\widehat{X}) > r$ if the given matrix $A$ is not optimal already.

(ii) Let $\phi(X) := \Pi_{\mathbb{H}}(\Pi_{\mathbb{M}_r}(X))$. Since $\Pi_{\mathbb{M}_r}(X)$ is differentiable at $\widehat{X}$, so is $\phi(X)$. Moreover, the directional derivative of $\phi(X)$ at $\widehat{X}$ along the direction $Y$ is given by

$$\nabla\phi(\widehat{X})Y = \Pi_{\mathbb{H}}(\nabla\Pi_{\mathbb{M}_r}(\widehat{X})Y) \quad \text{and} \quad \|\nabla\phi(\widehat{X})Y\| \leq \|\nabla\Pi_{\mathbb{M}_r}(\widehat{X})Y\|. \tag{29}$$

The inequality above holds because $\Pi_{\mathbb{H}}(\cdot)$ is an orthogonal projection to a subspace and its operator norm is 1. The matrices in Proposition 1 have the following bounds.

$$\|\Phi^{\pm}_{i,r+j}\| \leq \frac{1}{\sqrt{2}}\Big(\|\mathbf{u}_{r+j}\mathbf{v}_i^T\| + \|\mathbf{u}_i\mathbf{v}_{r+j}^T\|\Big) \leq \frac{1}{\sqrt{2}}(1+1) = \sqrt{2},$$
$$\|\langle Y,\ \Phi^{\pm}_{i,r+j}\rangle\Phi^{\pm}_{i,r+j}\| \leq \|\Phi^{\pm}_{i,r+j}\|^2\|Y\| \leq 2\|Y\|.$$

Therefore,

$$\begin{aligned}&\left\| \sum_{\substack{1\leq i\leq r\\ 1\leq j\leq p}} \left[\frac{\sigma_{r+j}}{\sigma_i - \sigma_{r+j}}\langle Y,\ \Phi^{+}_{i,r+j}\rangle\Phi^{+}_{i,r+j} - \frac{\sigma_{r+j}}{\sigma_i + \sigma_{r+j}}\langle Y,\ \Phi^{-}_{i,r+j}\rangle\Phi^{-}_{i,r+j}\right]\right\| \\ &\leq 4\sum_{\substack{1\leq i\leq r\\ 1\leq j\leq p}} \frac{\sigma_{r+j}}{\sigma_i - \sigma_{r+j}}\|Y\| \leq 4pr\frac{\sigma_{r+1}}{\sigma_r - \sigma_{r+1}}\|Y\| \leq \frac{w_0^2}{2\rho}\|Y\| = \frac{1}{2}\epsilon_0\|Y\|. \end{aligned} \tag{30}$$

In the above, we used the fact that $\psi(t) := t/(\sigma_r - t)$ is an increasing function of $t$ for $t < \sigma_r$. Proposition 1, (29) and (30) imply

$$\|\nabla\phi(\widehat{X})Y\| \leq \|\Pi_{\mathcal{T}_{\mathbb{M}_r}(\widehat{X})}(Y)\| + \epsilon_0/2\|Y\| \leq \|Y\| + \epsilon_0/2\|Y\| \leq (1+\epsilon_0/2)\|Y\|.$$

The second equality above used the fact that the operator norm of $\Pi_{\mathcal{T}_{\mathbb{M}_r}(\widehat{X})}$ is not greater than 1 due to $\mathcal{T}_{\mathbb{M}_r}(\widehat{X})$ being a subspace. Since $\phi(\cdot)$ is differentiable at $\widehat{X}$, there exists $\epsilon > 0$ such that

$$\|\phi(X) - \phi(\widehat{X}) - \nabla\phi(\widehat{X})(X - \widehat{X})\| \leq \frac{1}{4}\epsilon_0\|X - \widehat{X}\|, \quad \forall\ X \in \mathcal{B}_\epsilon(\widehat{X}).$$

Therefore,

$$\|\phi(X) - \phi(\widehat{X})\| \leq \|\phi(X) - \phi(\widehat{X}) - \nabla\phi(\widehat{X})(X - \widehat{X})\| + \|\nabla\phi(\widehat{X})(X - \widehat{X})\|$$

$$\leq \frac{1}{4}\epsilon_0\|X-\widehat{X}\| + (1+\epsilon_0/2)\|X-\widehat{X}\| = (1+3\epsilon_0/4)\|X-\widehat{X}\|.$$

Now we are ready to quantify the error between $X^\nu$ and $\widehat{X}$ whenever $X^\nu \in \mathcal{B}_\epsilon(\widehat{X})$.

$$\begin{aligned}\|X^{\nu+1}-\widehat{X}\| &= \left\| \frac{\rho}{\rho + W^{(2)}} \circ (\phi(X^\nu)-\phi(\widehat{X})) \right\| \leq \frac{\rho}{\rho + w_0^2}\|\phi(X^\nu)-\phi(\widehat{X})\| \\ &\leq \frac{1+3\epsilon_0/4}{1+\epsilon_0}\|X^\nu-\widehat{X}\| = c\|X^\nu-\widehat{X}\|.\end{aligned}$$

Consequently, $X^{\nu+1} \in \mathcal{B}_\epsilon(\widehat{X})$. Since $\{X^\nu\}_{\nu\in\mathcal{K}}$ converges to $\widehat{X}$, $X^\nu$ will eventually falls in $\mathcal{B}_\epsilon(\widehat{X})$, which implies that the whole sequence $\{X^\nu\}$ will converge to $\widehat{X}$ and eventually converges at a linear rate. □

***Remark 3*** (Implication on MAP) When the weight matrix $W = 0$, pMAP reduces to MAP according to (9). Theorem 2(i) implies

$$\|X^{\nu+1} - \Pi_{\mathbb{M}_r}(X^{\nu+1})\|^2 - \|X^\nu - \Pi_{\mathbb{M}_r}(X^\nu)\|^2 \leq -\|X^{\nu+1}-X^\nu\|^2, \tag{31}$$

which obviously implies

$$\|X^{\nu+1} - \Pi_{\mathbb{M}_r}(X^{\nu+1})\| \leq \|X^\nu - \Pi_{\mathbb{M}_r}(X^\nu)\|. \tag{32}$$

The decrease property (32) was known in [5, Eq.(4.1)] and was used there to ascertain that MAP is a descent algorithm. Our improved bound (31) says a lightly more that the decrease each step in the function $\|X - \Pi_{\mathbb{M}_r}(X)\|$ is strict unless the update becomes unchanged. In this case ($W = 0$), the penalty parameter is just a scaling factor in the objective, hence the KKT result in Theorem 2(ii) does not apply to MAP. This probably explains why it is difficult to establish similar results for MAP.

***Remark 4*** (On linear convergence) In the general context of matrix completion, Lai and Varghese [20] established a local linear convergence of MAP under the following two assumptions. We describe them in terms of the Hankel matrix completion. (i) The partially observed data **a** can be completed to a rank $r$ Hankel matrix $M$. (ii) A transversality condition (see [20, Thm. 2]) holds at $M$. We emphasize that the result of [20] is a local result that requires that the initial point of MAP is close enough to $M$ and the rank $r$ assumption of $M$ is also crucial to their analysis, which also motivated our proof. In contrast, our result is a global one and enjoys a linear convergence rate near the limit under a more realistic assumption $\sigma_r \gg \sigma_{r+1}$. One may have noticed that the convergence rate $c$ though strictly less than 1 may be close to 1. This is often numerically observed that MAP often converges slowly. But the more important point here is that in such a situation it ensures that the whole sequence converges. This global convergence justifies the widely used stopping criterion $\|X^{\nu+1} - X^\nu\| \leq \epsilon$.

## 4 Extension to complex-valued matrices

The results obtained in the previous sections are for real-valued matrices and they can be extended to complex-valued matrices by employing what is known as the Wirtinger calculus [30]. We note that not all algorithms for Hankel matrix optimization have a straightforward extension from the real case to the complex case, see [6] for comments on some algorithms. We explain our extension below.

Suppose $f : \mathbb{C}^n \mapsto \mathbb{R}$ is a real-valued function in the complex domain. We write $\mathbf{z} \in \mathbb{C}^n$ as $\mathbf{z} = \mathbf{x} + j\mathbf{y}$ with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. The conjugate $\bar{\mathbf{z}} := \mathbf{x} - j\mathbf{y}$. Then we can write the function $f(\mathbf{z})$ in terms of its real variables $\mathbf{x}$ and $\mathbf{y}$. With a slight abuse of notation, we still denote it as $f(\mathbf{x}, \mathbf{y})$. In the case where the optimization of $f(\mathbf{z})$ can be equivalently represented as optimization of $f$ in terms of its real variables, the partial derivatives $\partial f(\mathbf{x}, \mathbf{y})/\partial \mathbf{x}$ and $\partial f(\mathbf{x}, \mathbf{y})/\partial \mathbf{y}$ would be sufficient. For other cases where algorithms are preferred to be executed in the complex domain, then the Wirtinger calculus [30] is more convenient to use and it is well explained (and derived) in [19]. The $\mathbb{R}$-derivative and the conjugate $\mathbb{R}$-derivative of $f$ in the complex domain are defined respectively by

$$\frac{\partial f}{\partial \mathbf{z}} = \frac{1}{2}\left(\frac{\partial f}{\partial \mathbf{x}} - \mathbf{j}\frac{\partial f}{\partial \mathbf{y}}\right), \quad \frac{\partial f}{\partial \bar{\mathbf{z}}} = \frac{1}{2}\left(\frac{\partial f}{\partial \mathbf{x}} + \mathbf{j}\frac{\partial f}{\partial \mathbf{y}}\right).$$

The $\mathbb{R}$-derivatives in the complex domain play the same role as the derivatives in the real domain because the following two first-order expansions are equivalent:

$$\begin{aligned} f(\mathbf{x} + \Delta\mathbf{x}, \mathbf{y} + \Delta\mathbf{y}) &= f(\mathbf{x}, \mathbf{y}) + \langle \partial f/\partial \mathbf{x},\ \Delta\mathbf{x}\rangle \\ &\quad + \langle \partial f/\partial \mathbf{y},\ \Delta\mathbf{y}\rangle + o(\|\Delta\mathbf{x}\| + \|\Delta\mathbf{y}\|) \\ f(\mathbf{z} + \Delta\mathbf{z}) &= f(\mathbf{z}) + 2\mathbf{Re}(\langle @f/@\mathbf{z},\ \mathbf{z}\rangle) + o(\|\mathbf{z}\|). \end{aligned} \tag{33}$$

Here, we treat the partial derivatives as column vectors and $\mathbf{Re}(\mathbf{x})$ is the real part of $\mathbf{x}$. Note that in the first-order expansion in $f(\mathbf{z} + \Delta\mathbf{z})$ used the conjugate $\mathbb{R}$-derivative. Hence, we define the complex gradient to be $\nabla f(\mathbf{z}) := 2\partial f/\partial \bar{\mathbf{z}}$, when it exists. When $f$ is not differentiable, we can extend the subdifferential of $f$ from the real case to the complex case by generalizing (33).

In order to extend Theorem 1, we need to characterize $\partial d_{\mathbb{M}_r}(X)$ in the complex domain. We may follow the route of [26] to conduct the extension. For example, we may define the regular subgradient of $d_{\mathbb{M}_r}(X)$ [26, Def. 8.3] to its complex counterpart by replacing the conjugate-gradient in the first-order expansion in (33) by a regular subgradient. We then define subdifferential through regular subgradients. With this definition in the complex domain, we may extend [26, (8.53)] to derive formulae for $\partial d_{\mathbb{M}_r}(X)$. What we needed in the proof of Theorem 1 is $(X - \Pi_{\mathbb{M}_r}(X))/d_{\mathbb{M}_r}(X) \in \partial d_{\mathbb{M}_r}(X)$ when $X \notin \mathbb{M}_r$. The proof of this result follows a straightforward extension of the corresponding part in [26, (8.53)] and if reproduced here would take up much space. Hence we omit it.

In order to extend the results in Sect. 3, we need the subdifferential of $h_r(X)$ in order to majorize $g_r(X)$. Since $h_r(X)$ is convex, its subdifferential is easy to define. We

note that the inner product for the complex matrix space $\mathbb{C}^{k\times\ell}$ is defined as $\langle A,\ B\rangle = \mathrm{Trace}(A^H B)$, where $A$ is the Hermitian conjugate, i.e., $A^H := \overline{A}^T$.

**Definition 3** The subdifferential $\partial h_r(X)$ is defined as

$$\partial h_r(X) = \left\{ S \in \mathbb{C}^{k\times\ell} \mid h_r(Z) \geq h_r(X) + \mathbf{Re}\Big(\langle \mathrm{S},\ \mathrm{Z}-\mathrm{X}\rangle\Big) \right\}.$$

The following result is really what we needed in order to extend the results in Sect. 3 to the complex domain.

**Proposition 2** *For any $X \in \mathbb{C}^{k\times\ell}$, we have $\mathcal{P}_{\mathbb{M}_r}(X) \subset \partial h_r(X)$.*

***Proof*** Let $\Pi_{\mathbb{M}_r}(X)$ stand for any element in $\mathcal{P}_{\mathbb{M}_r}(X)$. It is easy to verify the following identities:

$$\|X-Z\|^2 = \|X\|^2 + \|Z\|^2 - 2\mathbf{Re}\Big(\langle \mathrm{X},\ \mathrm{Z}\rangle\Big) = \|\mathrm{X}\|^2 + \|\mathrm{Z}\|^2 - 2\mathbf{Re}\Big(\langle \mathrm{Z},\ \mathrm{X}\rangle\Big) \quad (34)$$

We use (11) and (34) to compute

$$\begin{aligned}
&h_r(Z) - h_r(X) - \mathbf{Re}\Big(\dot{}_{\mathbb{M}_r}(\mathrm{X}),\ \mathrm{Z}-\mathrm{X}\Big)\\
&= \frac{1}{2}\|\Pi_{\mathbb{M}_r}(Z)\|^2 - \frac{1}{2}\|\Pi_{\mathbb{M}_r}(X)\|^2 + \underbrace{\frac{1}{2}\|\Pi_{\mathbb{M}_r}(X)-Z\|^2 - \frac{1}{2}\|\Pi_{\mathbb{M}_r}(X)\|^2 - \frac{1}{2}\|Z\|^2}_{=-\mathbf{Re}(\langle \dot{}_{\mathbb{M}_r}(\mathrm{X}),\ \mathrm{Z}\rangle)}\\
&\quad + \underbrace{\frac{1}{2}\|\Pi_{\mathbb{M}_r}(X)\|^2 + \frac{1}{2}\|X\|^2 - \frac{1}{2}\|\Pi_{\mathbb{M}_r}(X)-X\|^2}_{=\mathbf{Re}(\langle \dot{}_{\mathbb{M}_r}(\mathrm{X}),\ \mathrm{X}\rangle)}\\
&= \frac{1}{2}\|\Pi_{\mathbb{M}_r}(Z)\|^2 - \frac{1}{2}\|Z\|^2 + \frac{1}{2}\|\Pi_{\mathbb{M}_r}(X)-Z\|^2\\
&\quad - \underbrace{\left(\frac{1}{2}\|\Pi_{\mathbb{M}_r}(X)\|^2 - \frac{1}{2}\|X\|^2 + \frac{1}{2}\|\Pi_{\mathbb{M}_r}(X)-X\|^2\right)}_{=0}\\
&= \frac{1}{2}\|\Pi_{\mathbb{M}_r}(Z)\|^2 - \frac{1}{2}\|Z\|^2 + \frac{1}{2}\|\Pi_{\mathbb{M}_r}(X)-Z\|^2\\
&\geq \frac{1}{2}\|\Pi_{\mathbb{M}_r}(Z)\|^2 - \frac{1}{2}\|Z\|^2 + \frac{1}{2}\|\Pi_{\mathbb{M}_r}(Z)-Z\|^2 = 0.
\end{aligned}$$

This proves the claim. □

A direct consequence is that

$$\partial g_r(X) = X - \partial h_r(X) \supset \mathcal{P}_{\mathbb{M}_r}(X)$$

and the majorization $g_r(X)$ through the subdifferential of $h_r(X)$ holds. The rest of the extension is straightforward and we do not repeat it here.

## 5 Numerical experiments

In this section we test two popular problems (time series denoising in real domain and incomplete signal completion in complex domain) to demonstrate the numerical performance of pMAP. The time series denoising problem aims to extract the noiseless data from polluted observations by removing the noise components, while incomplete signal completion problem tries to approximate the missing data in an incomplete complex valued signal.

In both numerical experiments, a solver is terminated when any of the following conditions is met

$$\frac{|F_\rho(X^{\nu+1}) - F_\rho(X^\nu)|}{\max\{1, F_\rho(X^\nu)\}} \leq ftol, \quad g_r(X^{\nu+1}) \leq gtol \quad \text{or} \quad \frac{\|X^{\nu+1} - X^\nu\|}{\|X^\nu\|} \leq tol.$$

here $tol$, $ftol$, $gtol$ are set at $10^{-5}$, $10^{-7}$ and $10^{-8}$, respectively. A solver will also be terminated if it reaches the maximum iterations, setting at 200. All codes used were written in MATLAB (2019a) and run on a laptop equipped with a 7th Generation Intel Core i5-7200U CPU and 8GB memory card.

### 5.1 Time series denoising

#### 5.1.1 Experiment introduction

In the first experiment we compare the proposed pMAP with some leading solvers including Cadzow's method (Cadzow, [14]) and Douglas-Rachford iterations (DRI, [6]) for real-valued time series de-noising. In this test we randomly generate noiseless time series $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ via the following process:

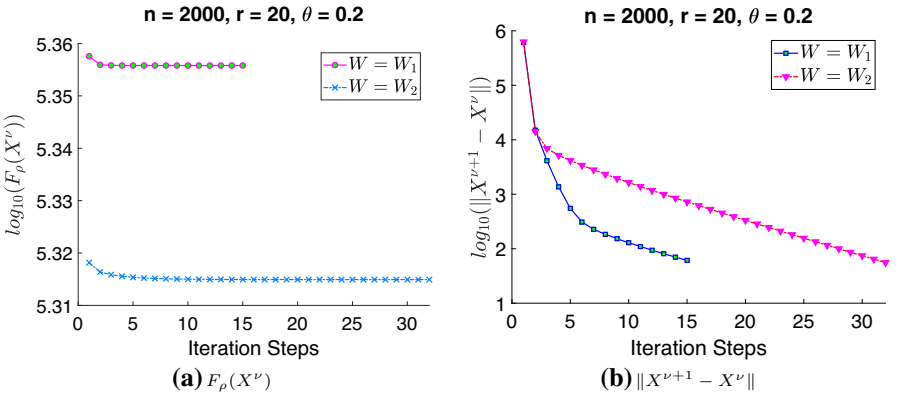
$$a_t = \sum_{s=1}^{r} d_s(1+\alpha_s)^t cos(2\pi t/\beta_s - \tau_s), \quad \text{for } t = 1, 2, \ldots, n$$

where all $d_s$, $\alpha_s$, $\beta_s$ and $\tau_s$ follow uniform distribution as $d_s \sim U[0, 10^3)$, $\alpha_s \sim U[-10^{-3}, 10^{-3})$, $\beta_s \sim U[6, 18)$ and $\tau_s \sim U[-\pi, \pi)$. It is known that for any $\{l, k\}$ such that $l + k - 1 = n$, the rank of Hankel matrix $A = \mathcal{T}(\mathbf{a}) \in \mathbb{R}^{l\times k}$ must be $2r$ when both $l$ and $k$ are no smaller than $2r$. We then construct the noisy time series $\mathbf{y}$ by adding the noises series $\epsilon$ to $\mathbf{a}$ as $\mathbf{y} = \mathbf{a} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$ is the noise component and $\epsilon_t = \theta \frac{e_t}{\|e\|_2}\|a\|$. Here $e_t$ is the white noise with mean 0 and variance 1. We considered two scenarios: $\{n, r\} = \{1000, 10\}$ and $\{2000, 20\}$. For each scenario we test three noise levels at $\theta = 0.1$, 0.2 and 0.5.

We further consider two weight choices:

1. $\{W_1\}_{i,j} = 1$, for $i = 1, \ldots, l$ and $j = 1, \ldots, k$;
2. $\{W_2\}_{i,j} = \frac{1}{i+j-1}$, for $i = 1, \ldots, l$ and $j = 1, \ldots, k$.

Both weights are standardised for comparison purpose (i.e., $W/\|W\|$ was used). Note that Cadzow's method can only handle $W_1$.




**(a)** $F_\rho(X^\nu)$ **(b)** $\|X^{\nu+1} - X^\nu\|$

**Fig. 1** Plot of $F_\rho(X^\nu)$ and $\|X^{\nu+1} - X^\nu\|$ by pMAP with $\rho$ fixed

### 5.1.2 Demonstration of convergence

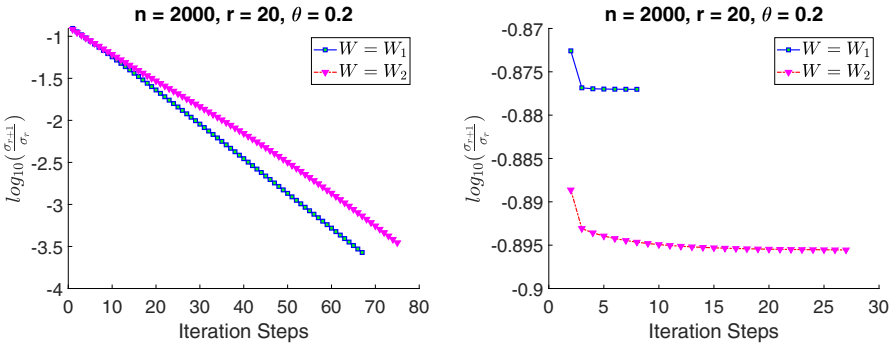
Before coming to the numerical comparison, we first demonstrate the convergence behaviour of Algorithm 1 under different updating strategy of $\rho$. We plot the sequences of $F_\rho(X^\nu)$ in Fig. 1a with both $W_1$ and $W_2$. It can be observed that in both cases, the functional value $F_\rho(X^\nu)$ decreases and converges. We further plot the sequence of $\|X^{\nu+1} - X^\nu\|$ in Fig. 1b using the same example. We find that the sequence of $\|X^{\nu+1} - X^\nu\|$ is also decreasing and converges to zero, which is consistent with Theorem 2.

The behaviours of $\frac{\sigma_{r+1}}{\sigma_r}$ are shown in Fig. 2 with respect to different $\rho$ updating strategies. In this and later experiments, $\rho$ is initialised as $\rho^0 = 10^{-2} \times m/n^2$ where $n$ denotes the length of a time series and $m$ stands for the amount of known observations, which equals to $n$ in this test. As shown in Fig. 2a, $\frac{\sigma_{r+1}}{\sigma_r}$ approaches zero with increasing $\rho$, which means $g_r(X^\nu)$ goes to zero as well. By contrast if $\rho$ is fixed as $\rho^\nu = \rho^0$ at each iterate, Fig. 2b shows that $\frac{\sigma_{r+1}}{\sigma_r}$ decreases much slower than the first strategy. As a result, we will update $\rho$ by $\rho^{\nu+1} = 1.1\rho^\nu$ at each iterate when $\rho^\nu \leq n \times \min(W)$, where $\min(W)$ is the minimal weights in $W$. The behaviour of convergence of Algorithm 1 appears consistence for other choices.

### 5.1.3 Numerical results

The numerical results are reported in Table 1 including the number of iterations (Iter), cpu time for computation (Time), root of mean square error (RMSE) for each solver which is calculated as

$$\text{RMSE} := \sqrt{\sum_{i \in \mathcal{I}} (\hat{x}_i - a_i)^2 / |\mathcal{I}|}.$$




**(a)** $\frac{\sigma_{r+1}}{\sigma_r}$ with increasing $\rho$ **(b)** $\frac{\sigma_{r+1}}{\sigma_r}$ with fixed $\rho$

**Fig. 2** Plot of $\frac{\sigma_{r+1}}{\sigma_r}$ at each iterate by pMAP. $\rho$ is updated by $\rho^{\nu+1} = 1.1\rho^{\nu}$ in (**a**) and fixed without updating in (**b**)

where $\hat{\mathbf{x}} = \{\hat{x}_1, \ldots, \hat{x}_n\}$ is obtained as $\hat{\mathbf{x}} = \mathcal{H}^{-1}(\widehat{X})$ and $\widehat{X}$ is the estimated result from a certain solver. $\mathcal{I}$ denotes the index set of all data to be predicted while $|\mathcal{I}|$ stands for the size of $\mathcal{I}$. Apparently smaller RMSEs indicate better solution qualities. Because the Cadzow method does not allow users to select arbitrary weights apart from $W_1$, we will not report the numerical results for the Cadzow method under $W_2$.

We also report success rate (SR) of each solver in Table 1. One instance is successfully de-noised if the relative gap between the RMSE of approximated solution and the best possible RMSE is smaller than a pre-defined threshold, i.e.,

$$\frac{\text{RMSE}_{\text{approx}} - \text{RMSE}_{\text{best}}}{\text{RMSE}_{\text{best}}} \leq \text{threshold}$$

In this experiment we set threshold $= 10\%$. For any combinations of $\{n/r/\theta\}$, all data reported in Table 1 are the mean values over 50 randomly generated tests.

Our first observation on Table 1 is that when applying $W = W_1$, pMAP reports the best results in 3 examples out of 6 while DRI performs the best in the rest 3 examples. In general, Cadzow, DRI and pMAP have very similar performance on estimation accuracy under $W_1$ because they are MAP-based algorithms.

When the weight matrix is set as $W_2$, a significant improvement on the estimation accuracy can be observed for DRI and pMAP comparing to the case $W = W_1$. This result matches our expectation because $W_2$ assumes that all data have equal importance by sharing the same weight, while $W_1$ implies that data in the middle of a time series are more weighted than the data at both ends.

For all $\{n/r/\theta\}$ combinations, our proposed solver with $W_2$ always generated the estimation results with lowest RMSEs. It is also important to mention that our pMAP algorithm enjoys the most robust convergence result among all candidate solvers. As a result, we conclude that our proposed pMAP algorithm is competitive and effective in solving real-valued time series denoising problems.

**Table 1** Experiment results for Cadzow iteration, DRI (Douglas-Rachford iterations) and our proposed pMAP, including iterations (Iter), CPU time in seconds (Time), Root of mean square error (RMSE) and success rate (SR). Results in this table are the average of 50 trials

| $n/r/\theta$ | $W$ | | pMAP | Cadzow | DRI |
|---|---|---|---|---|---|
| 1000/10/0.1 | $W_1$ | Iter | 64.46 | 8.96 | 200.00 |
| | | Time | 3.77 | 0.31 | 20.68 |
| | | RMSE | 39.00 | 38.67 | 38.17 |
| | | SR | 1.00 | 1.00 | 0.98 |
| | $W_2$ | Iter | **71.72** | | 200.00 |
| | | Time | **3.14** | | 20.69 |
| | | RMSE | **33.43** | | 37.95 |
| | | SR | **1.00** | | 0.30 |
| 1000/10/0.2 | $W_1$ | Iter | 72.00 | 8.38 | 200.00 |
| | | Time | 4.56 | 0.27 | 21.14 |
| | | RMSE | 39.67 | 39.69 | 39.69 |
| | | SR | 1.00 | 1.00 | 1.00 |
| | $W_2$ | Iter | **78.36** | | 200.00 |
| | | Time | **3.70** | | 20.87 |
| | | RMSE | **34.19** | | 39.46 |
| | | SR | **1.00** | | 0.36 |
| 1000/10/0.5 | $W_1$ | Iter | 81.74 | 12.50 | 200.00 |
| | | Time | 4.95 | 0.44 | 20.41 |
| | | RMSE | 84.49 | 83.97 | 83.81 |
| | | SR | 1.00 | 1.00 | 0.98 |
| | $W_2$ | Iter | **88.80** | | 200.00 |
| | | Time | **4.04** | | 20.40 |
| | | RMSE | **73.68** | | 83.37 |
| | | SR | **1.00** | | 0.26 |
| 2000/20/0.1 | $W_1$ | Iter | 58.00 | 9.56 | 200.00 |
| | | Time | 20.39 | 2.66 | 233.61 |
| | | RMSE | 30.25 | 30.33 | 30.26 |
| | | SR | 1.00 | 1.00 | 1.00 |
| | $W_2$ | Iter | **64.60** | | 200.00 |
| | | Time | **20.44** | | 233.08 |
| | | RMS | **25.16** | | 30.14 |
| | | SR | **1.00** | | 0.10 |
| 2000/20/0.2 | $W_1$ | Iter | 65.00 | 14.96 | 200.00 |
| | | Time | 25.69 | 4.17 | 234.72 |
| | | RMSE | 52.71 | 52.26 | 52.20 |
| | | SR | 1.00 | 1.00 | 1.00 |
| | $W_2$ | Iter | **71.98** | | 200.00 |
| | | Time | **24.07** | | 238.24 |
| | | RMSE | **44.59** | | 52.01 |

**Table 1** continued

| $n/r/\theta$ | $W$ | | pMAP | Cadzow | DRI |
|---|---|---|---|---|---|
| | | SR | **1.00** | | 0.22 |
| 2000/20/0.5 | $W_1$ | Iter | 74.00 | 11.80 | 200.00 |
| | | Time | 29.47 | 3.40 | 234.85 |
| | | RMSE | 147.78 | 147.22 | 147.07 |
| | | SR | 1.00 | 1.00 | 1.00 |
| | $W_2$ | Iter | **81.20** | | 200.00 |
| | | Time | **26.98** | | 235.46 |
| | | RMSE | **127.24** | | 146.65 |
| | | SR | **1.00** | | 0.28 |

## 5.2 Spectral sparse signal recovery

### 5.2.1 Experiment introduction

In this experiment, we consider the problem of recovering missing values in an incomplete spectral sparse signal. We refer to [2,4] and the references therein for its background in recovering signals which are spectrally sparse via off-grid methodologies. We follow the suggestions in [2] to generate the experiment data $\mathbf{a} = \{a_1, a_2, \ldots a_n\}$ where

$$a_t = \sum_{s=1}^{r} d_s e^{2\pi j \omega_s t}, \quad \text{for } t \in \{0, 1, \ldots, n\}$$

where $j = \sqrt{-1}$, $r$ is the model order, $\omega_s$ is the frequency of each sinusoid and $d_s \neq 0$ is the weight of each sinusoid. Both $\omega_s$ and $d_s$ are randomly sampled following uniform distributions, as $\omega_s \sim U[0, 1)$ and $d_s \sim U[0, 2\pi)$. Indexes of missing data are randomly sampled following uniform distribution. In this experiment we introduce three sub-tests focusing on different purposes.

**Test a: Incomplete signal recovery without noises** In this sub-test we assume only a subset $\Omega$ of the sampling points $\{1, \ldots, n\}$ are observed and we aim to recover the signal by estimating the missing data. Here all observed data are noiseless. We use success rate (SR) to measure the performance of the candidate methods in incomplete signal recovery. We say the signal is successfully recovered if

$$\frac{\|\hat{\mathbf{x}} - a\|}{\|\mathbf{a}\|} \leq 10^{-3}$$

where $\hat{\mathbf{x}}$ is the estimated signal.

In this test, signal length $n$ is set to 499, 999 or 1999, respectively. The percentage of known observations $\frac{m}{n}$ is set to 30% or 60% where $m$ stands for the amount of known observations. All combinations of $\{n/m/r\}$ used in this test can be found in Table 2. To compare the performance of our proposed method, we further introduce

three state-of-art solvers in spectral sparse signal recovery including Atomic Norm Minimization (ANM [28]), Fast Iterative Hard Thresholding (FIHT [3]) and Projected Gradient Descent (PGD [2]). $W$ for each solver is defined as

$$W_{i,j} = \begin{cases} 1/\sqrt{i+j-1} & \text{for } i+j-1 = 1, \ldots, k-1 \\ 1/\sqrt{k} & \text{for } i+j-1 = k, \ldots, n-k+1 \\ 1/\sqrt{(n-i-j+2)} & \text{for } i+j-1 = n-k+2, \ldots, n, \end{cases}$$

if $a_{i+j-1}$ is observed, and $W_{i,j} = 0$ if it is missing.

**Test b: Incomplete signal recovery with noises** This sub-test still aims to recover an incomplete spectral sparse signal, but some observed data are noisy while others are noiseless. Here we follow the signal generating process the same as in **Test a**, however, $\frac{1}{3}m$ observations are polluted by random noises, as $y_i = a_i + \epsilon_i$ where $\epsilon_i = \theta \frac{e_t}{\|e\|_2}\|a\|$. Here $e_t$ is a complex stranded normal random variable and the noise level $\theta$ is set at 0.2. We assume the index set of polluted observations is known in advance.

The weight matrix $W$ is set as follows. We give polluted observations very small weights (say 1) and noiseless observations are assigned to much larger weights such as 100. The missing data in the signal are given zero weight. This weighting scheme is more reasonable than other choices as we place higher confidence in noiseless data, less confidence in noisy data and no confidence in missing values. It is worth noting that this weight matrix setting is for pMAP only because FIHT and PGD do not support flexible weight choices. The rest settings of this sub-test such as the definition of $\{n, m, r\}$ are interpreted the same way as in **Test a**.

**Test c: Recover missing data with inaccurately estimated rank** In both **Test a** and **Test b**, we assume the objective rank $r$ to be a known parameter. However, obtaining the true information of objective rank is quite challenging in many real applications. So in this sub-test we will examine the performances of candidate solvers in recovering the incomplete spectral sparse signals while the objective rank $r$ is incorrectly estimated. Signal length $n$ is set as 3999 and we assume that 30% data in a signal are randomly or accurately observed without noises. True rank $r$ is set as 15 but assumed to be unknown in this test. It means for each solver, we will try different estimated rank $\hat{r}$ ranges from 6 to 30. Success rates (SR) over 50 instances are reported to measure the performance of each solver.

### 5.2.2 Numerical results

**Test (a)** The numerical results of this test are listed in Table 2 including total iterations (Iter), CPU time in seconds (Time), RMSE and success rate (SR) for each solver. Among all solvers, ANM enjoys the best global convergence result because it is a convex relaxation method. However, the computational cost of ANM is much higher than the rest solvers and it runs out of memories when $n$ is larger than 500. At the same time, it fails to generate better results comparing with other solvers. Hence we do not report its performances in the rest part of this test problem. Although DRI performs slightly better than the Cadzow method in terms of accuracy, both of these two solvers failed to successfully recover any incomplete signals.

**Table 2** Numerical results for six different solvers on the incomplete signal recovery experiment including iterations (Iter), computational time (Time), estimation error (RMSE) and success recovery rate (SR). Results in this table are the average of 50 trials. *Experiment results ANM when $n \geq 999$ are not available because they run out of memory

| n/m/r | | Cadzow | DRI | ANM | FIHT | PGD | pMAP |
|---|---|---|---|---|---|---|---|
| *m/n = 30%* | | | | | | | |
| 499/150/10 | Iter | 10.04 | 200 | - | 28.42 | 50.36 | **45.36** |
| | Time | 0.32 | 9.88 | 1.5E03 | 0.24 | 0.50 | **2.15** |
| | RMSE | 9.9E−01 | 3.4E−01 | 2.7E−04 | 2.0E−02 | 2.3E−04 | **4.8E−04** |
| | SR | 0 | 0 | 0.96 | 0.94 | 0.94 | **1** |
| 499/150/20 | Iter | 13.18 | 200 | - | 20.84 | **121.32** | 104.78 |
| | Time | 0.57 | 10.60 | 1.7E03 | 0.27 | **1.74** | 5.07 |
| | RMSE | 1.0E+00 | 6.0E−01 | 7.1E−04 | 5.2E−01 | **3.8E−03** | 1.8E−01 |
| | SR | 0 | 0 | 0.76 | 0.08 | **0.78** | 0.58 |
| 999/300/30 | Iter | 12.56 | 200 | | **74.36** | 108.96 | **80.78** |
| | Time | 4.49 | 60.75 | | **1.44** | 2.78 | **32.63** |
| | RMSE | 1.0E+00 | 5.0E−01 | | **3.8E−02** | 1.3E−03 | **3.1E−02** |
| | SR | 0 | 0 | | **0.92** | 0.78 | **0.92** |
| 999/300/40 | Iter | 14.02 | 200 | | 14.66 | 160.84 | **130.42** |
| | Time | 7.29 | 61.39 | | 0.52 | 4.91 | **72.14** |
| | RMSE | 1.0E+00 | 6.0E−01 | | 5.3E−01 | 6.4E−03 | **1.2E−01** |
| | SR | 0 | 0 | | 0 | 0.5 | **0.68** |
| 1999/600/60 | Iter | 12.8 | 200 | | 102.88 | 153.14 | **111.44** |
| | Time | 28.82 | 606.69 | | 6.08 | 10.43 | **263.06** |
| | RMSE | 1.0E+00 | 5.1E−01 | | 9.2E−04 | 3.7E−03 | **8.4E−03** |
| | SR | 0 | 0 | | 0.94 | 0.54 | **0.96** |
| 1999/600/80 | Iter | 14.68 | 200 | | 2 | 189.82 | **182.46** |
| | Time | 46.17 | 607.84 | | 0.88 | 17.84 | **596.34** |
| | RMSE | 1.02E+00 | 6.13E−01 | | 5.62E−01 | 1.06E−02 | **2.05E−01** |
| | SR | 0 | 0 | | 0 | 0.18 | **0.28** |

**Table 2** continued

| n/m/r | | Cadzow | DRI | ANM | FIHT | PGD | pMAP |
|---|---|---|---|---|---|---|---|
| *m/n = 60%* | | | | | | | |
| 499/300/20 | Iter | 10.86 | 200 | – | **14.82** | 76.28 | **24.82** |
| | Time | 0.44 | 9.74 | 1.8E03 | **0.17** | 1.07 | **1.20** |
| | RMSE | 9.7E−01 | 1.7E−01 | 7.1E−05 | **8.8E−06** | 5.0E−04 | **3.9E−04** |
| | SR | 0 | 0 | 0.98 | **1** | 0.86 | **1** |
| 499/300/40 | Iter | 14.88 | 200 | - | **27.08** | 173.06 | **41.18** |
| | Time | 1.25 | 9.74 | 2.3E03 | **0.60** | 3.85 | **3.76** |
| | RMSE | 9.6E−01 | 3.5E−01 | 9.6E−04 | **2.6E−05** | 6.0E−03 | **5.1E−04** |
| | SR | 0 | 0 | 0.76 | **1** | 0.4 | **1** |
| 999/600/60 | Iter | 14.2 | 200 | | **20.8** | 178.6 | **40.64** |
| | Time | 12.09 | 61.60 | | **0.96** | 7.43 | **36.39** |
| | RMSE | 9.6E−01 | 2.6E−01 | | **1.8E−05** | 4.8E−03 | **5.6E−04** |
| | SR | 0 | 0 | | **1** | 0.32 | **1** |
| 999/600/80 | Iter | 16.12 | 200 | | **26.82** | 198.1 | **51.6** |
| | Time | 20.16 | 61.85 | | **1.77** | 10.74 | **66.79** |
| | RMSE | 9.45E−01 | 3.37E−01 | | **2.58E−05** | **9.41E−03** | **6.37E−04** |
| | SR | 0 | 0 | | **1** | 0.04 | **1** |
| 1999/1200/120 | Iter | 14.16 | 200 | | **21.38** | 198.7 | **58.34** |
| | Time | 73.92 | 597.26 | | **3.86** | 28.78 | **311.07** |
| | RMSE | 9.6E−01 | 2.6E−01 | | **1.7E−05** | 8.5E−03 | **7.6E−04** |
| | SR | 0 | 0 | | **1** | 0.06 | **1** |
| 1999/1200/160 | Iter | 16.22 | 200 | | **27.6** | 200 | **75** |
| | Time | 123.15 | 612.82 | | **7.50** | 39.52 | **577.77** |
| | RMSE | 9.5E−01 | 3.5E−01 | | **2.5E−05** | 1.1E−02 | **8.6E−04** |
| | SR | 0 | 0 | | **1** | 0.02 | **1** |

We further note the fact that in some cases FIHT stopped within a few iteration steps and this behaviour may lead to inferior solutions. Similar behaviours were also reported in another recent research (Fig. 3, [31]). Also with the increasing $r$, the performance of PGD declines when the ratio between $m$ and $r$ keeps fixed. It is because PGD has some assumptions on the lower bound of $m$ with respect to $r$ (Theorem 2.1, [2]), which may not hold in some cases. On the other hand, pMAP performs the best in 11 cases out of 12 in terms of the SR.

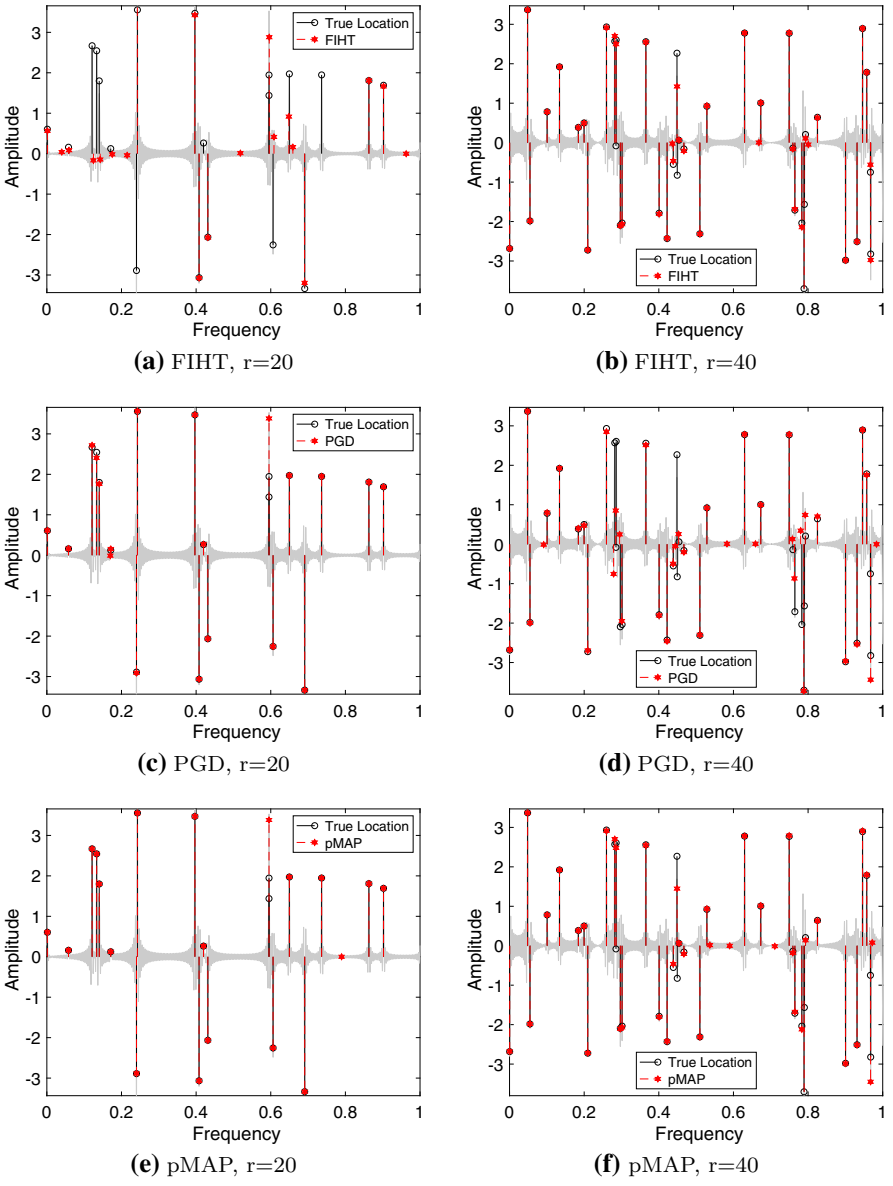
We find that FIHT is more computational efficient than pMAP. This is the result of using the subspace technique in FIHT (Alg. 2 in [3]), which is designed to approximate the matrix projections on to a low rank subspace and therefore reduces the computational time from $O(n^3)$ (SVD) to $O(nr^2)$. Although this technique can be adapted to our framework so as to reduce the computational cost, there would be no theoretical guarantee for its convergence results. We leave this potential improvement for future researches.

Some applications require to estimate the coefficients of a spectral sparse signal based on the reconstructed signal, including the amplitude $d_s$ and the frequency $\omega_s$ for all $s \in [1, 2, \ldots, r]$. We use the method in [6, Fig.2] to reconstruct the coefficients of recovered signals in this experiment. The reconstruction results are plotted in Fig. 3 over two randomly selected instances. It is easy to observe that when $r = 20$, FIHT incorrectly estimated most of coefficients with significant errors while both PGD and pMAP could successfully recover most of them. On the other hand when $r$ increases to 40, PGD performed worse than both FIHT and pMAP. There are several coefficients that were accurately estimated by FIHT and pMAP, but failed to be recovered by PGD (those unrecovered data points were shown by pointing arrows in Fig. 3d).

**Test b)** Table 3 lists the numerical results for this test including Iter, Time, RMSE and SR for five solvers. Due to the interference of noises, we increase the threshold of success rate to $10^{-2}$ to make sure the numerical results are comparable and meaningful.

Experimental results show that our proposed pMAP significantly outperforms the rest four candidate solvers in all tests. All the competitive solvers failed to recover any signals successfully in all cases (i.e., $SR = 0$ on average). We observed that those solvers were capable of generating points with RMSE at a level of $10^{-2}$, but encountered extreme difficulty in driving RMSE below $10^{-2}$. In contrast, the SR of pMAP is at least 0.92. It might be due to the fact that in DRI, FIHT and PGD, the weight of each observations can not be customised, which means these solvers have to give equal weights to both noisy observations and noiseless observations. As a result, their estimation results are significantly affected by noisy observations. This test clearly demonstrates the advantage of pMAP by allowing customized weighting schemes.

**Test (c)** The numerical results of recovering missing data with inaccurately estimated rank experiment are plotted in Fig. 4 for each solver. When $\hat{r}$ is smaller than 15, success rate for all solvers are zero. It indicates that none of these solvers can recover the incomplete signal successfully when there is a lack of coefficients information. With $\hat{r}$ exactly equals to 15, all three methods including FIHT, PGD and pMAP can achieve 100% recovery rate. However when $\hat{r} > 15$, one can expect varying performances of the three solvers. The success rates of both PGD and FIHT gradually decline with the increasing $\hat{r}$ and they finally reach around 40% when $\hat{r} = 30$. One




**(a)** FIHT, r=20 **(b)** FIHT, r=40

**(c)** PGD, r=20 **(d)** PGD, r=40

**(e)** pMAP, r=20 **(f)** pMAP, r=40

**Fig. 3** Spectral sparse signal coefficients reconstruction results by FIHT, PGD and pMAP, setting $\{n/m/r\} = \{499/150/20\}$ in (**a, c, e**) and $\{n/m/r\} = \{499/300/40\}$ in (**b, d, f**), respectively. Black circles stand for the true locations of coefficients while red stars stand for the estimated locations of coefficients

**Table 3** Numerical results for Cadzow, FIHT, PGD, DRI and our proposed pMAP on the noisy signal recovery experiment, including iterations (Iter), CPU time in seconds (Time), root of mean square error (RMSE) and success rate (SR). Results in this table are the average of 50 trials

| $n/r$ | | Cadzow | DRI | FIHT | PGD | pMAP |
|---|---|---|---|---|---|---|
| 499/5 | Iter | 7.92 | 200.00 | 7.64 | 24.18 | **20.86** |
| | Time | 0.16 | 8.62 | 0.05 | 0.19 | **0.80** |
| | RMSE | 9.86E−01 | 5.95E−02 | 2.63E−02 | 2.63E−02 | **2.19E−03** |
| | SR | 0 | 0 | 0 | 0 | **1** |
| 499/10 | Iter | 8.96 | 200.00 | 14.18 | 27.48 | **28.56** |
| | Time | 0.22 | 9.51 | 0.11 | 0.22 | **1.00** |
| | RMSE | 9.84E−01 | 9.58E−02 | 3.71E−02 | 3.70E−02 | **3.14E−03** |
| | SR | 0 | 0 | 0 | 0 | **1** |
| 499/20 | Iter | 11.34 | 200.00 | 28.52 | 77.74 | **43.64** |
| | Time | 0.58 | 9.32 | 0.31 | 1.12 | **1.93** |
| | RMSE | 9.73E−01 | 1.71E−01 | 5.70E−02 | 5.64E−02 | **6.76E−03** |
| | SR | 0 | 0 | 0 | 0 | **0.96** |
| 999/10 | Iter | 8.24 | 200.00 | 7.60 | 17.44 | **29.54** |
| | Time | 1.01 | 58.22 | 0.10 | 0.24 | **5.46** |
| | RMSE | 9.86E−01 | 6.16E−02 | 2.58E−02 | 2.58E−02 | **1.39E−03** |
| | SR | 0 | 0 | 0 | 0 | **1** |
| 999/20 | Iter | 9.18 | 200.00 | 16.96 | 53.62 | **40.08** |
| | Time | 2.12 | 58.09 | 0.26 | 1.03 | **10.26** |
| | RMSE | 9.84E−01 | 1.07E−01 | 3.84E−02 | 3.83E−02 | **2.80E−03** |
| | SR | 0 | 0 | 0 | 0 | **0.98** |
| 999/40 | Iter | 11.70 | 200.00 | 40.20 | 127.40 | **58.56** |
| | Time | 5.77 | 57.51 | 1.07 | 3.84 | **30.70** |
| | RMSE | 9.64E−01 | 1.78E−01 | 5.70E−02 | 5.64E−02 | **4.46E−03** |
| | SR | 0 | 0 | 0 | 0 | **0.96** |
| 1999/20 | Iter | 8.56 | 200.00 | 9.78 | 31.38 | **45.34** |
| | Time | 6.04 | 582.12 | 0.22 | 0.96 | **35.32** |
| | RMSE | 9.97E−01 | 6.70E−02 | 2.56E−02 | 2.56E−02 | **9.33E−04** |
| | SR | 0 | 0 | 0 | 0 | **1** |
| 1999/40 | Iter | 9.58 | 200.00 | 25.04 | 79.34 | **58.32** |
| | Time | 13.52 | 580.73 | 0.99 | 3.67 | **87.16** |
| | RMSE | 9.92E−01 | 1.04E−01 | 3.86E−02 | 3.85E−02 | **1.78E−03** |
| | SR | 0 | 0 | 0 | 0 | **0.98** |
| 1999/80 | Iter | 11.88 | 200.00 | 64.78 | 169.96 | **85.40** |
| | Time | 36.53 | 582.52 | 5.39 | 15.18 | **270.06** |
| | RMSE | 9.74E−01 | 1.82E−01 | 5.70E−02 | 5.62E−02 | **4.03E−03** |
| | SR | 0 | 0 | 0 | 0 | **0.92** |

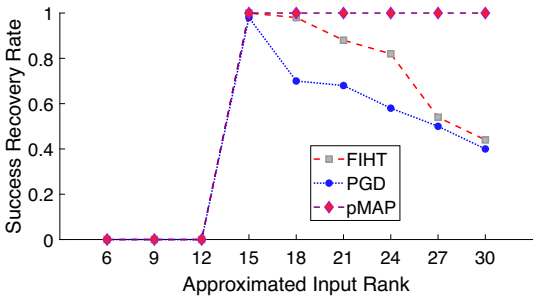
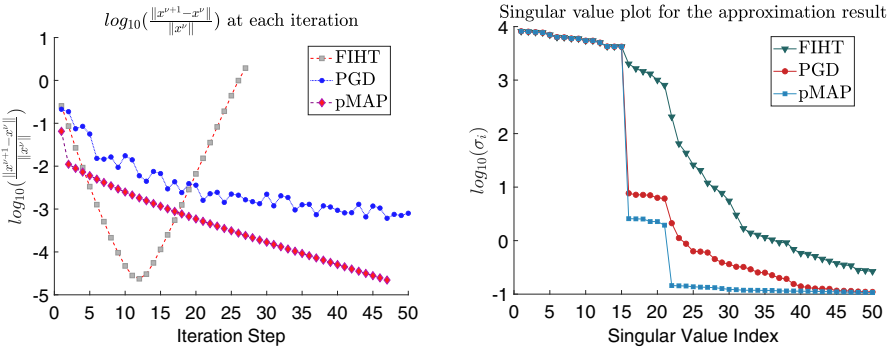



**Fig. 4** SR when the input rank is misappropriated for FIHT, PGD and pMAP. $n$ is set as 3999 and 30% observations are known. True rank $r$ is 15. Results in this figure are the average of 50 trials




**(a)** Plot of $log_{10}(\frac{\|x^{\nu+1}-x^{\nu}\|}{\|x^{\nu}\|})$ ar each iteration

**(b)** Solution singular values

**Fig. 5** Performance comparison for candidate solvers in incomplete signal recovering when the input rank is incorrectly estimated. True rank is 15 and input rank is 21. **a** Plots the relative gap between $x^{\nu}$ and $x^{\nu+1}$ for each solver at each iteration, while **b** plots the singular values of final solution by each solver. Both figures are with log base 10 scale

the other hand, SR of our proposed pMAP stays at 100% for any $\hat{r}$ no smaller than 15, which indicates that pMAP is more robust to the overestimation of objective rank than the other two solvers. One may wonder what has caused the failure of FIHT and PGD in this case. We explain it below.

Figure 5 compares the PGD, FIHT and pMAP using a random incomplete signal recovery example in terms of the relative error between the two consecutive iterates $\mathbf{x}^{\nu+1}$ and $\mathbf{x}^{\nu}$. Figure 5a shows that the failures of both FIHT and PGD in recovering incomplete signal was caused by the non-convergence behaviour. Figure 5b illustrates the distribution of singular values at the final iterates for each algorithm. A noticeable feature is that FIHT terminates too early because it cut off too many singular values that are not negligible, while PGD and pMAP cut off all comparably small eigenvalues. However, PGD suffers from non-monotonic convergence as shown in Fig. 5a. In contrast, the performance of pMAP just fits the situation studied in Theorem 3, which ensures locally linear convergence when the cut-off singular values are negligible comparing to the first 15 largest ones.

## 6 Conclusion

In this paper, we studied the problem of approximating a low rank Hankel matrix from the noisy and/or incomplete observation matrix under arbitrary weights. It is very challenging to tackle this problem due to its non-convexity. We introduced the framework of majorization minimization method such that this problem can be tackled iteratively with non-increasing objective function values. We further showed that the subproblem enjoys a closed form solution, which can be efficiently computed. We demonstrated the global optimal convergence property of our approach pMAP by assuming that the penalty parameter goes to infinity. We also showed that this method will at least converge to an $\epsilon$-approximate KKT point linearly if the penalty parameter $\rho$ is above a threshold. This method can be extended to tackle complex-valued matrices because the majorization $g_r(X)$ through the subdifferential of $h_r(X)$ holds in the complex-valued case. In the computational experiments for both series denoising and signal completion problems, pMAP usually outperforms other state-of-the-art solvers in terms of approximation accuracy within reasonable computing times.

One of the important topics to be further investigated is whether the computing cost of pMAP can be improved. For example, the subspace optimization technique used in [3] can reduce the computing time significantly compared with partial SVD, which is the major source of computing cost in pMAP. However, this technique at its current form is likely to break the established convergence result of pMAP, simply because it can not guarantee a closed form solution of each subproblem. Another interesting future research topic is extending our proposed pMAP framework to other rank-minimization related problems with similar structures, such as robust matrix completion and robust principal component analysis.

One referee suggested to consider the exact penalty using the distance function itself rather than its squared from, and use the majorization-minimization approach to the penalized problem. While we recognized the benefit of using the squared distance function as well surveyed in [18], we also think the suggested approach is worth serious investigation. The focus of the difficulty is on how to design an efficient majorization function for the distance function. We leave this to our next research topic.

**Acknowledgements** The authors would like to thank the AE and TE for their detailed comments on our coding and implementation. We are also grateful to the two anonymous referees for their constructive comments, which have helped to improve the quality of the paper. This research was partially supported by the National Natural Science Foundation of China (12011530155).

**Funding** The 2nd author's research is supported by the Ministry of Science and Technology, Taiwan (MOST 110-2115-M-003-003-MY2). This research was partially supported by the National Natural Science Foundation of China (12011530155).

**Data availability statement** This paper has associated data in a data repository. Matlab code and data used in the numerical experiments are available at https://doi.org/10.5281/zenodo.5807757.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Code availability** pMAP v1.0.0 is available under GNU General Public License. The URLs are contained in this published paper.



## References

1. Cadzow, J.A.: Signal enhancement: a composite property mapping algorithm. IEEE Trans. Acoust. Speech Signal Process. **36**, 49–62 (1988)
2. Cai, J.-F., Wang, T., Wei, K.: Spectral compressed sensing via projected gradient descent. SIAM J. Optim. **28**, 2625–2653 (2018)
3. Cai, J.-F., Wang, T., Wei, K.: Fast and provable algorithms for spectrally sparse signal reconstruction via low-rank Hankel matrix completion. Appl. Comput. Harmon. Anal. **46**, 94–121 (2019)
4. Chen, Y., Chi, Y.: Robust spectral compressed sensing via structured matrix completion. IEEE Trans. Inf. Theory **60**, 6576–6601 (2014)
5. Chu, M.T., Funderlic, R.E., Plemmons, R.J.: Structural low rank approximation. Linear Algebra Appl. **366**, 157–172 (2003)
6. Condat, L., Hirabayashi, A.: Cadzow denoising upgraded: a new projection method for the recovery of Dirac pulse from noisy linear measurements. Sample Theory Signal Image Process. **14**, 17–47 (2015)
7. Drusvyatskiy, D., Ioffe, A.D., Lewis, A.S.: Transversality and alternating projections for nonconvex sets. Found. Comput. Math. **15**, 1637–1651 (2015)
8. De Moor, B.: Total least squares for affinely structured matrices and the noisy realization problem. IEEE Trans. Signal Process. **42**, 3104–3113 (1994)
9. Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika **1**, 211–218 (1936)
10. Fazel, M., Pong, T.K., Sun, D.F., Tseng, P.: Hankel matrix rank minimization with applications to system identification and realization. SIAM J. Matrix Anal. Appl. **34**, 946–977 (2013)
11. Feppon, F., Lermusianux, P.J.: A geometric approach to dynamical model-order reduction. SIAM J. Matrix Anal. Appl. **39**, 510–538 (2018)
12. Gao, Y.: Structured low rank matrix optimization problems: a majorized penalty approach. Ph.D. thesis, National University of Singapore (2010)
13. Gao, Y., Sun, D.F.: A majorized penalty approach for calibrating rank constrained correlation matrix problems. Technical report, National University of Singapore (2010)
14. Gillard, J.: Cadzow's basic algorithm, alternating projections and singular spectrum analysis. Stat. Infer. **3**, 335–343 (2010)
15. Gillard, J., Usevich, K.: Structured low-rank matrix completion for forecasting in time series. Int. J. Forecast. **34**, 582–597 (2018)
16. Gillard, J., Zhigljavsky, A.: Weighted norms in subspace-based methods for time series analysis. Numer. Linear Algebra Appl. **23**, 947–967 (2016)
17. Golyandida, N., Nekrutkin, V., Zhigljavsky, A.: Analysis of Time Series Structure: SSA and Related Techniques. Chapman & Hall/CRC Press, Boca Raton (2001)
18. Keys, K.L., Zhou, H., Lange, K.: Proximal distance algorithms: theory and examples. J. Mach. Learn. Res. **20**, 1–38 (2019)
19. Kreutz-Delgado, K.: The complex gradient operator and the CR-calculus. University of California, San Diego, Version UCSD-ECE275CG-2009v1.0
20. Lai, M.-J., Varghese, A.: On convergence of the alternating projection method for matrix completion and sparse recovery problems. arXiv:1711.02151v1 (2017)
21. Li, Q., Qi, H.-D.: A sequential semismooth Newton method for the nearest low-rank correlation matrix problem. SIAM J. Optim. **21**, 1641–1666 (2011)

22. Liu, T., Lu, Z., Chen, X., Dai, Y.-H.: An exact penalty method for semidefinite-box constrained low-rank matrix optimization problems. IMA J. Numer. Anal. (2019) **(to appear)**
23. Markovsky, I.: Low Rank Approximation: Algorithms, Implementation, Applications. Springer, New York (2012)
24. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York (2000)
25. Qi, H.-D., Shen, J., Xiu, N.: A sequential majorization method for approximating weighted time series of finite rank. Stat. Interface **11**, 615–630 (2018)
26. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, New York (2004)
27. Shen, X., Mitchell, J.E.: A penalty method for rank minimization problems in symmetric matrices. Comput. Optim. Appl. **71**, 353–380 (2018)
28. Tang, G., Bhaskar, B.N., Shah, P., Recht, B.: Compressed sensing off the grid. IEEE Trans. Inf. Theory **59.11**, 7465–7490 (2013)
29. Usevich, K.: On signal and extraneous roots in singular spectrum analysis. Stat. Interface **3**, 281–295 (2010)
30. Wiringer, W.: Zurformalen theorie der funktionen von mehr complexen veränderlichen. Math. Ann. **97**, 357–375 (1927)
31. Ying, J., Cai, J.-F., Guo, D., Tang, G.: Vandermonde factorization of Hankel matrix for complex exponential signal recovery—application in fast NMR spectroscopy. IEEE Trans. Signal Process. **66**, 5520–5533 (2018)
32. Zhou, S., Xiu, N., Qi, H.-D.: A fast matrix majorization-projection method for penalized stress minimization with box constraints. IEEE Trans. Signal Process. **66**, 4331–4346 (2018)
33. Zhou, S., Xiu, N., Qi, H.-D.: Robust Euclidean embedding via EDM optimization. Math. Program. Comput. **12**, 337–387 (2020)