Multi-Channel Hankel Matrix Completion through Nonconvex Optimization

Shuai Zhang, Yingshuai Hao, Meng Wang, *Member, IEEE*, Joe H. Chow, *Fellow, IEEE*

***Abstract*—This paper studies the multi-channel missing data recovery problem when the measurements are generated by a dynamical system. A new model, termed multi-channel low-rank Hankel matrices, is proposed to characterize the intrinsic low-dimensional structures in multi-channel time series. The data recovery problem is formulated as a nonconvex optimization problem, and two fast algorithms (AM-FIHT and RAM-FIHT), both with linear convergence rates, are developed to recover the missing points with provable performance guarantees. The required number of observations is significantly reduced compared with conventional low-rank completion methods. Our methods are verified through numerical experiments on synthetic data and recorded synchrophasor data in power systems.**

***Index Terms*—low-rank matrix completion, nonconvex optimization, Hankel matrix, linear dynamic systems, synchrophasor data**

I. Introduction

Missing data recovery is an important task in various applications such as covariance estimation from partially observed correlations in remote sensing [10], multi-class learning in machine learning [3], [8], the Netflix Prize [1] problem and other similar questions in collaborative filtering [19]. Moreover, the recent framework of super-resolution enables accurate signal recovery from sparsely sampled measurements [9]. Example applications include magnetic resonance imaging (MRI) [21], [25], [43] and target localization in radar imaging [12], [47]. In power system monitoring, Phasor Measurement Units (PMU) [39] can measure voltage and current phasors directly at various locations and transmit the measurements to the operator for state estimation [2], [14] or disturbance identification [31]. Some PMU data points, however, do not reach the operator due to PMU malfunction or communication congestions. These missing data points should be recovered for the subsequent applications on PMU data [18].

Since practical datasets often have intrinsic low-dimensional structures, the missing data recovery problem can be formulated as a low-rank matrix completion problem, which is nonconvex due to the rank constraint. Its convex relaxation, termed Nuclear Norm Minimization (NNM) problem, has been extensively investigated [8], [11], [16], [20]. Given an $n_c \times n$ ($n_c \leq n$) matrix with rank r ($r \ll n$), as long as $O(rn \log^2 n)$[1] entries are observed, one can recover the remaining entries accurately by solving NNM [8], [11], [20].

Although elegant theoretical analyses exist, convex approaches like NNM have high computational complexity and poor convergence rate. For example, to decompose an $n_c \times n$ matrix, the per-iteration complexity of the best specialized implementation is $O(n_c^2 n)$ [36]. To reduce the computational complexity, first-order algorithms like [24] have been developed to solve the non-convex problem directly. Despite the numerical superiority, the theoretical analyses of the convergence and recovery performance of these nonconvex methods are still open problems. Only a few recent work such as [7], [24] provided such analyses on a case-by-base basis.

The low-rank matrix model, however, does not capture the temporal correlations in time series. A permutation of measurements at different time steps would result in different time series, but the rank of the data matrix remains the same. As a result, low-rank matrix completion methods require at least r entries in each column/row to recover the missing points and would fail if a complete column/row was lost. They cannot recover simultaneous data losses among all channels. Simultaneous data losses are not uncommon in power systems due to communication congestions.

There is limited study of the coupling of low-dimensional models and temporal correlations. Parametric models like hidden Markov models [33], [35] and autoregression (AR) models [22], [34] are employed to model temporal correlations. The accuracy of the algorithms depends on the correct estimation of model parameters, and no theoretical analysis is reported.

This paper develops a new model to characterize the intrinsic structures of multiple time series that are generated by a linear dynamical system. Our model of *multi-channel low-rank Hankel matrix* characterizes the temporal correlations in time series like PMU data without directly modeling the dynamical systems and estimating the system parameters. Our model can also be viewed as an extension of the single-channel low-rank Hankel matrix model with a $\Theta(r)$ degree of freedom in [7], [12] to an n_c-channel matrix with a $\Theta(n_c r)$ degree of freedom. It can also characterize spectrally sparse signals in applications like radar imaging [41] and magnetic resonance imaging [30].

Building upon the FIHT algorithm [7], this paper proposes two fast algorithms, termed accelerated multi-channel fast iterative hard thresholding (AM-FIHT) and robust AM-FIHT (RAM-FIHT) for multi-channel low-rank Hankel matrix completion. They can recover missing points for simultaneous data losses. The heavy ball method [29], [40] is employed to accelerate the convergence rate, and the acceleration is evaluated theoretically and numerically. Our algo-

The first two authors contributed equally to this paper. The authors are with the Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY, 12180. Email: {zhangs21, haoy2, wangm7, chowj}@rpi.edu. Phone: 518-276-3842. Fax: 518-276-6261.

Partial and preliminary results have appeared in [49].

[1] $f(n) = O(g(n))$ means that if for some constant $C > 0$, $f(n) \leq Cg(n)$ holds when n is sufficiently large. $f(n) = \Theta(g(n))$ means that for some constants $C_1 > 0$ and $C_2 > 0$, $C_1 g(n) \leq f(n) \leq C_2 g(n)$ holds when n is sufficiently large.

rithms converge linearly with a low per iteration complexity $O(r^2 n_c n + r n_c n \log n + r^3)$ to the original matrix (noiseless measurements) or a sufficiently close matrix depending on the noise level (noisy measurements). Theoretical analyses of FIHT with only noiseless measurements are reported [7]. Moreover, the recovery is successful as long as the number of observed measurements is $O(r^2 \log^2 n)$, significantly lower than $O(rn \log^2 n)$ for general rank-r matrices. This number is also a constant fraction of the required number of measurements by applying the single-channel Hankel matrix completion methods like FIHT [7] to each channel separately.

The rest of the paper is organized as follows. Sections II-III describe the problem formulation and the connection with the existing work. Sections IV-V present our algorithms and the theoretical analyses. Section VI records the numerical results on synthetic data and recorded PMU data. Section VII concludes the paper. All the proofs are summarized in Appendix.

Notation: Vectors are bold lowercase, matrices are bold uppercase, and scalars are in normal font. For example, $\boldsymbol{Z}$ is a matrix and $\boldsymbol{z}$ is vector. $\boldsymbol{Z}_{i*}$ denotes the ith row of $\boldsymbol{Z}$, and Z_{ij} denotes the (i,j)-th entry of $\boldsymbol{Z}$. $\boldsymbol{I}$ and $\boldsymbol{e}_i$ denote the identity matrix and the ith standard basis vector. $\boldsymbol{Z}^T$ and $\boldsymbol{Z}^*$ denote the transpose and conjugate transpose of $\boldsymbol{Z}$, so as $\boldsymbol{z}^T$ and $\boldsymbol{z}^*$. The inner product between two vectors is $\langle \boldsymbol{z}_1, \boldsymbol{z}_2 \rangle = \boldsymbol{z}_2^* \boldsymbol{z}_1$, and corresponding l_2 norm is $\|\boldsymbol{z}\| = \langle \boldsymbol{z}, \boldsymbol{z} \rangle^{1/2}$. For matrices, the inner product is defined as $\langle \boldsymbol{Z}_1, \boldsymbol{Z}_2 \rangle = \mathrm{Tr}(\boldsymbol{Z}_2^* \boldsymbol{Z}_1)$. $\|\boldsymbol{Z}\|_F$ stands for the Frobenius norm with $\|\boldsymbol{Z}\|_F = \langle \boldsymbol{Z}, \boldsymbol{Z} \rangle^{1/2}$. The spectral norm of matrix $\boldsymbol{Z}$ is denoted by $\|\boldsymbol{Z}\|$. The maximum entry (in absolute value) of $\boldsymbol{Z}$ is denoted as $\|\boldsymbol{Z}\|_\infty$. Linear operators on matrix spaces will be denoted by calligraphic letters. In particular, $\mathcal{I}$ is the identity operator. The spectral norm of a linear operator $\mathcal{A}$ on matrix spaces is denoted as $\|\mathcal{A}\| = \sup_{\langle \boldsymbol{Z}, \boldsymbol{Z} \rangle \le 1} \|\mathcal{A}\boldsymbol{Z}\|_F$. The adjoint operator of $\mathcal{A}$ is denoted as $\mathcal{A}^*$, which satisfies $\langle \mathcal{A}\boldsymbol{Z}_1, \boldsymbol{Z}_2 \rangle = \langle \boldsymbol{Z}_1, \mathcal{A}^* \boldsymbol{Z}_2 \rangle$.

II. Problem Formulation

Consider an n_p-th order linear dynamical system after an impulse response. Let $\boldsymbol{s}_t \in \mathbb{C}^{n_p}$ and $\boldsymbol{x}_t \in \mathbb{C}^{n_c}$ denote deviations of state variables and observations at time t from the equilibrium point. Then we have

$$\boldsymbol{s}_{t+1} = \boldsymbol{A}\boldsymbol{s}_t, \quad \boldsymbol{x}_t = \boldsymbol{C}\boldsymbol{s}_t, \quad t = 1, 2, \cdots, n, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{C}^{n_p \times n_p}$, and $\boldsymbol{C} \in \mathbb{C}^{n_c \times n_p}$. Let $\boldsymbol{X}$ contain the measurements from time 1 to n,

$$\boldsymbol{X} = [\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots,\ \boldsymbol{x}_n] \in \mathbb{C}^{n_c \times n}. \tag{2}$$

Further, the Hankel matrix of $\boldsymbol{X}$ is defined as

$$\mathcal{H}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_{n_2} \\ \boldsymbol{x}_2 & \boldsymbol{x}_3 & \cdots & \boldsymbol{x}_{n_2+1} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{n_1} & \boldsymbol{x}_{n_1+1} & \cdots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{C}^{n_c n_1 \times n_2}, \tag{3}$$

where $n_1 + n_2 = n + 1$.

Suppose $\boldsymbol{A}$ could be diagonalized, denoted by $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^{-1}$, where $\boldsymbol{P} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \cdots, \boldsymbol{l}_{n_p}]$, $\boldsymbol{P}^{-1} = [\boldsymbol{r}_1, \boldsymbol{r}_2, \cdots, \boldsymbol{r}_{n_p}]^*$, and $(\cdot)^*$ stands for the conjugate transpose. $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_{n_p})$ contains the eigenvalues of $\boldsymbol{A}$. Then

$$\begin{aligned} \boldsymbol{x}_{t+1} = \boldsymbol{C}\underbrace{\boldsymbol{A}\cdots\boldsymbol{A}}_{t \text{ times}}\boldsymbol{s}_1 = \boldsymbol{C}\boldsymbol{A}^t\boldsymbol{s}_1 = \boldsymbol{C}\boldsymbol{P}\boldsymbol{\Lambda}^t\boldsymbol{P}^{-1}\boldsymbol{s}_1 \\ = \sum_{i=1}^{n_p} \lambda_i^t \boldsymbol{r}_i^* \boldsymbol{s}_1 \boldsymbol{C}\boldsymbol{l}_i. \end{aligned} \tag{4}$$

All n_p modes of the system are considered in (4). In practice, a mode might be highly damped ($|\lambda_i| \approx 0$), or not excited by the input ($|\boldsymbol{r}_i^* \mathbf{s_1}| \approx 0$), or not directly measured ($\|\boldsymbol{C}\boldsymbol{l}_i\| \approx 0$). If only r ($r \ll n$) out of n modes are significant, assuming these modes to be λ_1,..., λ_r for simplicity, we have

$$\boldsymbol{x}_{t+1} \simeq \sum_{i=1}^{r} \lambda_i^t \boldsymbol{r}_i^* \boldsymbol{s}_1 \boldsymbol{C}\boldsymbol{l}_i. \tag{5}$$

Then the corresponding Hankel matrix can be written as

$$\mathcal{H}\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_{n_2} \\ \boldsymbol{x}_2 & \boldsymbol{x}_3 & \cdots & \boldsymbol{x}_{n_2+1} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_{n_1} & \boldsymbol{x}_{n_1+1} & \cdots & \boldsymbol{x}_n \end{bmatrix} = \boldsymbol{P}_L \boldsymbol{\Gamma} \boldsymbol{P}_R^T, \tag{6}$$

where

$$\boldsymbol{P}_L = \begin{bmatrix} \boldsymbol{I}_{n_c} & \boldsymbol{I}_{n_c} & \cdots & \boldsymbol{I}_{n_c} \\ \lambda_1 \boldsymbol{I}_{n_c} & \lambda_2 \boldsymbol{I}_{n_c} & \cdots & \lambda_r \boldsymbol{I}_{n_c} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{n_1-1} \boldsymbol{I}_{n_c} & \lambda_2^{n_1-1} \boldsymbol{I}_{n_c} & \cdots & \lambda_r^{n_1-1} \boldsymbol{I}_{n_c} \end{bmatrix} \in \mathbb{C}^{n_c n_1 \times n_c r}, \tag{7}$$

$$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{r}_1^* \boldsymbol{x}_1 \boldsymbol{C}\boldsymbol{l}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{r}_2^* \boldsymbol{x}_1 \boldsymbol{C}\boldsymbol{l}_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{r}_r^* \boldsymbol{x}_1 \boldsymbol{C}\boldsymbol{l}_r \end{bmatrix} \in \mathbb{C}^{n_c r \times r}, \tag{8}$$

and

$$\boldsymbol{P}_R = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_r \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{n_2-1} & \lambda_2^{n_2-1} & \cdots & \lambda_r^{n_2-1} \end{bmatrix} \in \mathbb{C}^{n_2 \times r}, \tag{9}$$

where $\boldsymbol{I}_{n_c} \in \mathbb{C}^{n_c \times n_c}$ is the identity matrix. One can check that both $\boldsymbol{X}$ and $\mathcal{H}\boldsymbol{X}$ are rank r matrices[2].

Let $\boldsymbol{N} \in \mathbb{C}^{n_c \times n}$ denote the measurement noise. $\boldsymbol{M} = \boldsymbol{X} + \boldsymbol{N}$ denotes the noisy measurements. Some entries of $\boldsymbol{M}$ are not observed due to data losses. Let $\widehat{\Omega}$ denote the index set of observed entries. The objective of missing data recovery is to reconstruct the missing data based on the observed entries $\mathcal{P}_{\widehat{\Omega}}(\boldsymbol{M})$. Since the rank of $\mathcal{H}\boldsymbol{X}$ is r, the data recovery problem can be formulated as

$$\min_{\boldsymbol{Z} \in \mathbb{C}^{n_c \times n}} \left\| \mathcal{P}_{\widehat{\Omega}}(\boldsymbol{Z} - \boldsymbol{M}) \right\|_F^2 \quad \text{subject to } \mathrm{rank}(\mathcal{H}\boldsymbol{Z}) = r, \tag{10}$$

where $\mathcal{P}_{\widehat{\Omega}}(\cdot)$ is the sampling operator with $(\mathcal{P}_{\widehat{\Omega}}(\boldsymbol{Z}))_{ij} = Z_{ij}$ if $(i,j) \in \widehat{\Omega}$ and 0 otherwise. (10) is a nonconvex problem due

[2]We assume $\mathcal{H}\boldsymbol{X}$ is exactly rank r throughout the paper. The methods analyses can be extended to approximately low-rank matrices with minor modifications. If $\mathcal{H}\boldsymbol{X}$ is approximately low-rank, i.e., its rank-r approximation error is very small, we seek to find the best rank-r approximation to $\mathcal{H}\boldsymbol{X}$. Then the recovery error is at least the approximation error.

to the rank constraint. It reduces to the conventional matrix completion problem when $n_1 = 1$.

Clearly, the recovery is impossible if $\boldsymbol{X}$ is in the null space of $\mathcal{P}_{\widehat{\Omega}}(\cdot)$. Here we follow the standard incoherence assumption in low-rank matrix completion [11].

Definition 1. *A matrix $\boldsymbol{Z} \in \mathbb{C}^{l_1 \times l_2}$ with singular value decomposition (SVD) as $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, is said to be incoherent with parameter μ if*

$$\max_{1\le k_1\le l_1} \left\|\boldsymbol{e}_{k_1}^*\boldsymbol{U}\right\|^2 \le \frac{\mu r}{l_1},\ \max_{1\le k_2\le l_2} \left\|\boldsymbol{e}_{k_2}^*\boldsymbol{V}\right\|^2 \le \frac{\mu r}{l_2}, \quad (11)$$

where $\boldsymbol{e}_{k_1}, \boldsymbol{e}_{k_2}$ are the coordinate unit vectors.

The incoherence definition guarantees that the singular vectors of the matrix are sufficiently spread, and $\mathcal{P}_{\widehat{\Omega}}(\cdot)$ samples enough information about the matrix. We focus on recovering μ-incoherence matrices in this paper.

III. Background and Related Work

The low-rank property of a Hankel matrix is also recently exploited in the direction of arrival (DOA) problem in array signal processing [12], [47], MRI image recovery from undersampled measurements [21], [37], [48], video inpainting [15] and system identification [17]. To see the connection with our model, the kth row of $\boldsymbol{X}$ in (2), denoted by $\boldsymbol{X}_{k*}$, can be equivalently viewed as the discrete samples of a spectrally sparse signal $g_k(t)$, which is a weighted sum of r damped or undamped sinusoids at $t = \{0, ..., n-1\}$, where

$$g_k(t) = \sum_{i=1}^{r} d_{k,i} e^{(2\pi\imath f_i - \tau_i)t}, k = 1, ..., n_c, \quad (12)$$

and f_i and $d_{k,i}$ are the frequency and the normalized complex amplitude of the ith sinusoid, respectively. $\imath$ is the imaginary unit. The connection between (12) and (2) is that $\lambda_i = e^{2\pi\imath f_i - \tau_i}$ and $d_{k,i} = \boldsymbol{r}_i^* \boldsymbol{s}_1 \boldsymbol{C}_{k*} \boldsymbol{l}_i$.

The signal of interest itself in array signal processing is spectrally sparse. In MRI imaging, if a signal reduces to a sparse linear combination of Dirac delta functions under some transformations, then its Fourier transform is a sum of a few sinusoids [25], [38], [48]. Most existing work on low-rank Hankel matrices studied single-channel signals, i.e., $n_c = 1$ in our setup. References [12], [26], [37], [38] considered 2-dimensional (2-D) and higher-dimensional signals, while a 2-D signal is still a sum of r 2-D sinusoids, and the degree of freedom is still $\Theta(r)$. The focus of this paper is multi-channel signals with $n_c > 1$. Each signal is a weighted sum of the same set of r sinusoids, while the weights $d_{k,i}$ are different for each channel $k = 1, ..., n_c$. The degree of the freedom of (12) is $\Theta(n_c r)$.

The multi-channel signal in (12) is related to the multiple measurement vector (MMV) problem [13]. References [27], [47] considered data recovery of MMV when the signals are linear combinations of undamped sinusoids, i.e., $\tau_i = 0$ for all i in (12). The data recovery is achieved in [27], [47] through atomic norm minimization, which requires solving large-scale semidefinite programs. Besides the high computational complexity, it is not clear how the atomic norm can be extended to handle damped sinusoids, i.e., $\tau_i \neq 0$. References [4], [25] studied multi-channel signal recovery using Hankel structures and can thus handle damped sinusoids. Despite the numerical evaluations, there is no theoretical analysis of the recovery guarantee in [4], [25]. This paper provides analytical recovery guarantees for multi-channel damped and undamped sinusoids.

The recovery of a low-rank Hankel matrix can be formulated as a convex optimization, for example, nuclear norm minimization for missing data recovery [15], [17], [25], [37], [45], [48] and minimizing a weighted sum of the nuclear norm and the ℓ_1 norm for bad data correction [26]. Since it is computationally challenging to solve these convex problems for high-dimensional Hankel matrices, fast algorithms to recover missing points in single-channel [7] and multi-channel Hankel matrices [5], [15] are proposed recently. Although numerical results are reported in [5], [15], only [7] provides the theoretical performance analysis of the proposed fast iterative hard thresholding (FIHT) algorithm for single-channel Hankel matrix recovery. FIHT is a projected gradient descent method. In each iteration, the algorithm updates the estimate along the gradient descent direction and then projects it to a rank-r matrix. To reduce the computational complexity, instead of solving singular value decomposition (SVD) directly, FIHT first projects a matrix onto a $2r$-dimensional subspace and then computes the SVD of the rank-$2r$ matrix. The per-iteration complexity of FIHT is $O(r^2 n + rn \log n + r^3)$.

Motivated by PMU data analysis in power systems, this paper connects dynamical systems with low-rank Hankel matrices. It develops fast data recovery algorithms for multi-channel Hankel matrices with provable performance guarantees.

IV. Data Recovery Algorithms

Here we describe two algorithms to solve (10) and defer the theoretical analyses to Section V. One is accelerated multi-channel fast iterative hard thresholding algorithm (AM-FIHT), and the other one is robust AM-FIHT (RAM-FIHT). Both algorithms are built upon the FIHT [7] with some major differences. First, FIHT recovers the missing points of one spectrally sparse signal, while (R)AM-FIHT recovers the missing points of n_c signals simultaneously. The simultaneous recovery can reduce the required number of measurements, as quantified in Theorem 5. Second, (R)AM-FIHT has a heavy-ball step [29], [40], e.g., term $\beta(\boldsymbol{W}_{l-1} - \boldsymbol{W}_{l-2})$ in line 5 of Algorithm 1 and line 14 of Algorithm 2, while FIHT does not. The basic idea of the heavy ball method is to compute the search direction using a linear combination of the gradient at the current iterate and the update direction in the previous step, rather than being memoryless of the past iterates' trajectory [29]. We will show analytically that with the heavy-ball step, AM-FIHT converges faster while maintaining the recovery accuracy (Theorem 2). Third, we provide the theoretical guarantee of data recovery when the measurements are noisy (Theorem 4), while [7] only has the performance guarantee of FIHT using noiseless measurements.

In both algorithms, $\boldsymbol{M}$, $\boldsymbol{X}_l$, $\boldsymbol{G}_l \in \mathbb{C}^{n_c \times n}$, and $\boldsymbol{W}_l$, $\Delta\boldsymbol{W}_l$, $\boldsymbol{L}_l \in \mathbb{C}^{n_c n_1 \times n_2}$. $\boldsymbol{L}_l$ is a rank-r matrix and its SVD is denoted as $\boldsymbol{L}_l = \boldsymbol{U}_l \boldsymbol{\Sigma}_l \boldsymbol{V}_l^*$, where $\boldsymbol{U}_l \in \mathbb{C}^{n_c n_1 \times r}$, $\boldsymbol{V}_l \in \mathbb{C}^{n_2 \times r}$ and $\boldsymbol{\Sigma_l} \in \mathbb{C}^{r \times r}$. $\mathcal{S}_l$ is the tangent subspace of the rank-r Rieman-

nian manifold at $\boldsymbol{L}_l$, and for any matrix $\boldsymbol{Z} \in \mathbb{C}^{n_c n_1 \times n_2}$, the projection of $\boldsymbol{Z}$ onto $\mathcal{S}_l$ is defined as

$$\mathcal{P}_{\mathcal{S}_l}(\boldsymbol{Z}) = \boldsymbol{U}_l\boldsymbol{U}_l^*\boldsymbol{Z} + \boldsymbol{Z}\boldsymbol{V}_l\boldsymbol{V}_l^* - \boldsymbol{U}_l\boldsymbol{U}_l^*\boldsymbol{Z}\boldsymbol{V}_l\boldsymbol{V}_l^*. \tag{13}$$

$\mathcal{Q}_r$ finds the best rank-r approximation as

$$\mathcal{Q}_r(\boldsymbol{Z}) = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*, \tag{14}$$

if $\boldsymbol{Z} = \sum_i \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^*$ is the SVD of $\boldsymbol{Z}$ with $\sigma_1 \geq \sigma_2 \geq \cdots$. $\mathcal{H}^\dagger$ is the Moore-Penrose pseudoinverse of $\mathcal{H}$. For any matrix $\boldsymbol{Z} \in \mathbb{C}^{n_c n_1 \times n_2}$, $(\mathcal{H}^\dagger \boldsymbol{Z}) \in \mathbb{C}^{n_c \times n}$ satisfies

$$\langle \mathcal{H}^\dagger \boldsymbol{Z}, \boldsymbol{e}_k \boldsymbol{e}_t^* \rangle = \frac{1}{w_t} \sum_{k_1+k_2=t+1} Z_{(k_1-1)n_c+k,k_2}, \tag{15}$$

where $w_t = \#\{(k_1,k_2) | k_1+k_2 = t+1, 1 \leq k_1 \leq n_1, 1 \leq k_2 \leq n_2\}$ as the number of elements in the t-th anti-diagonal of an $n_1 \times n_2$ matrix.

Algorithm 1 AM-FIHT for Data Recovery from Noiseless Measurements

Require: $\mathcal{P}_{\widehat{\Omega}}(\boldsymbol{M})$, n_1, n_2, r

1: Set $\boldsymbol{W}_{-2} = \boldsymbol{0}$, $\boldsymbol{W}_{-1} = p^{-1}\mathcal{H}\mathcal{P}_{\widehat{\Omega}}(\boldsymbol{M})$, $\boldsymbol{L}_0 = \mathcal{Q}_r(\boldsymbol{W}_{-1})$;
2: Initialize $\boldsymbol{X}_0 = \mathcal{H}^\dagger \boldsymbol{L}_0$;
3: **for** $l = 0, 1, \cdots$ **do**
4: $\quad \boldsymbol{G}_l = \mathcal{P}_{\widehat{\Omega}}(\boldsymbol{M} - \boldsymbol{X}_l)$;
5: $\quad \boldsymbol{W}_l = \mathcal{P}_{\mathcal{S}_l}\left(\mathcal{H}(\boldsymbol{X}_l + p^{-1}\boldsymbol{G}_l) + \beta(\boldsymbol{W}_{l-1} - \boldsymbol{W}_{l-2})\right)$;
6: $\quad \boldsymbol{L}_{l+1} = \mathcal{Q}_r(\boldsymbol{W}_l)$;
7: $\quad \boldsymbol{X}_{l+1} = \mathcal{H}^\dagger \boldsymbol{L}_{l+1}$;
8: **end for**
9: **return** $\boldsymbol{X}_l$

The key steps in AM-FIHT are as follows. Here the measurements are noiseless, thus $\boldsymbol{M} = \boldsymbol{X}$. In each iteration, we first update current $\boldsymbol{X}_l$ along the gradient descent direction $\boldsymbol{G}_l$, with a step size $p^{-1} = \frac{n_c n}{m}$, where m is the number of observed entries. To improve the convergence rate, the update is further combined with additional heavy-ball term $\beta(\boldsymbol{W}_{l-1} - \boldsymbol{W}_{l-2})$, which represents the update direction in the previous iteration. Next, $\mathcal{H}(\boldsymbol{X}_l + p^{-1}\boldsymbol{G}_l) + \beta(\boldsymbol{W}_{l-1} - \boldsymbol{W}_{l-2})$ is projected to a rank-r matrix. To reduce the computational complexity, we first project it to the $2r$-dimensional space $\mathcal{S}_l$ and then apply SVD on the rank-$2r$ matrix [7], instead of directly computing its SVD. The rank-r matrix $\boldsymbol{L}_{l+1}$ is obtained in line 7 by thresholding the singular values of the rank-$2r$ matrix $\boldsymbol{W}_l$. Finally, $\boldsymbol{X}_{l+1}$ is updated by $\mathcal{H}^\dagger \boldsymbol{L}_{l+1}$.

The analysis of the computational complexity of AM-FIHT is similar to that of FIHT [7] with some modifications for the n_c-channel signal and the heavy-ball step. Details are in the supplementary materials. The computational complexity of solving SVD of a matrix in $\mathbb{C}^{n_c n_1 \times n_2}$ is generally $O(n_c n^2 r)$. Due to the low rank structure of the matrices in $\mathcal{S}_l$, the SVD of $\boldsymbol{W}_l \in \mathbb{C}^{n_c n_1 \times n_2}$ can be computed in $O(n_c n r^2 + r^3)$ via QR decompositions and SVD on a $2r \times 2r$ matrix [7]. Moreover, it is not necessary to construct Hankel matrices following (3) explicitly. The matrix multiplication of $\boldsymbol{U}_l^* \mathcal{H} \boldsymbol{X}_l \in \mathbb{C}^{r \times n_2}$ and $(\mathcal{H}\boldsymbol{X}_l)\boldsymbol{V}_l \in \mathbb{C}^{n_c n_1 \times r}$ in line 5 can be completed via fast convolution algorithms with $O(n_c n r \log(n))$ flops, instead of the conventional complexity of $O(n_c n^2 r)$. Similar analysis can be applied to line 7, which costs $O(n_c n r \log(n))$ flops to compute $\boldsymbol{X}_{l+1}$ from the SVD of $\boldsymbol{L}_{l+1}$ directly.

With the heavy ball term, since the SVDs of $\boldsymbol{W}_{l-1}$ and $\boldsymbol{W}_{l-2}$ have been obtained in the last two steps, we compute $\mathcal{P}_{\mathcal{S}_l}(\boldsymbol{W}_{l-1}) - \mathcal{P}_{\mathcal{S}_l}(\boldsymbol{W}_{l-2})$ in line 5. From (13), the computation of $\boldsymbol{U}_l\boldsymbol{U}_l^*\boldsymbol{Z}\boldsymbol{V}_l\boldsymbol{V}_l^*$ plays the dominant part in computing $\mathcal{P}_{\mathcal{S}_l}(\boldsymbol{Z})$. Let $\boldsymbol{W}_l = \boldsymbol{U}_{\boldsymbol{W}_l}\boldsymbol{\Sigma}_{\boldsymbol{W}_l}\boldsymbol{V}_{\boldsymbol{W}_l}^*$ denote the SVD of $\boldsymbol{W}_l$, where $\boldsymbol{U}_{\boldsymbol{W}_l} \in \mathbb{C}^{n_c n_1 \times 2r}, \boldsymbol{V}_{\boldsymbol{W}_l} \in \mathbb{C}^{n_2 \times 2r}$. Then computing $\boldsymbol{U}_l^*\boldsymbol{U}_{\boldsymbol{W}_{l-1}}$ and $\boldsymbol{V}_{\boldsymbol{W}_{l-1}}^*\boldsymbol{V}_l$ requires $O(n_c n r^2)$ and $O(n r^2)$ flops, respectively. Computing $\boldsymbol{U}_l^*\boldsymbol{U}_{\boldsymbol{W}_{l-1}}\boldsymbol{\Sigma}_{\boldsymbol{W}_{l-1}}\boldsymbol{V}_{\boldsymbol{W}_{l-1}}^*\boldsymbol{V}_l$ further requires $O(r^3)$ flops.

From the above analysis, line 4 requires $O(n_c n)$ flops. The complexity of line 5 is $O(n_c n r \log(n) + n_c n r^2 + r^3)$. Line 6 requires $O(n_c n r^2 + r^3)$ flops, and line 7 requires $O(n_c n r \log(n))$ flops. Thus, the total per-iteration complexity of AM-FIHT is $O(r^2 n_c n + r n_c n \log n + r^3)$.

Algorithm 2 RAM-FIHT

Require: $\mathcal{P}_{\widehat{\Omega}}(\boldsymbol{M})$, n_1, n_2, r, μ, and β.

1: Partition $\widehat{\Omega}$ into $L+1$ disjoint sets $\widehat{\Omega}_0, \widehat{\Omega}_1, \cdots, \widehat{\Omega}_L$ of equal size $\widehat{m}$, let $\widehat{p} = \frac{\widehat{m}}{n}$.
2: Set $\boldsymbol{W}_{-2} = \boldsymbol{0}$, $\boldsymbol{W}_{-1} = \widehat{p}^{-1}\mathcal{H}\mathcal{P}_{\widehat{\Omega}_0}(\boldsymbol{M})$, $\boldsymbol{L}_0 = \mathcal{Q}_r(\boldsymbol{W}_{-1})$;
3: **for** $l = 0, 1, \cdots, L-1$ **do**;
4: $\quad [\boldsymbol{U}_l, \boldsymbol{\Sigma}_l, \boldsymbol{V}_l]$ =SVD$(\boldsymbol{L}_l)$;
5: $\quad$ **for** $i = 1, 2, \cdots, n_c n_1$ **do**
6: $\quad\quad (\boldsymbol{A}_l)_{i*} = \frac{(\boldsymbol{U}_l)_{i*}}{\|(\boldsymbol{U}_l)_{i*}\|} \min\left\{ \|(\boldsymbol{U}_l)_{i*}\|, \sqrt{\frac{\mu r}{n_c n_1}} \right\}$;
7: $\quad$ **end for**
8: $\quad$ **for** $i = 1, 2, \cdots, n_2$ **do**
9: $\quad\quad (\boldsymbol{B}_l)_{i*} = \frac{(\boldsymbol{V}_l)_{i*}}{\|(\boldsymbol{V}_l)_{i*}\|} \min\left\{ \|(\boldsymbol{V}_l)_{i*}\|, \sqrt{\frac{\mu r}{n_2}} \right\}$;
10: $\quad$ **end for**
11: $\quad \boldsymbol{L}_l' = \boldsymbol{A}_l\boldsymbol{\Sigma}_l\boldsymbol{B}_l^*$;
12: $\quad \widehat{\boldsymbol{X}_l} = \mathcal{H}^\dagger \boldsymbol{L}_l'$;
13: $\quad \boldsymbol{G}_l = \mathcal{P}_{\widehat{\Omega}_{l+1}}(\boldsymbol{M} - \widehat{\boldsymbol{X}_l})$;
14: $\quad \boldsymbol{W}_l = \mathcal{P}_{\mathcal{S}_l'}(\mathcal{H}(\widehat{\boldsymbol{X}}_l + \widehat{p}^{-1}\boldsymbol{G}_l) + \beta(\boldsymbol{W}_{l-1} - \boldsymbol{W}_{l-2}))$;
15: $\quad \boldsymbol{L}_{l+1} = \mathcal{Q}_r(\boldsymbol{W}_l)$;
16: **end for**
17: **return** $\boldsymbol{X}_L = \mathcal{H}^\dagger \boldsymbol{L}_L$;

RAM-FIHT differs from AM-FIHT mainly in resampling (line 1) and trimming (lines 5-10). The resampling and trimming are used in [7] to improve the initialization of FIHT. Here we apply these ideas in the data recovery algorithm and prove in Theorem 4 that the resulting RAM-FIHT can recover the matrix even when the observed measurements are noisy. There is no analytical analysis of FIHT on noisy measurements in [7]. Moreover, compared with AM-FIHT, we provide tighter bounds of the required number of observations for RAM-FIHT (comparing Theorems 1 and 3).

In RAM-FIHT, the sampling set $\widehat{\Omega}$ is divided into L disjoint subsets $\widehat{\Omega}_l$'s. During the l-th iteration, $\boldsymbol{L}_l$ is updated using the observed entries in $\widehat{\Omega}_l$, instead of using all the entries in $\widehat{\Omega}$ as in AM-FIHT. The partition of the sampling set is a standard technique in analyzing matrix completion (MC) problems [42]. The disjointness of $\boldsymbol{L}_l$'s in different iterations simplifies

the theoretical analyses[3], since it ensures the independence between $\boldsymbol{X}_l$ and $\boldsymbol{X}_{l+1}$. The trimming procedure ensures that the estimate in each iteration remains close to μ-incoherent, which in turn helps to obtain tighter bounds of the recovery performance in Theorem 3. We remark that the resampling and trimming steps in RAM-FIHT are introduced mainly to simplify the theoretical analyses and obtain tighter bounds, while we observe numerically that AM-FIHT and RAM-FIHT perform similarly in Section VI. The per iteration computational complexity of RAM-FIHT is $O(r^2 n_c n + r n_c n \log n + r^3)$.

V. Theoretical Analyses

The theoretical analyses of the convergence rates and recovery accuracy of AM-FIHT and RAM-FIHT are summarized in the following four theorems. All the proofs are deferred to the Appendix. Theorem 1 records the recovery performance of AM-FIHT using noiseless measurements with $\beta = 0$. Theorem 2 shows that the convergence rate of AM-FIHT can be further improved by using a small positive β. Theorems 3 and 4 discuss the recovery performance of RAM-FIHT from noiseless and noisy measurements, respectively. We also compare the recovery performance with recovering missing points on each individual row of $\boldsymbol{X}$ separately and quantify the performance gain of our algorithms in Theorem 5.

Theorem 1. *(AM-FIHT with noiseless measurements.) Assume $\mathcal{H}\boldsymbol{X}$ is μ-incoherent. Let $0 < \varepsilon_0 < \frac{1}{10}$ be a numerical constant and $\nu = 6\varepsilon_0 < 1$. Then with probability at least $1 - 3n_c n^{-2}$, the iterates $\boldsymbol{X}_l$'s generated by AM-FIHT with $\beta = 0$ satisfy*

$$\|\boldsymbol{X}_l - \boldsymbol{X}\|_F \leq \nu^{l-1} \|\boldsymbol{L}_0 - \mathcal{H}\boldsymbol{X}\|_F, \tag{16}$$

provided that

$$m \geq C_1 \max\left\{ \frac{\mu c_s r \log(n)}{\varepsilon_0^2}, \frac{1+\varepsilon_0}{\varepsilon_0} (n_c \mu c_s n)^{\frac{1}{2}} \kappa r \log^{\frac{3}{2}}(n) \right\} \tag{17}$$

for some constant $C_1 > 0$, where $\kappa = \frac{\sigma_{\max}(\mathcal{H}\boldsymbol{X})}{\sigma_{\min}(\mathcal{H}\boldsymbol{X})}$ denotes the condition number of $\mathcal{H}\boldsymbol{X}$ and $c_s = \max\{\frac{n}{n_1}, \frac{n}{n_2}\}$.

Theorem 1 indicates that if the number of noiseless observations is $O(r n_c^{1/2} n^{1/2} \log^{3/2}(n))$, then AM-FIHT is guaranteed to recover $\boldsymbol{X}$ exactly. Moreover, from (16), the iterates generated by AM-FIHT converge linearly to the groundtruth $\boldsymbol{X}$, and the rate of convergence is ν. Since $\boldsymbol{X}$ is rank r, if one directly applies a conventional low-rank matrix completion method such as NNM ([8], [11], [20]), the required number of observations is $O(rn \log^2(n))$. Thus, when n is large, by exploiting the low-rank Hankel structure of correlated time series, the required number of measurements is significantly reduced. Note that the degree of freedom of $\boldsymbol{X}$ is $\Theta(n_c r)$, as one can see from (12), the required number of observations by Theorem 1 is suboptimal due to the dependence upon n. This results from the artefacts in our proof techniques. We will provide a tighter bound for RAM-FIHT in Theorem 3.

The required number of measurements depends on c_s, which is minimized when $n_1 = n_2 = \frac{n+1}{2}$. In practice, the selection of n_1 and n_2 of the Hankel matrix is also affected by the accuracy of the low-rank approximation.

[3]In fact, we only need the mutual independence among the subsets $\widehat{\Omega}_l$'s, and the disjoint partition is a sufficient condition.

We set β as 0 in Theorem 1 to simplify the analyses. The improvement of the convergence rate by using a positive β is quantified in the following theorem.

Theorem 2. *(Faster convergence with a heavy-ball step) Given any $\beta \in [0, \tau)$ for some $\tau > 0$, let $\boldsymbol{X}_l$'s denote the convergent iterates returned by AM-FIHT. There exists an integer s_0, a constant $q \in (0,1)$ that depends on β such that*

$$\|\boldsymbol{X}_{s_0+k} - \boldsymbol{X}\|_F \leq c(\delta)(q(\beta) + \delta)^k, \ \forall k \geq 0 \tag{18}$$

holds for any $\delta \in (0, 1-q(\beta))$ and a positive $c(\delta)$ that depends on δ. Moreover,

$$q(0) > q(\beta), \ \forall \beta \in (0, \tau). \tag{19}$$

The exact expressions of q and τ are deferred to the proofs in Appendix (equation (53)). Theorem 2 indicates that by adding a heavy-ball term, when close enough to the groundtruth $\boldsymbol{X}$, the iterates converge linearly to $\boldsymbol{X}$, and the rate of convergence is $q(\beta) + \delta$. Moreover, from (19), with a small positive β, the iterates converge faster than those without the heavy-ball step. Such improvement is numerically evaluated in Section VI.

Theorem 3. *(RAM-FIHT with noiseless measurements) Assume $\mathcal{H}\boldsymbol{X}$ is μ-incoherent. Let $0 < \varepsilon_0 < \frac{1}{2}$ and*

$$L = \left\lceil \varepsilon_0^{-1} \log\left(\frac{\sigma_{\max}(\mathcal{H}\boldsymbol{X})}{128\kappa^3 \varepsilon} \right) \right\rceil. \tag{20}$$

Define $\nu = 2\varepsilon_0 < 1$. Then with probability at least $1 - (2L+3)n_c n^{-2}$, for any arbitrarily small constant $\varepsilon > 0$, the iterates $\boldsymbol{L}_l$'s and $\boldsymbol{X}_L$ generated by RAM-FIHT with $\beta = 0$ satisfy

$$\begin{aligned} &\|\boldsymbol{L}_l - \boldsymbol{X}\|_F \leq \nu^l \|\boldsymbol{L}_0 - \mathcal{H}\boldsymbol{X}\|_F, 1 \leq l \leq L, \\ \text{and } &\|\boldsymbol{X}_L - \boldsymbol{X}\|_F \leq \nu^L \|\boldsymbol{L}_0 - \mathcal{H}\boldsymbol{X}\|_F \leq \varepsilon, \end{aligned}$$

provided that

$$m \geq C_2 \varepsilon_0^{-3} \mu c_s \kappa^6 r^2 \log(n) \log\left(\frac{\sigma_{\max}(\mathcal{H}\boldsymbol{X})}{\kappa^3 \varepsilon} \right) \tag{21}$$

for some constant $C_2 > 0$.

Theorem 3 shows that the iterates of RAM-FIHT converge to the groundtruth $\boldsymbol{X}$ with a linear convergence rate, and the number of required measurements is further reduced from that needed by AM-FIHT. To see this, note that $\sigma_{\max}(\mathcal{H}\boldsymbol{X}) \leq \sqrt{n_c n} \|\boldsymbol{X}\|_\infty$. If $\|\boldsymbol{X}\|_\infty$ is a constant, and select $\varepsilon = O(n^{-\alpha})$ with a positive constant α, we have $L = O(\log(n))$ from (20) and $m \geq O(r^2 \log^2 n)$ from (21). Compared with the bound of $O(r n_c^{1/2} n^{1/2} \log^{3/2}(n))$ in Theorem 1, the dependence on n is significantly reduced to $\log^2 n$, while the dependence on r is worse, from r to r^2. Since r is usually very small, and n is much larger, $O(r^2 \log^2 n)$ by Theorem 3 is tighter than $O(r n_c^{1/2} n^{1/2} \log^{3/2}(n))$ by Theorem 1. Since the degree of freedom of $\Theta(n_c r)$, we suspect that the bound could be improved further using better proof techniques than ours.

Theorem 4. *(RAM-FIHT with noisy measurements) Assume $\mathcal{H}\boldsymbol{X}$ is μ-incoherent and*

$$\|\boldsymbol{N}\|_\infty \leq \frac{\varepsilon_0 \|\mathcal{H}\boldsymbol{X}\|}{2048\kappa^3 r^{1/2} n_c^{1/2} n}. \tag{22}$$

Let $L = \left\lceil \varepsilon_0^{-1} \log\left(\frac{\sigma_{\max}(\mathcal{H}\boldsymbol{X})}{128\kappa^3\varepsilon}\right)\right\rceil$ *and* $0 < \varepsilon_0 < \frac{1}{4}$. *Define* $\nu = 2\varepsilon_0 < \frac{1}{2}$. *Then with probability at least* $1-(3L+3)n_c n^{-2}$ *and for any arbitrarily small constant* $\varepsilon > 0$, *the iterates* $\boldsymbol{L}_l$*'s* $(l = 1, ..., L)$ *generated by RAM-FIHT with* $\beta = 0$ *satisfies*

$$\begin{aligned}\|\boldsymbol{L}_l - \mathcal{H}\boldsymbol{X}\|_F \leq & \nu^l \|\boldsymbol{L}_0 - \mathcal{H}\boldsymbol{X}\|_F \\ & + 128 n_c^{1/2} n \|\boldsymbol{N}\|_\infty + 8 r^{1/2} \|\mathcal{H}\boldsymbol{N}\|,\end{aligned}$$

and

$$\|\boldsymbol{L}_L - \mathcal{H}\boldsymbol{X}\|_F \leq \varepsilon + 128 n_c^{1/2} n \|\boldsymbol{N}\|_\infty + 8 r^{1/2} \|\mathcal{H}\boldsymbol{N}\|, \tag{23}$$

provided that

$$m \geq C_3 \varepsilon_0^{-3} \mu c_s \kappa^6 r^3 \log(n) \log\left(\frac{\sigma_{\max}(\mathcal{H}\boldsymbol{X})}{\kappa^3 \varepsilon}\right) \tag{24}$$

for some constant $C_3 > 0$.

Theorem 4 explores the performance of RAM-FIHT in the noisy case. Note that $\|\mathcal{H}\boldsymbol{X}\|_\infty = \|\boldsymbol{X}\|_\infty$, and

$$\frac{n_c^{1/2} n}{\mu c_s r} \|\mathcal{H}\boldsymbol{X}\|_\infty \leq \|\mathcal{H}\boldsymbol{X}\| \leq n_c^{1/2} n \|\mathcal{H}\boldsymbol{X}\|_\infty .$$

If μ and r are both constants, (22) implies that $\|\boldsymbol{N}\|_\infty$ can be as large as a constant fraction of $\|\boldsymbol{X}\|_\infty$. When the number of observations is at least $O(r^3 \log^2(n))$, the error between the ground truth and the iterates returned by RAM-FIHT is controlled by the noise level. To evaluate the optimality of this error bound, consider a special case that $\boldsymbol{X}$ is a constant matrix with each entry being c, and $\boldsymbol{N}$ is a constant matrix with each entry being $-c$. Then every observation is zero, and the estimated matrix from partial observations by any recovery method would be a zero matrix. Then the recovery error is $\|\mathcal{H}\boldsymbol{N}\|_F = \sqrt{n_c n_1 n_2}|c| = \sqrt{n_c n_1 n_2}\|\boldsymbol{N}\|_\infty$. The sum of the second and the third term in the right hand side of (23) is bounded by $(128 c_s + 8 r^{1/2})\sqrt{n_c n_1 n_2}\|\boldsymbol{N}\|_\infty$. Thus, the error bound of RAM-FIHT is in the same order of the minimum error by any method.

Comparison with single-channel missing data recovery. FIHT [7] is a single-channel Hankel matrix completion method. When $n_c = 1$, Theorems 1 and 3 reduce to the results in [7]. One can apply FIHT to recover the missing points of each row of $\boldsymbol{X}$ and solve n_c data recovery problems separately. Let $\mathcal{H}\boldsymbol{X}_{k*}$ denote the single-channel Hankel matrix constructed from the kth row of $\boldsymbol{X}$. Suppose $\mathcal{H}\boldsymbol{X}_{k*}$ is μ_0-incoherent for every $1 \leq k \leq n_c$. Then setting $n_c = 1$ in Theorems 1 and 3 (or using Theorems 1 and 2 in [7]), we know that if each $\mathcal{H}\boldsymbol{X}_{k*}$ is recovered separately, the required number of measurements is proportional to $\sqrt{\mu_0}$ (AM-FIHT) or μ_0 (RAM-FIHT). Then the total number of observations to recover $\boldsymbol{X}$ is proportional to $n_c\sqrt{\mu_0}$ or $n_c\mu_0$. In contrast, the required number of observations by our methods is proportional to $\sqrt{n_c\mu}$ (AM-FIHT) or μ (RAM-FIHT). Thus, the ratio of the number of measurements by our method to FIHT is $\sqrt{\frac{\mu}{n_c\mu_0}}$ (or $\frac{\mu}{n_c\mu_0}$). We can show that our method only requires a constant fraction of the measurements by using FIHT through the following theorem.

Theorem 5.

$$\frac{\mu}{n_c\mu_0} < 1. \tag{25}$$

If it further holds that $(1-\delta)|\widehat{d}| \leq |d_{k,i}| \leq (1+\delta)|\widehat{d}|, \forall k \in \{1, ..., n_c\}, i \in \{1, ..., r\}$ *for some* $\delta \in (0, 1)$ *and* $\widehat{d} \in \mathbb{C}$, *where* $d_{k,i} = \boldsymbol{r}_i^* \boldsymbol{s}_1 \boldsymbol{C}_{k*} \boldsymbol{l}_i$, *we have*

$$\frac{\mu}{n_c\mu_0} \leq \frac{1}{1 + (n_c - 1)\frac{(1-\delta)^2}{\kappa_L^2(1+\delta)^2}}, \tag{26}$$

where κ_L *is the conditional number of* $\boldsymbol{P}_L$ *when* $n_c = 1$.

Theorem 5 indicates that the required number of measurements is reduced when collectively processing $\boldsymbol{X}$. Note that μ_0 is independent of the amplitude parameters $d_{k,i}$'s and depends only on the separations of the frequencies f_i's in (12). As a direct corollary of Theorem 2 in [28], if the separation among frequencies f_i's is at least $1/c_s n$, then μ_0 is a constant. In contrast, μ depends on both $d_{k,i}$'s and f_i's. (25) shows that μ is always less than $n_c\mu_0$. Moreover, in the special case that $d_{k,i}$'s are all in a small range, $\mu/(n_c\mu_0)$ can be reduced to approximately κ_L^2/n_c from (26). With well separated frequencies, the maximum and minimum singular values of $\boldsymbol{P}_L$ when $n_c = 1$ are both proportional to $\sqrt{n_1}$ [28]. That implies κ_L is a constant. Then, κ_L^2/n_c is in the order of $1/n_c$ for large n_c, and we have $\mu/\mu_0 = O(1/n_c)$. Combining these results with the arguments before Theorem 3, one can see that the required number of measurements is significantly reduced by collective processing.

VI. Numerical Results

We test the numerical performance of AM-FIHT and RAM-FIHT. The simulations are implemented in MATLAB on a desktop with 3.4 GHz Intel Core i7 and 16 GB memory. In all the experiments, we delete some data points in the datasets and test the recovery performance. We consider three modes of missing data patterns, as illustrated in Fig. 1. Given a data loss percentage,

- Mode 1: Data losses occur randomly and independently across time and channels.
- Mode 2: At randomly selected time instants, the data points in all channels are lost simultaneously.
- Mode 3: Starting from a randomly selected time instant, in half of the channels that are randomly selected, the data points are lost simultaneously and consecutively lost.

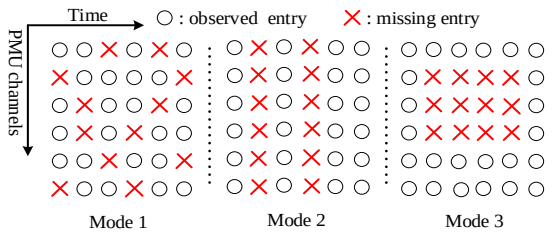

Fig. 1: Three modes of missing data

A. Numerical experiments on synthetic data

We test the recovery performance on synthetic spectrally sparse signals. Each row of matrix $\boldsymbol{X} \in \mathbb{C}^{n_c \times n}$ is a weighted sum of r sinusoids as shown in (12). Each f_i is randomly selected from $(0, 1)$. τ_i is 0 for all i. The complex coefficient $d_{k,i}$ has its angle randomly selected from $(0, 2\pi)$ and its magnitude chosen as $1 + 10^{a_{k,i}}$, where $a_{k,i}$ is randomly selected from $(0, 1)$.

1) (R)AM-FIHT with noiseless measurements: We first compare the performance of AM-FIHT and RAM-FIHT with noiseless measurements. For RAM-FIHT, instead of dividing the observation set into disjoint subsets, we use the entire observation set in every iteration. Hence, RAM-FIHT differs from AM-FIHT in the trimming step, and the thresholding is set as the ground truth μ throughout this section. AM-FIHT is tested with both $\beta = 0$ and $\beta = (1-p)^2/5$, while only $\beta = (1-p)^2/5$ is tested on RAM-FIHT. An algorithm terminates if

$$\|\mathcal{P}_{\widehat{\Omega}}(\boldsymbol{X}_l - \boldsymbol{X}_{l-1})\|_F / \|\mathcal{P}_{\widehat{\Omega}}(\boldsymbol{X}_{l-1})\|_F \leq 10^{-6} \qquad (27)$$

is satisfied before reaching the maximum iteration number, which is set as 300 here.

Figs. 2, 3, and 4 show the recovery phase transitions of AM-FIHT and RAM-FIHT with missing data patterns following different modes. $n = 300$, $n_1 = 150$, and $n_c = 30$. The x-axis is the fraction of observations $p = \frac{m}{n_c n}$. The y-axis is the rank r. For each fixed p and r, we generate 100 independent realizations of synthetic data matrices and data erasures. We say the recovery is successful in a test case if

$$\|\mathcal{P}_{\widehat{\Omega}^c}(\boldsymbol{X}_l - \boldsymbol{X})\|_F / \|\mathcal{P}_{\widehat{\Omega}^c}(\boldsymbol{X})\|_F < 10^{-3} \qquad (28)$$

holds when the algorithm terminates after l-th iteration, and $\widehat{\Omega}^c$ is the complement of $\widehat{\Omega}$. A white block corresponds to 100% success, and a black one means failures in all 100 tests.

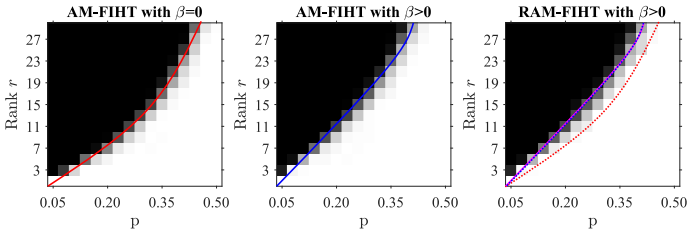

Fig. 2: Phase transition under Mode 1

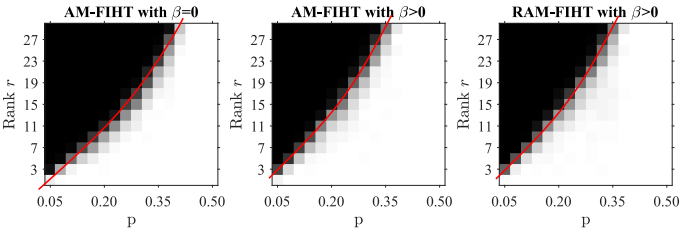

Fig. 3: Phase transition under Mode 2

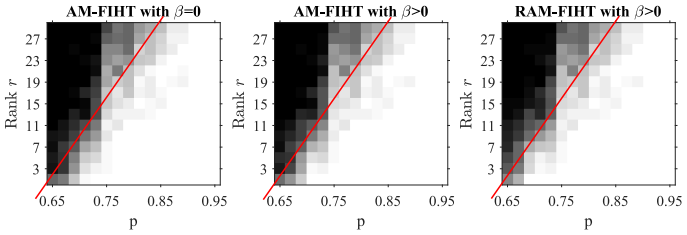

Fig. 4: Phase transition under Mode 3

Auxiliary solid lines (red for AM-FIHT with $\beta = 0$, blue for AM-FIHT with $\beta > 0$, and magenta for RAM-FIHT) are added in Figs. 2, 3, and 4 to highlight the phase transition. In the subfigures for RAM-FIHT, the phase transition curves for AM-FIHT are repeated in dotted curves to compare. Both AM-FIHT and RAM-FIHT with $\beta > 0$ preform very similarly, as the blue dotted line and the magenta solid line coincide in all three modes.

The phase transition threshold of $\beta > 0$ is higher than that of $\beta = 0$ for all the modes. The recovery improvement by the heavy-ball step can be intuitively explained as follows. Note that Theorem 2 shows that the heavy ball can speed up the convergence by reducing $q(0)$ to $q(\beta)$. With a certain percentage of data losses, it might hold that $q(\beta) < 1 < q(0)$, which indicates that the iterates with $\beta > 0$ are still convergent, while those with $\beta = 0$ may be divergent.

One can see from the phase transition lines that the required ratio of observations is approximately linear in r when other parameters are fixed. Note that the degree of freedom of the signal is $\Theta(n_c r)$. Although our bound of the required number of measurements $O(r^2 \log n)$ in Theorem 3 is not order-wise optimal due to the artefacts of the proof, the required number of measurements in numerical experiments is approximately linear in the degree of freedom.

2) Comparison with existing algorithms: We next study the recovery performance with both noiseless and noisy measurements. Here, rank r is fixed as 15, and $n = 600, n_1 = 300, n_c = 20$. All the other setups remain the same. We compare our methods with FIHT [7] and Singular Value Thresholding (SVT) [6]. Since FIHT recovers the missing points in a single channel, we convert each row of $\boldsymbol{X}$ to a Hankel matrix with the size of 300×301 and apply FIHT separately. SVT solves the convex NNM problem approximately, and the algorithm is applied on the original observation matrix and the constructed Hankel matrix, respectively. The relative recovery error is calculated as $\|\mathcal{P}_{\widehat{\Omega}^c}(\boldsymbol{X}_l - \boldsymbol{X})\|_F / \|\mathcal{P}_{\widehat{\Omega}^c}(\boldsymbol{X})\|_F$.

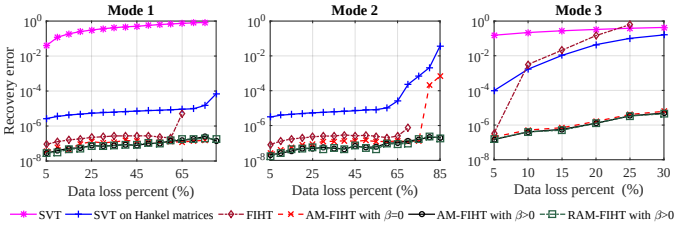

Fig. 5: Performance comparison of recovery methods in noiseless setting

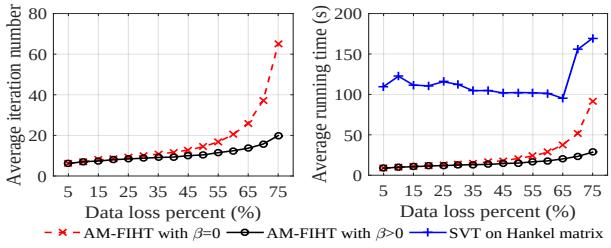

Fig. 6: Running time comparison in Mode 1

Fig. 5 compares the relative recovery error of convergent tests by different methods with noiseless measurements and different data loss patterns. (R)AM-FIHT with a nonzero β performs the best among all the methods. As the original data matrix is not low-rank, SVT fails in all cases. When applied to the constructed Hankel matrix, SVT exhibits better

performance, however, the recovery errors are still much larger compared with (R)AM-FIHT. SVT also needs the much longer running time, as shown in Fig. 6. To achieve the error bound of 10^{-5}, SVT requires around 100 iterations at a time cost of 100 seconds, while AM-FIHT with a nonzero β only takes less than 12 seconds to obtain an error bound of 10^{-7}. With 65% data loss in Mode 1, 13.3% tests of FIHT diverge. In contrast, all tests of AM-FIHT are convergent, even in the case with 75% data loss. A nonzero β also increases the percentage of convergent tests. With 80% of data loss, only 76.7% tests of AM-FIHT with $\beta = 0$ converge, while all the tests of AM-FIHT with $\beta > 0$ are convergent. Moreover, AM-FIHT performs much better than FIHT and SVT in Mode 3.

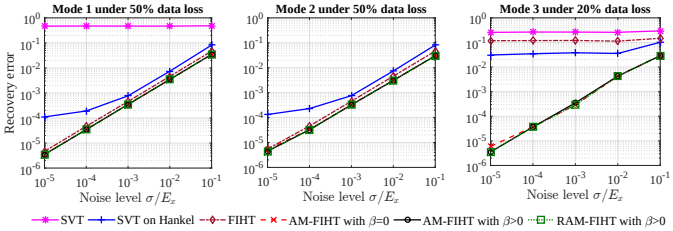

Fig. 7: Performance comparison of recovery methods in noisy setting under 50% data loss in modes 1, 2 and 20% data loss in mode 3

When measurements are noisy, every entry of $\boldsymbol{N}$ is independently drawn from Gaussian $\mathcal{N}(0, \sigma^2)$, where σ is the standard deviation. Fig. 7 shows the relative recovery error of convergent tests against the relative noise level σ/E_x, where E_x is the average energy of $\boldsymbol{X}$ calculated as $E_x = ||\boldsymbol{X}||_F/\sqrt{n_c n}$. The data loss percentage is fixed as 50% in modes 1, 2 and 20% for mode 3, respectively. In all three modes, AM-FIHT and RAM-FIHT perform similarly and achieve the smallest error among all the methods. The relative recovery error is proportional to the relative noise level, with a ratio between 0.3 to 0.4. FIHT is sightly worse than these two methods in modes 1 and 2, but its performance degrades significantly in mode 3. SVT has a better performance when applied on the Hankel matrix instead of the data matrix, but it is still worse than (R)AM-FIHT.

B. Numerical experiments on actual PMU data

The low-rank property of the Hankel PMU data matrix is verified on a recorded PMU dataset in Central New York [18]. 11 voltage phasors are measured at a rate of 30 samples per second. Fig. 8 shows the recorded voltage magnitudes and angles of a 10-second dataset that contains a disturbance at 2.3s. Fig. 9 shows the approximation errors of $\boldsymbol{X}$ and $\mathcal{H}\boldsymbol{X}$ by rank-r matrices with varying n_1 and r. The approximation error of $\boldsymbol{X}$ with the rank-r matrix $\mathcal{Q}_r(\boldsymbol{X})$ is defined as $\|\boldsymbol{X} - \mathcal{Q}_r(\boldsymbol{X})\|_F/\|\boldsymbol{X}\|_F$, and likewise for $\mathcal{H}\boldsymbol{X}$. One can see from Fig. 9 that all these data matrices can be approximated by rank-8 matrices with negligible errors.

We set $n_1 = 8, r = 8$ and test both $\beta = 0$ and $\beta = (1-p)/5$ in the simulation. Fig. 10 shows the percentage of convergent runs out of 100 runs for different algorithms. Fig. 11 compares the average recovery error of convergent runs. Overall, AM-FIHT with $\beta > 0$ achieves a small recovery error, tolerates a high data loss rate, and does not require much computation. For example, when the data loss rate is 55% in Mode 2, AM-FIHT with $\beta = (1-p)/5$ converges every time. The number of iterations is 47.2 on average, and the running time is 0.62 seconds. It takes 4.34 seconds to run 400 iterations of SVT on Hankel matrices. AM-FIHT with $\beta = 0$ diverges for 95% of the runs. FIHT diverges completely. Similar result to Fig. 6 about the average iteration numbers of AM-FIHT with $\beta = 0$ and $\beta > 0$ from respective successful trials is observed as well, thus a small positive β helps improve the convergence rate.

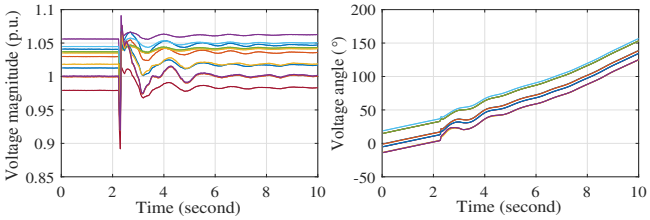

Fig. 8: The measured voltage phasors of 11 channels

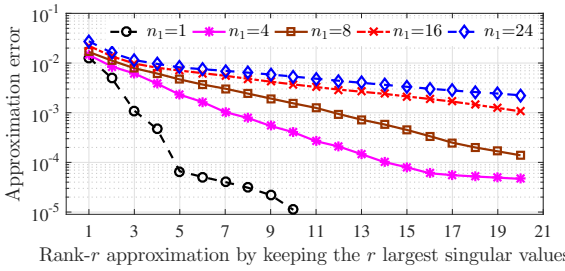

Fig. 9: The approximation errors of the data block and the Hankel matrices

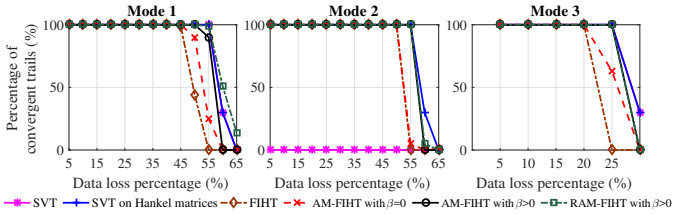

Fig. 10: Percentage of convergent trials of recovery algorithms

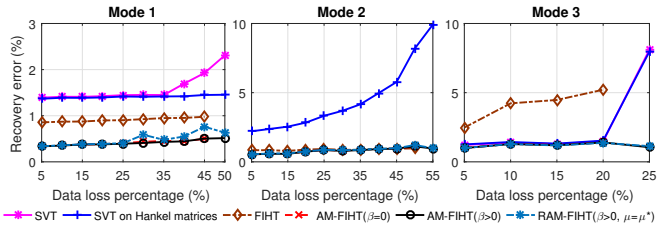

Fig. 11: Performance comparison of recovery algorithms

There are minor differences between AM-FIHT ($\beta > 0$) and RAM-FIHT ($\beta > 0$) in mode 1. RAM-FIHT tolerates a slightly higher data loss percentage, while its average recovery error of convergent runs is slightly larger than that of AM-FIHT. AM-FIHT and RAM-FIHT perform similarly in modes 2 and 3.

VII. Conclusion and Discussions

This paper characterizes the intrinsic low-dimensional structures of correlated time series through multi-channel low-rank Hankel matrices. Two iterative hard thresholding algorithms

with linear convergence rates are proposed to solve the non-convex missing data recovery problem. Our bound of the required number of observed entries for successful recovery is $O(r^2 \log^2 n)$, significantly smaller than $O(rn \log^2 n)$ by conventional low-rank matrix completion methods. Our bound is slightly larger than the degree of the freedom $\Theta(n_c r)$, and we suspect that the bound can be improved with better proof techniques. The convergence rate is proved to be accelerated further by adding a heavy-ball step, which also increases the tolerable missing data percentage numerically.

One motivating application of our methods is power system synchrophasor data recovery. Other applications include array signal processing and MRI image recovery. This paper provides the first analytical characterization of multi-channel Hankel matrix completion methods, while existing works mostly focused on single-channel Hankel matrix recovery. One future direction is to study data recovery from both data losses and corruptions, where partial measurements contain significant errors. The bad data should be first located and removed before recovering the missing points.

ACKNOWLEDGEMENT

This research is supported in part by ARO W911NF-17-1-0407, NSF 1508875, EPRI 1007316, IBM Corporation, and the ERC Program of NSF and DoE under the supplement to NSF EEC-1041877 and the CURENT Industry Partnership Program. We thank New York Power Authority for providing recorded PMU datasets.

REFERENCES

[1] "The Netflix Prize," http://netflixprize.com/, June 2006.

[2] N. H. Abbasy and H. M. Ismail, "A unified approach for the optimal pmu location for power system state estimation," *IEEE Trans. Power Syst.*, vol. 24, no. 2, pp. 806–813, 2009.

[3] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in *Proc. the 24th International Conference on Machine Learning*. ACM, 2007, pp. 17–24.

[4] A. Balachandrasekaran, V. Magnotta, and M. Jacob, "Recovery of damped exponentials using structured low rank matrix completion," *IEEE Trans. Med. Imag.*, vol. 36, no. 10, pp. 2087–2098, 2017.

[5] A. Balachandrasekaran, G. Ongie, and M. Jacob, "Accelerated dynamic mri using structured low rank matrix completion," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 1858–1862.

[6] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[7] J.-F. Cai, T. Wang, and K. Wei, "Fast and provable algorithms for spectrally sparse signal reconstruction via low-rank hankel matrix completion," *Applied and Computational Harmonic Analysis*, 2017.

[8] E. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2053 –2080, 2010.

[9] E. J. Candés and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Communications on Pure and Applied Mathematics*, vol. 67, no. 6, pp. 906–956, 2014.

[10] E. J. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, 2010.

[11] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[12] Y. Chen and Y. Chi, "Robust spectral compressed sensing via structured matrix completion," *IEEE Trans. Inf. Theory*, vol. 60, no. 10, pp. 6576–6601, 2014.

[13] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, 2005.

[14] J. De La Ree, V. Centeno, J. S. Thorp, and A. G. Phadke, "Synchronized phasor measurement applications in power systems," *IEEE Trans. Smart Grid*, vol. 1, no. 1, pp. 20–27, 2010.

[15] T. Ding, M. Sznaier, and O. I. Camps, "A rank minimization approach to video inpainting," in *IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[16] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, 2002.

[17] M. Fazel, T. K. Pong, D. Sun, and P. Tseng, "Hankel matrix rank minimization with applications to system identification and realization," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 3, pp. 946–977, 2013.

[18] P. Gao, M. Wang, S. G. Ghiocel, J. H. Chow, B. Fardanesh, and G. Stefopoulos, "Missing data recovery by exploiting low-dimensionality in power system synchrophasor measurements," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 1006–1013, 2016.

[19] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[20] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1548–1566, 2011.

[21] J. P. Haldar, "Low-rank modeling of local k-space neighborhoods (loraks) for constrained mri," *IEEE Trans. Med. Imag.*, vol. 33, no. 3, pp. 668–681, 2014.

[22] E. C. Hall and R. M. Willett, "Online convex optimization in dynamic environments," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 4, pp. 647–662, 2015.

[23] R. A. Horn and C. R. Johnson, "Matrix analysis," *Cambridge University Press, New York*, 1985.

[24] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the 45th Annual ACM Symposium on Theory of Computing*, 2013, pp. 665–674.

[25] K. H. Jin, D. Lee, and J. C. Ye, "A general framework for compressed sensing and parallel mri using annihilating filter based low-rank hankel matrix," *IEEE Transactions on Computational Imaging*, vol. 2, no. 4, pp. 480–495, 2016.

[26] K. H. Jin and J. C. Ye, "Sparse and low-rank decomposition of a hankel structured matrix for impulse noise removal," *IEEE Trans. Image Process.*, vol. PP, no. 99, pp. 1–1, 2017.

[27] Y. Li and Y. Chi, "Off-the-grid line spectrum denoising and estimation with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 64, no. 5, pp. 1257–1269, March 2016.

[28] W. Liao and A. Fannjiang, "Music for single-snapshot spectral estimation: Stability and super-resolution," *Applied and Computational Harmonic Analysis*, vol. 40, no. 1, pp. 33–67, 2016.

[29] J. Liu, A. Eryilmaz, N. B. Shroff, and E. S. Bentley, "Heavy-ball: A new approach to tame delay and convergence in wireless network optimization," in *The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 2016, pp. 1–9.

[30] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[31] J. Ma, P. Zhang, H. Fu, B. Bo, and Z. Dong, "Application of phasor measurement unit on locating disturbance source for low-frequency oscillation," *IEEE Trans. Smart Grid*, vol. 1, no. 3, pp. 340–346, 2010.

[32] J. K. Merikoski and R. Kumar, "Inequalities for spreads of matrix sums and products," *Applied Mathematics E-Notes*, vol. 4, pp. 150–159, 2004.

[33] N. Mohammadiha and A. Leijon, "Nonnegative hmm for babble noise derived from speech hmm: Application to speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 998–1011, 2013.

[34] N. Mohammadiha, P. Smaragdis, G. Panahandeh, and S. Doclo., "A state-space approach to dynamic nonnegative matrix factorization," *IEEE Trans. Signal Process.*, vol. 63, no. 4, pp. 949–959, 2015.

[35] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 140–148.

[36] P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain, "Non-convex robust pca," in *NIPS*, 2014, pp. 1107–1115.

[37] G. Ongie, S. Biswas, and M. Jacob, "Convex recovery of continuous domain piecewise constant images from nonuniform fourier samples," *IEEE Trans. Signal Process.*, vol. 66, no. 1, pp. 236–250, 2018.

[38] G. Ongie and M. Jacob, "Off-the-grid recovery of piecewise constant images from few fourier samples," *SIAM Journal on Imaging Sciences*, vol. 9, no. 3, pp. 1004–1041, 2016.

[39] A. Phadke and J. Thorp, *Synchronized phasor measurements and their applications*. Springer, 2008.

[40] B. T. Polyak, "Introduction to optimization. 1987," *Optimization Software, Inc, New York*.

[41] L. C. Potter, E. Ertin, J. T. Parker, and M. Cetin, "Sparsity and compressed sensing in radar imaging," *Proc. IEEE*, vol. 98, no. 6, pp. 1006–1020, 2010.

[42] B. Recht, "A simpler approach to matrix completion," *The Journal of Machine Learning Research*, pp. 3413–3430, 2011.

[43] P. J. Shin, P. E. Z. Larson, M. A. Ohliger, M. Elad, J. M. Pauly, D. B. Vigneron, and M. Lustig, "Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion," *Magnetic Resonance in Medicine*, vol. 72, no. 4, pp. 959–970, 2014.

[44] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[45] K. Usevich and P. Comon, "Hankel low-rank matrix completion: Performance of the nuclear norm relaxation," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 637–646, 2016.

[46] K. Wei, J.-F. Cai, T. F. Chan, and S. Leung, "Guarantees of riemannian optimization for low rank matrix completion," *arXiv preprint arXiv:1603.06610*, 2016.

[47] Z. Yang and L. Xie, "Exact joint sparse frequency recovery via optimization methods," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5145–5157, 2016.

[48] J. C. Ye, J. M. Kim, K. H. Jin, and K. Lee, "Compressive sampling using annihilating filter-based low-rank interpolation," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 777–801, 2017.

[49] S. Zhang, Y. Hao, M. Wang, and J. H. Chow, "Multi-channel missing data recovery by exploiting the low-rank hankel structures," in *Proc. CAMSAP*, 2017.

APPENDIX

A. Notations and assumptions

We first introduce notations used in the proof. For matrix $\boldsymbol{Z}_1 \in \mathbb{C}^{n_c\times n}$, we define the **Block Hankel Operator** $\widetilde{\mathcal{H}}$ as

$$\widetilde{\mathcal{H}}\boldsymbol{Z}_1 = \begin{bmatrix}\mathcal{H}\boldsymbol{Z}_1 & \mathcal{H}\boldsymbol{Z}_1 & \cdots & \mathcal{H}\boldsymbol{Z}_1\end{bmatrix} \in \mathbb{C}^{n_c n_1\times n_c n_2}.$$

$\widetilde{\mathcal{H}}\boldsymbol{Z}_1$ is an n_c-block Hankel matrix. $\widetilde{\mathcal{H}}^*$ is the adjoint operator of $\widetilde{\mathcal{H}}$. For any matrix $\boldsymbol{Z}_2 \in \mathbb{C}^{n_c n_1\times n_c n_2}$, $(\widetilde{\mathcal{H}}^*\boldsymbol{Z}_2) \in \mathbb{C}^{n_c\times n}$ satisfies

$$\langle\widetilde{\mathcal{H}}^*\boldsymbol{Z}_2, \boldsymbol{e}_k\boldsymbol{e}_t^*\rangle = \sum_{l=0}^{n_c-1}\sum_{k_1+k_2=t+1}\langle \boldsymbol{Z}, \boldsymbol{e}_{(k_1-1)n_c+k}\boldsymbol{e}_{k_2+ln_2}^*\rangle.$$

Define $\widetilde{\mathcal{D}}^2 := \widetilde{\mathcal{H}}^*\widetilde{\mathcal{H}}$, an operator from an $n_c\times n$ matrix $\boldsymbol{Z}$ to an $n_c\times n$ matrix with

$$\widetilde{\mathcal{D}}^2\boldsymbol{Z} = \sum_{t=1}^{n}\sum_{k=1}^{n_c} n_c w_t\langle \boldsymbol{Z}, \boldsymbol{e}_k\boldsymbol{e}_t^*\rangle\boldsymbol{e}_k\boldsymbol{e}_t^*,$$

where w_t is defined in (15). Then the Moore-Penrose pseudoinverse of $\widetilde{\mathcal{H}}$, denoted as $\widetilde{\mathcal{H}}^\dagger$, equals to $\widetilde{\mathcal{D}}^{-2}\widetilde{\mathcal{H}}^*$. Further, we define $\widetilde{\mathcal{G}} = \widetilde{\mathcal{H}}\widetilde{\mathcal{D}}^{-1}$, then the adjoint operator of $\widetilde{\mathcal{G}}$ is defined as $\widetilde{\mathcal{G}}^* = \widetilde{\mathcal{D}}^{-1}\widetilde{\mathcal{H}}^*$. Additionally,

$$\boldsymbol{Y} := \widetilde{\mathcal{D}}\boldsymbol{X} \quad \text{and} \quad \boldsymbol{Y}_l := \widetilde{\mathcal{D}}\boldsymbol{X}_l. \tag{29}$$

For any matrix $\boldsymbol{Z}\in\mathbb{C}^{n_c\times n}$, one can check that $\|\widetilde{\mathcal{H}}\boldsymbol{X}\| = \sqrt{n_c}\,\|\mathcal{H}\boldsymbol{X}\|$ and $\|\widetilde{\mathcal{H}}\boldsymbol{X}\|_F = \sqrt{n_c}\,\|\mathcal{H}\boldsymbol{X}\|_F$. Immediately, $\widetilde{\mathcal{H}}\boldsymbol{X}$ and $\mathcal{H}\boldsymbol{X}$ share the same conditional number κ. Moreover, it is clear that $\widetilde{\mathcal{G}}$ is a unit operator as $\widetilde{\mathcal{G}}^*\widetilde{\mathcal{G}} = \mathcal{I}$, and $\widetilde{\mathcal{H}}\boldsymbol{X} = \widetilde{\mathcal{G}}\boldsymbol{Y}$.

The following proofs will be established on Block Hankel Operator $\widetilde{\mathcal{H}}$. Consider AM-FIHT in terms of $\widetilde{\mathcal{H}}$, the initialization can be written as $\widetilde{\boldsymbol{L}}_0 = p^{-1}\mathcal{Q}_r(\widetilde{\mathcal{H}}\mathcal{P}_\Omega(\boldsymbol{X}))$. Further, the major steps can be represented as

$$\widetilde{\boldsymbol{W}}_l = \mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{H}}(\boldsymbol{X}_l + p^{-1}\boldsymbol{G}_l + \beta\Delta\widetilde{\boldsymbol{W}}), \quad \widetilde{\boldsymbol{L}}_{l+1} = \mathcal{Q}_r(\widetilde{\boldsymbol{W}}_l),$$

where $\widetilde{\mathcal{S}}_l$ is the tangent subspace at $\widetilde{\boldsymbol{L}}_l$. The resulting $\widetilde{\boldsymbol{X}}_l$ returned by AM-FIHT in terms of $\widetilde{\mathcal{H}}$ is given as $\widetilde{\mathcal{H}}^\dagger\widetilde{\boldsymbol{L}}_l$.

Moreover, by replacing $\mathcal{H}$ with $\widetilde{\mathcal{H}}$, AM-FIHT returns the same result as $\widetilde{\boldsymbol{X}}_l = \boldsymbol{X}_l$. In other words, we can show that

$$\widetilde{\boldsymbol{L}}_l = \begin{bmatrix}\boldsymbol{L}_l & \boldsymbol{L}_l & \cdots & \boldsymbol{L}_l\end{bmatrix}, \forall l \geq 0. \tag{30}$$

To see this, it is clear that (30) holds for $l = 0$ from the definition of $\widetilde{\boldsymbol{L}}_0$. Then suppose (30) holds when $l = k$. Immediately, we have $\widetilde{\boldsymbol{W}}_k = \begin{bmatrix}\boldsymbol{W}_k & \boldsymbol{W}_k & \cdots & \boldsymbol{W}_k\end{bmatrix}$. Let $\boldsymbol{W}_k = \sum_{i=1}^{\min(n_c n_1, n_2)}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^*$ be the SVD with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(n_c n_1, n_2)}$. Then for $l = k+1$,

$$\begin{aligned}\widetilde{\boldsymbol{L}}_{k+1} = \mathcal{Q}_r(\widetilde{\boldsymbol{W}}_k) &= \sum_{i=1}^{r}\sqrt{n_c}\sigma_i\boldsymbol{u}_i\frac{1}{\sqrt{n_c}}\begin{bmatrix}\boldsymbol{v}_i^* & \cdots & \boldsymbol{v}_i^*\end{bmatrix}\\ &= \sum_{i=1}^{r}\begin{bmatrix}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^* & \cdots & \sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^*\end{bmatrix}.\\ &= \begin{bmatrix}\boldsymbol{L}_{k+1} & \boldsymbol{L}_{k+1} & \cdots & \boldsymbol{L}_{k+1}\end{bmatrix}.\end{aligned}$$

Hence, the connection between $\boldsymbol{X}_l$ and $\widetilde{\boldsymbol{L}}_l$ can be given as

$$\begin{aligned}\|\boldsymbol{X}_l - \boldsymbol{X}\|_F = \|\widetilde{\mathcal{D}}^{-1}(\boldsymbol{Y}_l - \boldsymbol{Y})\|_F &\leq \frac{1}{\sqrt{n_c}}\|\boldsymbol{Y}_l - \boldsymbol{Y}\|_F\\ = \frac{1}{\sqrt{n_c}}\|\widetilde{\mathcal{G}}^*(\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y})\|_F &\leq \frac{1}{\sqrt{n_c}}\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F.\end{aligned} \tag{31}$$

For RAM-FIHT, similarly define an n_c-block matrix $\widetilde{\boldsymbol{L}}_l' = \begin{bmatrix}\boldsymbol{L}_l' & \boldsymbol{L}_l' & \cdots & \boldsymbol{L}_l'\end{bmatrix}$. From the discussion above, one can verify that (30) also holds in RAM-FIHT. Then the SVD of $\widetilde{\boldsymbol{L}}_l$ in (30) is $\widetilde{\boldsymbol{L}}_l = \widetilde{\boldsymbol{U}}_l\widetilde{\boldsymbol{\Sigma}}_l\widetilde{\boldsymbol{V}}_l^* = \boldsymbol{U}_l\big(\sqrt{n_c}\boldsymbol{\Sigma}_l\big)\big(\frac{1}{\sqrt{n_c}}[\boldsymbol{V}_l^*\cdots\boldsymbol{V}_l^*]\big)$. Then $\widetilde{\boldsymbol{A}}_l$ and $\widetilde{\boldsymbol{B}}_l$ are defined as

$$(\widetilde{\boldsymbol{A}}_l)_{i*} = \frac{(\widetilde{\boldsymbol{U}}_l)_{i*}}{\big\|(\widetilde{\boldsymbol{U}}_l)_{i*}\big\|}\min\left\{\big\|(\widetilde{\boldsymbol{U}}_l)_{i*}\big\|, \sqrt{\frac{\mu r}{n_c n_1}}\right\}, \tag{32}$$

$$(\widetilde{\boldsymbol{B}}_l)_{i*} = \frac{(\widetilde{\boldsymbol{V}}_l)_{i*}}{\big\|(\widetilde{\boldsymbol{V}}_l)_{i*}\big\|}\min\left\{\big\|(\widetilde{\boldsymbol{V}}_l)_{i*}\big\|, \sqrt{\frac{\mu r}{n_c n_2}}\right\}. \tag{33}$$

Sampling model with replacement. As shown in [42], due to the duplications, the number of observed entries in a sampling model with replacement is less than or equal to that in a sampling model without replacement. Thus, it is sufficient to study the sampling model with replacement. To distinguish $\widehat{\Omega}$ in (10), which represents the sampling set without replacement, let Ω be m unions of indices sampled uniformly from set $\{1,2,\cdots,n_c\}\times\{1,2,\cdots,n\}$ with replacement, and

$$\mathcal{P}_\Omega(\boldsymbol{Z}_1) = \sum_{a=1}^{m}\langle\boldsymbol{Z}_1, \boldsymbol{e}_{k_a}\boldsymbol{e}_{t_a}^*\rangle\boldsymbol{e}_{k_a}\boldsymbol{e}_{t_a}^*, \tag{34}$$

for any $\boldsymbol{Z}_1\in\mathbb{C}^{n_c\times n}$. By changing the sampling model, several critical lemmas can be derived from Bernstein Inequality.

Lemma 1 ([44], Theorem 1.6)**.** *Consider a finite sequence $\{\boldsymbol{Z}_k\}$ of independent, random matrices with dimensions $d_1\times d_2$. Assume that such random matrix satisfies*

$$\mathbb{E}(\boldsymbol{Z}_k) = 0 \quad \textit{and} \quad \|\boldsymbol{Z}_k\| \leq R \quad \textit{almost surely.}$$

Define

$$\sigma^2 := \max\left\{\Big\|\sum_k\mathbb{E}(\boldsymbol{Z}_k\boldsymbol{Z}_k^*)\Big\|, \Big\|\sum_k\mathbb{E}(\boldsymbol{Z}_k^*\boldsymbol{Z}_k)\Big\|\right\}.$$

Then for all $t\geq 0$,

$$\mathbb{P}\left\{\left\|\sum_k\boldsymbol{Z}_k\right\| \geq t\right\} \leq (d_1+d_2)\exp\Big(\frac{-t^2/2}{\sigma^2 + Rt/3}\Big).$$

Suppose that $t \le \sigma^2/R$, then the right hand side can be released as $(d_1+d_2)\exp(-\frac{3}{8}t^2/\sigma^2)$. Such kind of manipulation will be adopted in several proofs.

Note that the set of matrices

$$\{\widetilde{\boldsymbol{H}}_{k.t}|\widetilde{\boldsymbol{H}}_{k,t} = \frac{1}{\sqrt{n_c w_t}}\widetilde{\mathcal{H}}(\boldsymbol{e}_k\boldsymbol{e}_t^*), 1 \le k \le n_c, 1 \le t \le n\}$$

forms an orthonormal basis of the n_c-block Hankel matrix, where $\widetilde{\mathcal{H}}\boldsymbol{X} = \sum_{k=1}^{n_c}\sum_{t=1}^{n}\langle\widetilde{\boldsymbol{H}}_{k,t}, \widetilde{\mathcal{H}}\boldsymbol{X}\rangle\widetilde{\boldsymbol{H}}_{k,t}$. Then, for all $(k_a, t_a) \in \Omega$, $\mathcal{P}_\Omega$ is also used as the operator

$$\mathcal{P}_\Omega(\boldsymbol{Z}_2) = \sum_{a=1}^{m}\langle\boldsymbol{Z}_2, \widetilde{\boldsymbol{H}}_{k_a,t_a}\rangle\widetilde{\boldsymbol{H}}_{k_a,t_a},$$

for any $\boldsymbol{Z}_2 \in \mathbb{C}^{n_c n_1 \times n_c n_2}$. In spite of abuse of notation, the meaning of $\mathcal{P}_\Omega$ is clear from context. By such definition, $\mathcal{P}_\Omega(\widetilde{\mathcal{H}}\boldsymbol{Z}_1) = \widetilde{\mathcal{H}}\mathcal{P}_\Omega(\boldsymbol{Z}_1)$ for any $\boldsymbol{Z}_1 \in \mathbb{C}^{n\times n_c}$. Additionally, $\widetilde{\boldsymbol{H}}_{k,t}$ only has $n_c w_t$ nonzero entries of magnitude $1/\sqrt{n_c w_t}$, so $\|\widetilde{\boldsymbol{H}}_{k,t}\|_F = 1$. The following lemma can be established directly from the definition of incoherence.

Lemma 2. *Let $\widetilde{\mathcal{H}}\boldsymbol{X} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^*$ be the SVD of $\widetilde{\mathcal{H}}\boldsymbol{X}$. Assume $\widetilde{\mathcal{H}}\boldsymbol{X}$ is μ-incoherent. Then*

$$\left\|\boldsymbol{e}_{k_1}^*\widetilde{\boldsymbol{U}}\right\|^2 \le \frac{\mu c_s r}{n_c n} \quad , \quad \left\|\boldsymbol{e}_{k_2}^*\widetilde{\boldsymbol{V}}\right\|^2 \le \frac{\mu c_s r}{n_c n}, \tag{35}$$

$$\left\|\mathcal{P}_{\widetilde{\boldsymbol{U}}}(\widetilde{\boldsymbol{H}}_{k,t})\right\|_F^2 \le \frac{\mu c_s r}{n_c n} \quad , \quad \left\|\mathcal{P}_{\widetilde{\boldsymbol{V}}}(\widetilde{\boldsymbol{H}}_{k,t})\right\|_F^2 \le \frac{\mu c_s r}{n_c n}. \tag{36}$$

where $\boldsymbol{e}_{k_1}, \boldsymbol{e}_{k_2}$ are the coordinate unit vectors.

B. Supporting Lemmas for Theorem 1

We first present some supporting lemmas to prove Theorem 1. Lemma 3 shows that the maximum number of repetitions is bounded by $O(\log n)$ with high probability in uniform sampling. Lemma 4 derives the properties of $p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}$, and the random operator can be close enough to its mean $\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}$ with a significant large amount of observed entries. Lemma 5 connects the angle of two subspaces, represented as $\|\mathcal{P}_{\widetilde{\mathcal{S}}_l} - \mathcal{P}_{\widetilde{\mathcal{S}}}\|$, with $\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F$, and shows the angle decreases as $\boldsymbol{L}_l$ approaches to the ground truth. Lemma 6 indicates the distance between the initial point and ground truth. Lemmas 4 and 6 are built upon Lemmas 5 and 2 of [7], respectively, by extending from single-channel signals to multi-channel signals. When $n_c = 1$, Lemmas 4 and 6 are reduced to corresponding lemmas in [7].

Lemma 3. *Under sampling with replacement model, the maximum number of repetitions of any entry in Ω is less than $3\log(n)$ with probability at least $1 - n_c n^{-2}$ for $n \ge 12$.*

Lemma 4. *Let $\widetilde{\mathcal{S}}$ be the tangent subspace of $\widetilde{\mathcal{H}}\boldsymbol{X}$. Assume $\widetilde{\mathcal{H}}\boldsymbol{X}$ is μ-incoherent. Then with $m \ge 32\mu c_s r\log(n)$,*

$$\left\|\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}} - p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}\right\| \le \sqrt{\frac{32\mu c_s r\log(n)}{m}}$$

holds with probability at least $1 - n_c n^{-2}$.

Lemma 5 ([46], Lemma 4.1)**.** *Let $\boldsymbol{Z}_l$ be a rank-r matrix and $\mathcal{S}_l$ be the tangent subspace of $\boldsymbol{Z}_l$. If $\boldsymbol{Z}$ is also a rank-r matrix and its tangent subspace is denoted as $\mathcal{S}$, then*

$$\|(\mathcal{I} - \mathcal{P}_{\mathcal{S}_l})(\boldsymbol{Z}_l - \boldsymbol{Z})\|_F \le \frac{\|\boldsymbol{Z}_l - \boldsymbol{Z}\|_F^2}{\sigma_{\min}(\boldsymbol{Z})}, \tag{37}$$

$$\|\mathcal{P}_{\mathcal{S}_l} - \mathcal{P}_{\mathcal{S}}\| \le \frac{2\|\boldsymbol{Z}_l - \boldsymbol{Z}\|_F}{\sigma_{\min}(\boldsymbol{Z})}. \tag{38}$$

Lemma 6. *Assume $\widetilde{\mathcal{H}}\boldsymbol{X}$ is μ-incoherent. With the initial point $\widetilde{\boldsymbol{L}}_0 = \mathcal{Q}_r(p^{-1}\widetilde{\mathcal{H}}\mathcal{P}_\Omega(\boldsymbol{X}))$, if $m \ge 16\mu c_s r\log(n)$, we have*

$$\left\|\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{H}}\boldsymbol{X}\right\| \le \sqrt{\frac{64\mu c_s r\log(n)}{m}}\left\|\widetilde{\mathcal{H}}\boldsymbol{X}\right\|$$

holds with probability at least $1 - n_c n^{-2}$.

C. Proof of Theorem 1

The proof of Theorem 1 is extended from that of Theorem 3 in [7] with some modifications. The majority of the efforts are devoted to the "Inductive Step" to build the connections between $\boldsymbol{W}_{l-1}$ and $\boldsymbol{W}_l$ through (43). In (43), the major issue is to bound I_1, I_2, I_3 and I_4, and (44) and (47) provide critical steps in bounding these items. This inductive step is built upon a similar analysis for $\boldsymbol{L}_l$'s in [7]. Here we study $\boldsymbol{W}_l$'s instead of $\boldsymbol{L}_l$'s since the analysis of Theorem 2 is based on analyzing $\boldsymbol{W}_l$'s. Although Lemma 6 provides the theoretical bound for $\boldsymbol{L}_0$ directly, a similar result for $\boldsymbol{W}_0$ is lacking. Thus, some efforts to analyze $\boldsymbol{W}_0$ is needed in the "Base Case" part. (24) in Theorem 1 is obtained from (51), which provides the theoretical bound for the required number of observations to ensure successful recovery. We include detailed steps as follows for the completeness of the proof.

Proof of Theorem 1. As $\widetilde{\boldsymbol{L}}_{l+1} = \mathcal{Q}_r(\widetilde{\boldsymbol{W}}_l)$, $\widetilde{\boldsymbol{L}}_{l+1}$ is the best rank-r approximation to $\widetilde{\boldsymbol{W}}_l$. Then we have

$$\begin{aligned}\|\widetilde{\boldsymbol{L}}_{l+1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F &\le \|\widetilde{\boldsymbol{W}}_l - \widetilde{\boldsymbol{L}}_{l+1}\|_F + \|\widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \\ &\le 2\|\widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F.\end{aligned} \tag{39}$$

Therefore, it is sufficient to bound $\|\widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F$. Lemma 3 suggests that with probability at least $1 - n_c n^{-2}$,

$$\|\mathcal{P}_\Omega\| \le 3\log(n) \tag{40}$$

holds. Lemma 4 suggests as long as $m \ge 32\varepsilon_0^{-2}\mu c_s r\log(n)$,

$$\left\|\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}} - p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}\right\| \le \varepsilon_0 \tag{41}$$

holds with probability at least $1 - n_c n^{-2}$.

Now we will show that the following inequality holds by mathematical induction,

$$\frac{\|\widetilde{\boldsymbol{W}}_k - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})} \le \frac{p^{1/2}\varepsilon_0}{12\log(n)(1+\varepsilon_0)}. \tag{42}$$

Inductive Step: Suppose (42) holds for $k = l-1$. Recall that

$$\begin{aligned}\widetilde{\boldsymbol{W}}_l &= \mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{H}}(\boldsymbol{X}_l + p^{-1}\mathcal{P}_\Omega(\boldsymbol{X} - \boldsymbol{X}_l)) \\ &= \mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\boldsymbol{Y}_l + p^{-1}\mathcal{P}_\Omega(\boldsymbol{Y} - \boldsymbol{Y}_l)).\end{aligned}$$

Then, for all $l \geq 1$, we have

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\right\|_F = \left\|\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\boldsymbol{Y}_l + p^{-1}\mathcal{P}_\Omega(\boldsymbol{Y}-\boldsymbol{Y}_l) - \widetilde{\mathcal{G}}\boldsymbol{Y}\right\|_F \\
=&\left\|\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}\boldsymbol{Y} - \widetilde{\mathcal{G}}\boldsymbol{Y} + (\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}} - p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_\Omega)(\boldsymbol{Y}_l - \boldsymbol{Y})\right\|_F \\
\overset{(a)}{\leq}&\left\|(\mathcal{I}-\mathcal{P}_{\widetilde{\mathcal{S}}_l})(\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
&+\left\|(\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^* - p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*)(\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
\leq&\left\|(\mathcal{I}-\mathcal{P}_{\widetilde{\mathcal{S}}_l})(\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F + \left\|\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*(\mathcal{I}-\mathcal{P}_{\widetilde{\mathcal{S}}_l})(\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
&+\left\|(\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l} - p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l})(\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
&+p^{-1}\left\|\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*(\mathcal{I}-\mathcal{P}_{\widetilde{\mathcal{S}}_l})(\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
:=& I_1 + I_2 + I_3 + I_4,
\end{aligned} \tag{43}$$

where (a) holds since $\widetilde{\boldsymbol{L}}_l = \mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\boldsymbol{W}}_{l-1}$. By applying (37),

$$\begin{aligned}
&I_1 + I_2 + I_4 \\
\leq&\frac{2\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F^2}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})} + p^{-1}\left\|\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l}\right\|\frac{\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F^2}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})} \\
\leq&\frac{8\|\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F^2}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})} + 4p^{-1}\left\|\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l}\right\|\frac{\|\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F^2}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}.
\end{aligned} \tag{44}$$

Next, we will bound $\|\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l}\|$. For any $\boldsymbol{Z} \in \mathbb{C}^{n_c n_1 \times n_c n_2}$,

$$\begin{aligned}
\|\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}(\boldsymbol{Z})\|^2 =&\langle\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}(\boldsymbol{Z}), \mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}(\boldsymbol{Z})\rangle \\
\overset{(b)}{\leq}&3\log(n)\langle\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}(\boldsymbol{Z}), \mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}(\boldsymbol{Z})\rangle \\
=&3\log(n)\langle\boldsymbol{Z}, \mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}(\boldsymbol{Z})\rangle \\
\overset{(c)}{\leq}&3\log(n)(1+\varepsilon_0)p\,\|\boldsymbol{Z}\|_F^2\,.
\end{aligned} \tag{45}$$

where (b) holds due to (40), and (c) holds due to (41). Hence,

$$\left\|\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\right\| = \left\|\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}\right\| \leq \sqrt{3\log(n)(1+\varepsilon_0)p},$$

and

$$\begin{aligned}
&\left\|\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l}\right\| \leq \left\|\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*(\mathcal{P}_{\widetilde{\mathcal{S}}_l} - \mathcal{P}_{\widetilde{\mathcal{S}}})\right\| + \left\|\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}\right\| \\
\overset{(a)}{\leq}&3\log(n)\frac{2\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})} + \left\|\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}\right\| \\
\overset{(b)}{\leq}&3\log(n)\frac{p^{1/2}\varepsilon_0}{3\log(n)(1+\varepsilon_0)} + \sqrt{3\log(n)(1+\varepsilon_0)p} \\
\leq&3\log(n)(1+\varepsilon_0)p^{1/2},
\end{aligned} \tag{46}$$

where (a) holds due to (38) and (40), and (b) holds due to (39) and the inductive hypothesis.

Hence, $I_1 + I_2 + I_4 \leq 2\varepsilon_0\|\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F$. Moreover,

$$\begin{aligned}
&\left\|\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l} - p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l}\right\| \\
\leq&\left\|\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}} - p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}\right\| + \left\|(\mathcal{P}_{\widetilde{\mathcal{S}}} - \mathcal{P}_{\widetilde{\mathcal{S}}_l})\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l}\right\| \\
&+\left\|\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*(\mathcal{P}_{\widetilde{\mathcal{S}}} - \mathcal{P}_{\widetilde{\mathcal{S}}_l})\right\| + \left\|p^{-1}(\mathcal{P}_{\widetilde{\mathcal{S}}} - \mathcal{P}_{\widetilde{\mathcal{S}}_l})\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l}\right\| \\
&+\left\|p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*(\mathcal{P}_{\widetilde{\mathcal{S}}} - \mathcal{P}_{\widetilde{\mathcal{S}}_l})\right\| \\
\overset{(c)}{\leq}&\varepsilon_0 + \frac{2\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}\left(2 + p^{-1}\|\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_l}\| + p^{-1}\|\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\|\right) \\
\overset{(d)}{\leq}&4\varepsilon_0.
\end{aligned} \tag{47}$$

where (c) comes from Lemma 5, and (d) comes from (42) and (46). Then, I_3 can be bounded as

$$I_3 \leq 4\varepsilon_0\|\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F. \tag{48}$$

By putting pieces together, we have

$$\|\widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \leq \nu\|\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F. \tag{49}$$

Hence, (42) still holds for $k = l$.

Base Case: Let us assume

$$\|\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \leq \frac{p^{1/2}\varepsilon_0}{6\log(n)(1+\varepsilon_0)}. \tag{50}$$

Then, similar to $I_1 + I_2 + I_3 + I_4$ in (43), we have

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{W}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\right\|_F = \left\|\mathcal{P}_{\widetilde{\mathcal{S}}_0}\widetilde{\mathcal{G}}(\boldsymbol{Y}_0 + p^{-1}\mathcal{P}_\Omega(\boldsymbol{Y}-\boldsymbol{Y}_0)) - \widetilde{\mathcal{G}}\boldsymbol{Y}\right\|_F \\
\leq&\left\|(\mathcal{I}-\mathcal{P}_{\widetilde{\mathcal{S}}_0})(\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F + \left\|\mathcal{P}_{\widetilde{\mathcal{S}}_0}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*(\mathcal{I}-\mathcal{P}_{\widetilde{\mathcal{S}}_0})(\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
&+\left\|(\mathcal{P}_{\widetilde{\mathcal{S}}_0}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_0} - p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}_0}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}_0})(\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
&+p^{-1}\left\|\mathcal{P}_{\widetilde{\mathcal{S}}_0}\widetilde{\mathcal{G}}\mathcal{P}_\Omega\widetilde{\mathcal{G}}^*(\mathcal{I}-\mathcal{P}_{\widetilde{\mathcal{S}}_0})(\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
\leq&5\varepsilon_0\|\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F.
\end{aligned}$$

Since $\varepsilon_0 \in (0, 1/10)$, we have

$$\|\widetilde{\boldsymbol{W}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \leq \frac{1}{2}\|\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \leq \frac{p^{1/2}\varepsilon_0}{12\log(n)(1+\varepsilon_0)},$$

which completes the induction part.

Then the only thing is to check the assumption (50). Using Lemma 6 and $\|\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \leq \sqrt{2r}\|\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|$, with probability at least $1 - n_c n^{-2}$,

$$\frac{\|\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})} \leq \frac{\sqrt{2r}\|\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})} = \kappa\sqrt{\frac{128\mu c_s r^2\log(n)}{m}}.$$

Therefore, to guarantee (50), we need

$$\kappa\sqrt{\frac{128\mu c_s r^2\log(n)}{m}} \leq \frac{p^{1/2}\varepsilon_0}{6\log(n)(1+\varepsilon_0)}, \tag{51}$$

That is $m \geq C_1(1+\varepsilon_0)\varepsilon_0^{-1}n_c^{1/2}\mu^{1/2}c_s^{1/2}\kappa r n^{1/2}\log^{3/2}(n)$ with $C_1 = 48\sqrt{2}$.

Hence, with probability at least $1-(2l+1)n_c n^{-2}$, from (49),

$$\begin{aligned}
\|\boldsymbol{Y}_l - \boldsymbol{Y}\|_F = \|\widetilde{\mathcal{G}}^*(\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y})\|_F \leq& \|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \\
\leq& 2\|\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \leq 2\nu^{l-1}\|\widetilde{\boldsymbol{W}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \\
\leq& \nu^{l-1}\|\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F,
\end{aligned} \tag{52}$$

where $\boldsymbol{Y}_l = \widetilde{\mathcal{G}}^*\widetilde{\boldsymbol{L}}_l, \widetilde{\mathcal{G}}^*\widetilde{\mathcal{G}} = \mathcal{I}$ and $\|\widetilde{\mathcal{G}}^*\| \leq 1$. □

D. Proof of Theorem 2

First, we extend the eigenvalues and eigenvectors of linear operators on vector spaces to the eigenvalues and eigenmatrices of linear operators on matrix spaces, defined as follows.

Definition 2. *Let $\mathcal{A}$ denote a linear operator from $\mathbb{C}^{l_1\times l_2}$ to $\mathbb{C}^{l_1\times l_2}$, for any matrix $\boldsymbol{M}$ in the space and $\boldsymbol{M} \neq \boldsymbol{0}$, if $\mathcal{A}\boldsymbol{M} = \lambda\boldsymbol{M}$ holds, then $\boldsymbol{M}$ is one eigenmatrix of operator $\mathcal{A}$, and λ is the corresponding eigenvalue.*

Let $\mathcal{L}$ denote the following linear operator on the matrix space $\mathbb{C}^{n_c n_1 \times n_c n_2}$,

$$\mathcal{L} = \mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}} - p^{-1}\mathcal{P}_{\widetilde{\mathcal{S}}}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widetilde{\mathcal{S}}}.$$

We first introduce Lemmas 7, 8 and 9 that are useful in the proof of Theorem 2.

Lemma 7. *Suppose that for any $\epsilon > 0$, there always exists an integer s_ϵ such that for any integer $k \geq 0$, the iterates $\widetilde{\boldsymbol{W}}_{s_\epsilon+k}$ generated by AM-FIHT satisfy $\|\widetilde{\boldsymbol{W}}_{s_\epsilon+k} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \leq \epsilon$. Then with any $l > s_\epsilon + 1$, the updated rule can be denoted as*

$$\begin{bmatrix} \widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y} \\ \widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y} \end{bmatrix} = \begin{bmatrix} \mathcal{L}(\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}) + \beta\mathcal{P}_{\widetilde{\mathcal{S}}}(\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\boldsymbol{W}}_{l-2}) \\ \widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y} \end{bmatrix} + \widetilde{\boldsymbol{Z}}_{l-1},$$

where

$$\left\|\widetilde{\boldsymbol{Z}}_{l-1}\right\|_F = o\left(\left\|\begin{bmatrix} \widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y} \\ \widetilde{\boldsymbol{W}}_{l-2} - \widetilde{\mathcal{G}}\boldsymbol{Y} \end{bmatrix}\right\|_F\right).$$

Lemma 8. *All the eigenvalues of operator $\mathcal{L}$ are real numbers.*

Lemma 9 ([40], Lemma 2.1)**.** *Let $\mathcal{A}$ be a linear operator from $\mathbb{C}^{l_1\times l_2}$ to $\mathbb{C}^{l_1\times l_2}$, and let $\lambda_1, \cdots, \lambda_n$ be its eigenvalues, let $\rho(\mathcal{A}) = \max_{1\leq i\leq n}|\lambda_i|$, if $\rho(\mathcal{A}) < 1$, then there exists some constant $c(\delta)$ such that $\|\mathcal{A}^k\| \leq c(\delta)(\rho(\mathcal{A})+\delta)^k$ holds for all integers k, where $0 < \delta < 1 - \rho(\mathcal{A})$.*

Proof of Theorem 2. First, we claim that AM-FIHT is still convergent with a small $\beta \in (0, 1)$. Based on the proof of Theorem 1,

$$\|\widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \leq \nu\|\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F,$$

where $\nu = 6\varepsilon_0, 0 < \varepsilon_0 < 1/10$. A loose bound from a direct derivation with $\beta \neq 0$ is

$$\|\widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \leq (\nu+\beta)\|\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F + \beta\|\widetilde{\boldsymbol{W}}_{l-2} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F.$$

Thus if $\nu + 2\beta < 1$, i.e., $\beta \in (0, \frac{1}{5})$, the iteration is still convergent. Thus for any $\epsilon > 0$, we can always find such an l that $\|\widetilde{\boldsymbol{W}}_{l-2+k} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \leq \epsilon, \forall k \geq 0$. Then following Lemma 7, if we ignore $\widetilde{\boldsymbol{Z}}_{l-1}$, then

$$\widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y} = \mathcal{L}(\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}) + \beta\mathcal{P}_{\widetilde{\mathcal{S}}}(\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\boldsymbol{W}}_{l-2}).$$

Thus $\mathcal{P}_{\widetilde{\mathcal{S}}}(\widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}) = \widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}$. With $\mathcal{P}_{\widetilde{\mathcal{S}}}(\widetilde{\mathcal{G}}\boldsymbol{Y}) = \widetilde{\mathcal{G}}\boldsymbol{Y}$, we have $\mathcal{P}_{\widetilde{\mathcal{S}}}(\widetilde{\boldsymbol{W}}_l) = \widetilde{\boldsymbol{W}}_l$. The update rule of AM-FIHT can be further simplified as

$$\begin{aligned}\begin{bmatrix} \widetilde{\boldsymbol{W}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y} \\ \widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y} \end{bmatrix} =& \begin{bmatrix} \mathcal{L}(\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y}) + \beta(\widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\boldsymbol{W}}_{l-2}) \\ \widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y} \end{bmatrix} \\ :=& \widetilde{\mathcal{L}}\begin{bmatrix} \widetilde{\boldsymbol{W}}_{l-1} - \widetilde{\mathcal{G}}\boldsymbol{Y} \\ \widetilde{\boldsymbol{W}}_{l-2} - \widetilde{\mathcal{G}}\boldsymbol{Y} \end{bmatrix}.\end{aligned}$$

Following Lemma 4, we have $\|\mathcal{L}\| < 1$, if $m > 32\mu c_s r\log(n)$. Based on the definitions of $\rho(\mathcal{L})$ and $\|\mathcal{L}\|$, we have $\rho(\mathcal{L}) \leq \|\mathcal{L}\| < 1$. Our aim is to prove $\rho(\widetilde{\mathcal{L}}) < \rho(\mathcal{L})$. Let λ denote one nonzero eigenvalue of $\widetilde{\mathcal{L}}$, the corresponding eigenmatrix is $\begin{bmatrix}\boldsymbol{M}_1 \\ \boldsymbol{M}_2\end{bmatrix}$, then

$$\widetilde{\mathcal{L}}\begin{bmatrix}\boldsymbol{M}_1 \\ \boldsymbol{M}_2\end{bmatrix} = \begin{bmatrix}\mathcal{L}\boldsymbol{M}_1 + \beta(\boldsymbol{M}_1 - \boldsymbol{M}_2) \\ \boldsymbol{M}_1\end{bmatrix} = \lambda\begin{bmatrix}\boldsymbol{M}_1 \\ \boldsymbol{M}_2\end{bmatrix},$$

we have $\boldsymbol{M}_1 = \lambda\boldsymbol{M}_2, \mathcal{L}(\boldsymbol{M}_1) + \beta\boldsymbol{M}_1 - \beta\boldsymbol{M}_2 = \lambda\boldsymbol{M}_1$. Therefore, $\lambda\mathcal{L}(\boldsymbol{M}_2) + \lambda\beta\boldsymbol{M}_2 - \beta\boldsymbol{M}_2 = \lambda^2\boldsymbol{M}_2$. With $\lambda \neq 0$,

$$\mathcal{L}(\boldsymbol{M}_2) = (\lambda - \beta + \beta/\lambda)\boldsymbol{M}_2 := \eta_i\boldsymbol{M}_2,$$

thus $\boldsymbol{M}_2$ is an eigenmatrix of operator $\mathcal{L}$, with the corresponding eigenvalue as η_i. Lemma 8 shows $\eta_i \in \mathbb{R}$, then we have

$$\begin{aligned}&\lambda^2 - \eta_i\lambda - \beta\lambda + \beta = 0, \\ &\lambda_{i1} = \frac{\eta_i + \beta + \sqrt{(\eta_i+\beta)^2 - 4\beta}}{2}, \\ &\lambda_{i2} = \frac{\eta_i + \beta - \sqrt{(\eta_i+\beta)^2 - 4\beta}}{2}.\end{aligned}$$

Here we analyze in two cases:

1) for any η_i that satisfies $(\eta_i+\beta)^2 - 4\beta \leq 0$, the modulus $|\lambda_{i1}| = |\lambda_{i2}| = \sqrt{\beta}$.
2) for any η_i that satisfies $(\eta_i+\beta)^2 - 4\beta > 0, \eta_i$ cannot be zero for any $\beta \in (0, 1)$. With $\rho(\mathcal{L}) = \max_i|\eta_i| < 1$, it holds that $\eta_i < 1$. In this case, with $\beta \in (0, 1)$, we have

$$\begin{aligned}&(\eta_i+\beta)^2 - 4\beta = (\eta_i - \beta)^2 - 4(1-\eta_i)\beta < (\eta_i - \beta)^2, \\ &\max\{|\lambda_{i1}|, |\lambda_{i2}|\} = \frac{|\eta_i+\beta| + \sqrt{(\eta_i+\beta)^2 - 4\beta}}{2} \\ &< \frac{|\eta_i+\beta| + |\eta_i - \beta|}{2} = \max\{|\eta_i|, \beta\} \leq \max\{|\eta_i|, \sqrt{\beta}\}.\end{aligned}$$

Combining the two cases, if we choose a positive β that satisfies $\beta < \left(\max_i\{|\eta_i|\}\right)^2 = \rho^2(\mathcal{L})$, let

$$q(0) = \rho(\mathcal{L}), q(\beta) = \rho(\widetilde{\mathcal{L}}), \tau = \min\left\{1/5, \rho^2(\mathcal{L})\right\}, \quad (53)$$

then we have $q(0) > q(\beta), \forall\beta \in (0, \tau)$. □

E. Proof of Theorem 3

Lemma 10 derives the properties of $\widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega_{l+1}}\widetilde{\mathcal{G}}^*(\mathcal{P}_{\widetilde{\boldsymbol{U}}} - \mathcal{P}_{\widehat{\boldsymbol{U}}_l})$, and the random operator can be close enough to its mean $\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*(\mathcal{P}_{\widetilde{\boldsymbol{U}}} - \mathcal{P}_{\widehat{\boldsymbol{U}}_l})$ with a significant large amount of observed entries. Lemma 11 illustrates the relation between $\widetilde{\boldsymbol{L}}_l$ and $\widetilde{\boldsymbol{L}}_l'$ and gives the bound on the incoherence of $\widetilde{\boldsymbol{L}}_l'$, which is obtained after the trimming part (line 5 to 10). Lemmas 10 and 11 are built upon Lemmas 9 and 10 in [7] by extending from single-channel signals to multi-channel signals. Similar to the proof of Theorem 1, the proof of Theorem 3 is built upon that of Lemma 3 in [7], which is originally proposed as an initialization strategy. The major steps are devoted to bounding I_5, I_6 and I_7 in (58), and the corresponding results are presented in (59), (60), and (61). We include some details for the completeness of this proof.

Lemma 10. *Let $\widetilde{\boldsymbol{L}}_l' = \widehat{\boldsymbol{U}}_l\widehat{\boldsymbol{\Sigma}}_l\widehat{\boldsymbol{V}}_l^*$ and $\widetilde{\mathcal{G}}\boldsymbol{Y} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^*$ be the SVD of $\widetilde{\boldsymbol{L}}_l'$ and $\widetilde{\mathcal{G}}\boldsymbol{Y}$. Further let $\widehat{\mathcal{S}}_l$ be the tangent subspace of $\widetilde{\boldsymbol{L}}_l'$. Assume there exists a constant μ such that*

$$\left\|\mathcal{P}_{\widehat{\boldsymbol{U}}_l}\widetilde{\boldsymbol{H}}_{k,t}\right\|_F^2 \leq \frac{\mu c_s r}{n_c n}, \quad \left\|\mathcal{P}_{\widehat{\boldsymbol{V}}_l}\widetilde{\boldsymbol{H}}_{k,t}\right\|_F^2 \leq \frac{\mu c_s r}{n_c n},$$

and

$$\left\|\mathcal{P}_{\widetilde{\boldsymbol{U}}}\widetilde{\boldsymbol{H}}_{k,t}\right\|_F^2 \leq \frac{\mu c_s r}{n_c n}, \quad \left\|\mathcal{P}_{\widetilde{\boldsymbol{V}}}\widetilde{\boldsymbol{H}}_{k,t}\right\|_F^2 \leq \frac{\mu c_s r}{n_c n}.$$

for all $1 \le t \le n, 1 \le k \le n_c$. *Let* $\Omega_{l+1} = \{(k_a, t_a)|a = 1, \cdots, \widehat{m}\}$ *be a set of indices sampled with replacement. If* $\mathcal{P}_{\Omega_{l+1}}$ *is independent of* $\widetilde{\boldsymbol{U}}$, $\widetilde{\boldsymbol{V}}$, $\widehat{\boldsymbol{U}}_l$ *and* $\widehat{\boldsymbol{V}}_l$, *then*

$$\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\mathcal{I} - \widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}})\widetilde{\mathcal{G}}^*(\mathcal{P}_{\widetilde{\boldsymbol{U}}} - \mathcal{P}_{\widehat{\boldsymbol{U}}_l})\right\| \le \sqrt{\frac{160\mu c_s r\log(n)}{\widehat{m}}}$$

with probability at least $1 - n_c n^{-2}$, *if* $\widehat{m} \ge \frac{125}{18}\mu c_s r\log(n)$.

Lemma 11. *Let* $\widetilde{\boldsymbol{L}}_l = \widetilde{\boldsymbol{U}}_l\widetilde{\boldsymbol{\Sigma}}_l\widetilde{\boldsymbol{V}}_l^*$ *and* $\widetilde{\mathcal{G}}\boldsymbol{Y} = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^*$ *be the SVD of* $\widetilde{\boldsymbol{L}}_l$ *and* $\widetilde{\mathcal{G}}\boldsymbol{Y}$. *Assume*

$$\max_{k_1}||\widetilde{\boldsymbol{U}}_{k_1*}||^2 \le \frac{\mu c_s r}{n_c n} \text{ and } \max_{k_2}||\widetilde{\boldsymbol{V}}_{k_2*}||^2 \le \frac{\mu c_s r}{n_c n}.$$

Suppose $\widetilde{\boldsymbol{L}}_l$ *and* $\widetilde{\mathcal{G}}\boldsymbol{Y}$ *are both rank-*r *matrices satisfying*

$$\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le \frac{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{10\sqrt{2}}.$$

Then the matrix $\widetilde{\boldsymbol{L}}_l' = \widehat{\boldsymbol{U}}_l\widehat{\boldsymbol{\Sigma}}_l\widehat{\boldsymbol{V}}_l^*$, *denoting the SVD of* $\widetilde{\boldsymbol{L}}_l'$, *that is obtained after trimming in RAM-FIHT satisfies*

$$\|\widetilde{\boldsymbol{L}}' - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le 8\kappa\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F, \tag{54}$$

$$\max_{k_1,k_2}\left\{\|\widehat{\boldsymbol{U}}_{k_1*}\|^2, \|\widehat{\boldsymbol{V}}_{k_2*}\|^2\right\} \le \frac{100\mu c_s r}{81 n_c n}, \tag{55}$$

where κ *denotes the condition number of* $\widetilde{\mathcal{G}}\boldsymbol{Y}$.

Proof of Theorem 3. First, we show the following inequality holds with high probability by mathematical induction.

$$\|\widetilde{\boldsymbol{L}}_k - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le \frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{128\kappa^2}. \tag{56}$$

Inductive Step: Suppose (56) holds when $k = l$ and $l \ge 0$. Then (55) in Lemma 11 holds. Further, we can conclude

$$\left\|\mathcal{P}_{\widetilde{\boldsymbol{U}}_l}\widetilde{\boldsymbol{H}}_{k,t}\right\|_F^2 \le \frac{100\mu c_s r}{81 n_c n} \text{ and } \left\|\mathcal{P}_{\widehat{\boldsymbol{V}}_l}\widetilde{\boldsymbol{H}}_{k,t}\right\|_F^2 \le \frac{100\mu c_s r}{81 n_c n}. \tag{57}$$

Recall that $\boldsymbol{Y} = \widetilde{\mathcal{D}}\boldsymbol{X}$ and $\widetilde{\mathcal{G}}\boldsymbol{Y} = \widetilde{\mathcal{H}}\mathbf{X}$. Define $\widehat{\boldsymbol{Y}}_l = \widetilde{\mathcal{D}}\widehat{\boldsymbol{X}}_l$. Since the measurements are noiseless, then $\boldsymbol{M} = \boldsymbol{X}$ and

$$\widetilde{\mathcal{H}}(\widehat{\boldsymbol{X}}_l + \widehat{p}^{-1}\widetilde{\mathcal{P}}_{\Omega_{l+1}}(\boldsymbol{X} - \widehat{\boldsymbol{X}}_l)) = \widetilde{\mathcal{G}}(\widehat{\boldsymbol{Y}}_l + \widehat{p}^{-1}\widetilde{\mathcal{P}}_{\Omega_{l+1}}(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_l)).$$

Then,

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{L}}_{l+1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\right\|_F \le 2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\widehat{\boldsymbol{Y}}_l + \widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}}(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_l)) - \widetilde{\mathcal{G}}\boldsymbol{Y}\right\|_F \\
\le& 2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\boldsymbol{Y} - \widetilde{\mathcal{G}}\boldsymbol{Y}\right\|_F + 2\left\|(\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}} - \widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega_{l+1}})(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}_l)\right\|_F \\
=& 2\left\|(\mathcal{I} - \mathcal{P}_{\widehat{\mathcal{S}}_l})\widetilde{\mathcal{G}}\boldsymbol{Y}\right\|_F \\
&+ 2\left\|(\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^* - \widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega_{l+1}}\widetilde{\mathcal{G}}^*)(\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
\le& 2\left\|(\mathcal{I} - \mathcal{P}_{\widehat{\mathcal{S}}_l})(\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
&+ 2\left\|(\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widehat{\mathcal{S}}_l} - \widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega_{l+1}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widehat{\mathcal{S}}_l})(\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
&+ 2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\mathcal{I} - \widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}})\widetilde{\mathcal{G}}^*(\mathcal{I} - \mathcal{P}_{\widehat{\mathcal{S}}_l})(\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
:=& I_5 + I_6 + I_7.
\end{aligned} \tag{58}$$

where the first inequality comes from (39).
With (54) and (56), I_5 can be bounded as

$$I_5 \le \frac{2\|\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F^2}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})} \le \varepsilon_0\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F, \tag{59}$$

As for the item I_6, Lemma 4 along with (54) suggests

$$\begin{aligned}
I_6 &\le 2\sqrt{\frac{3200\mu c_s r\log(n)}{81\widehat{m}}}\|\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \\
&\le 16\kappa\sqrt{\frac{3200\mu c_s r\log(n)}{81\widehat{m}}}\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F
\end{aligned} \tag{60}$$

with probability at least $1 - n_c n^{-2}$. To bound I_7,

$$\begin{aligned}
&\Big(\mathcal{I} - \mathcal{P}_{\widehat{\mathcal{S}}_l}\Big)\Big(\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y}\Big) = \Big(\boldsymbol{I} - \widehat{\boldsymbol{U}}_l\widehat{\boldsymbol{U}}_l^*\Big)(-\widetilde{\mathcal{G}}\boldsymbol{Y})\Big(\boldsymbol{I} - \widehat{\boldsymbol{V}}_l\widehat{\boldsymbol{V}}_l^*\Big) \\
=&\Big(\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^* - \widehat{\boldsymbol{U}}_l\widehat{\boldsymbol{U}}_l^*\Big)\Big(\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y}\Big)\Big(\boldsymbol{I} - \widehat{\boldsymbol{V}}_l\widehat{\boldsymbol{V}}_l^*\Big) \\
=&\Big(\mathcal{P}_{\widetilde{\boldsymbol{U}}} - \mathcal{P}_{\widehat{\boldsymbol{U}}_l}\Big)(\mathcal{I} - \mathcal{P}_{\widetilde{\boldsymbol{V}}})\Big(\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y}\Big).
\end{aligned}$$

Hence, by Lemma 10, with probability at least $1 - n_c n^{-2}$,

$$\begin{aligned}
I_7 &= 2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\mathcal{I} - \widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}})\widetilde{\mathcal{G}}^*\Big(\mathcal{P}_{\widetilde{\boldsymbol{U}}} - \mathcal{P}_{\widehat{\boldsymbol{U}}_l}\Big)(\mathcal{I} - \mathcal{P}_{\widetilde{\boldsymbol{V}}})\Big(\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y}\Big)\right\|_F \\
&\le 2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\mathcal{I} - \widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}})\widetilde{\mathcal{G}}^*\Big(\mathcal{P}_{\widetilde{\boldsymbol{U}}} - \mathcal{P}_{\widehat{\boldsymbol{U}}_l}\Big)\right\|\left\|\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y}\right\|_F \\
&\le 16\kappa\sqrt{\frac{16000\mu c_s r\log(n)}{81\widehat{m}}}\left\|\widetilde{\boldsymbol{L}}_l' - \widetilde{\mathcal{G}}\boldsymbol{Y}\right\|_F.
\end{aligned} \tag{61}$$

Therefore, if $\widehat{m} \ge C_4\varepsilon_0^{-2}\mu c_s\kappa^2 r\log(n)$ for some constant C_4,

$$I_6 + I_7 \le 326\kappa\sqrt{\frac{\mu c_s r\log(n)}{\widehat{m}}}\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le \varepsilon_0\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F.$$

Putting pieces together gives

$$\|\widetilde{\boldsymbol{L}}_{l+1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le 2\varepsilon_0\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \tag{62}$$

with probability at least $1 - 2n_c n^{-2}$. Hence, (56) also holds when $k = l + 1$.

Base Case: Since $\widetilde{\boldsymbol{L}}_0 = \mathcal{Q}_r\big(\widehat{p}^{-1}\widetilde{\mathcal{H}}\mathcal{P}_{\Omega_0}(\boldsymbol{X})\big)$, we can follow the same idea in the proof of base case in Theorem 1. Thus, when $k = 0$, (56) is valid with probability at least $1 - n_c n^{-2}$ provided $\widehat{m} \ge C_5\varepsilon_0^{-2}\mu c_s\kappa^6 r^2\log(n)$ for some constant C_5.

Let $C_2 = \max\{C_4, C_5\}$. If $\widehat{m} \ge C_2\varepsilon_0^{-2}\mu c_s\kappa^6 r^2\log(n)$, then for each $l \ge 0$, we have

$$\|\widetilde{\boldsymbol{L}}_{l+1} - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le 2\varepsilon_0\|\widetilde{\boldsymbol{L}}_l - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F.$$

with probability at least $1 - 2n_c n^{-2}$. Directly the following inequality is obtained with probability $1 - (2L+1)n_c n^{-2}$,

$$\|\widetilde{\boldsymbol{L}}_L - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le \nu^L\|\widetilde{\boldsymbol{L}}_0 - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le \nu^L\frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{128\kappa^2}. \tag{63}$$

If we take $L = \left\lceil\varepsilon_0^{-1}\log\left(\frac{\sigma_{\max}(\mathcal{H}\boldsymbol{X})}{128\kappa^3\varepsilon}\right)\right\rceil$ with an arbitrarily small positive constant ε, since $\sigma_{\max}(\widetilde{\mathcal{G}}\boldsymbol{Y}) = \sqrt{n_c}\sigma_{\max}(\mathcal{H}\boldsymbol{X})$,

$$\|\widetilde{\boldsymbol{L}}_L - \widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le n_c^{1/2}\varepsilon, \tag{64}$$

which completes the proof of Theorem 3. □

F. Proof of Theorem 4

The proof of Theorem 4 is similar to Theorem 3, except some modification to handle the noise matrix $\boldsymbol{N}$. We first present the following lemma that will be useful in the proof.

Lemma 12. *Suppose* $m \ge 16\log(n)$, *then*

$$\left\|p^{-1}\widetilde{\mathcal{H}}\mathcal{P}_\Omega(\boldsymbol{N}) - \widetilde{\mathcal{H}}\boldsymbol{N}\right\| \le \sqrt{\frac{16\log(n)}{m}}n_c n\left\|\widetilde{\mathcal{H}}\boldsymbol{N}\right\|_\infty$$

with probability at least $1 - n_c n^{-2}$.

Proof of Theorem 4. For the noisy case where $\boldsymbol{M}=\boldsymbol{X}+\boldsymbol{N}$, we have assumed $\|\boldsymbol{N}\|_\infty \le \dfrac{\varepsilon_0\|\mathcal{H}\boldsymbol{X}\|}{2048\kappa^3 r^{1/2} n_c^{1/2} n}$. Recall (29), define $\boldsymbol{S}=\widetilde{\mathcal{D}}\boldsymbol{N}$, then

$$\|\widetilde{\mathcal{G}}\boldsymbol{S}\|_\infty = \|\widetilde{\mathcal{H}}\boldsymbol{N}\|_\infty = \|\boldsymbol{N}\|_\infty \le \frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{2048\kappa^2 r^{1/2} n_c n}.$$

Similar to the derivation of (39), we have

$$\begin{aligned}
&\|\widetilde{\boldsymbol{L}}_{l+1}-\widetilde{\mathcal{G}}\boldsymbol{Y}\| \le 2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\widehat{\boldsymbol{Y}}_l+\widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}}(\boldsymbol{Y}+\boldsymbol{S}-\widehat{\boldsymbol{Y}}_l))-\widetilde{\mathcal{G}}\boldsymbol{Y}\right\| \\
\le& 2\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\widehat{\boldsymbol{Y}}_l+\widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}}(\boldsymbol{Y}-\widehat{\boldsymbol{Y}}_l))-\widetilde{\mathcal{G}}\boldsymbol{Y}\right\| \\
&+2\|\widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega_{l+1}}(\boldsymbol{S})\|. \\
&\|\widetilde{\boldsymbol{L}}_{l+1}-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le \sqrt{2r}\|\widetilde{\boldsymbol{L}}_{l+1}-\widetilde{\mathcal{G}}\boldsymbol{Y}\| \\
\le& 2\sqrt{2r}\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\widehat{\boldsymbol{Y}}_l+\widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}}(\boldsymbol{Y}+\boldsymbol{S}-\widehat{\boldsymbol{Y}}_l))-\widetilde{\mathcal{G}}\boldsymbol{Y}\right\| \\
\le& 2\sqrt{2r}\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\widehat{\boldsymbol{Y}}_l+\widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}}(\boldsymbol{Y}-\widehat{\boldsymbol{Y}}_l))-\widetilde{\mathcal{G}}\boldsymbol{Y}\right\| \\
&+2\sqrt{2r}\|\widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega_{l+1}}(\boldsymbol{S})\| \\
\le& 2\sqrt{2r}\left\|(\mathcal{I}-\mathcal{P}_{\widehat{\mathcal{S}}_l})(\widetilde{\boldsymbol{L}}_l'-\widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
&+2\sqrt{2r}\left\|(\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widehat{\mathcal{S}}_l}-\widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega_{l+1}}\widetilde{\mathcal{G}}^*\mathcal{P}_{\widehat{\mathcal{S}}_l})(\widetilde{\boldsymbol{L}}_l'-\widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
&+2\sqrt{2r}\left\|\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}(\mathcal{I}-\widehat{p}^{-1}\mathcal{P}_{\Omega_{l+1}})\widetilde{\mathcal{G}}^*(\mathcal{I}-\mathcal{P}_{\widehat{\mathcal{S}}_l})(\widetilde{\boldsymbol{L}}_l'-\widetilde{\mathcal{G}}\boldsymbol{Y})\right\|_F \\
&+2\sqrt{2r}\|\widehat{p}^{-1}\mathcal{P}_{\widehat{\mathcal{S}}_l}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega_{l+1}}(\boldsymbol{S})\| \\
:=& \sqrt{2r}(I_5+I_6+I_7)+I_9,
\end{aligned} \tag{65}$$

where $I_5+I_6+I_7$ has been defined in (58).
Similar to the proof of Theorem 3, we show that the following inequality holds with high probability by induction.

$$\|\widetilde{\boldsymbol{L}}_k-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le \frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{128\sqrt{2}\kappa^2 r^{1/2}}. \tag{66}$$

Inductive Step: Suppose (66) holds when $k=l$ and $l\ge 0$. By Lemma 5 and (56), we have

$$\begin{aligned}
\sqrt{2r}I_5 \le& \frac{2\sqrt{2r}\|\widetilde{\boldsymbol{L}}_l'-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F^2}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})} \le \frac{128\sqrt{2}\kappa^2\sqrt{r}\|\widetilde{\boldsymbol{L}}_l-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F^2}{\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})} \\
\le& \varepsilon_0\|\widetilde{\boldsymbol{L}}_l-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F.
\end{aligned}$$

Similar to I_6 and I_7,

$$\sqrt{2r}(I_6+I_7) \le 326\kappa\sqrt{\frac{2\mu c_s r^2\log(n)}{\widehat{m}}}\|\widetilde{\boldsymbol{L}}_l-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F.$$

Hence, if $\widehat{m}\ge C_6\varepsilon_0^{-2}\mu c_s\kappa^2 r^2\log(n)$ for some constant C_6,

$$\sqrt{2r}(I_5+I_6+I_7) \le 2\varepsilon_0\|\widetilde{\boldsymbol{L}}_l-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F$$

with probability at least $1-2n_cn^{-2}$. On the other hand, if $m\ge 256r\log(n)$, then with probability at least $1-n_cn^{-2}$

$$\begin{aligned}
I_9 \le& 2\sqrt{2r}\|\widehat{p}^{-1}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega_{l+1}}(\boldsymbol{S})\| \\
\le& 2\sqrt{2r}\|\widehat{p}^{-1}\widetilde{\mathcal{G}}\mathcal{P}_{\Omega_{l+1}}(\boldsymbol{S})-\widetilde{\mathcal{G}}\boldsymbol{S}\|+2\sqrt{2r}\|\widetilde{\mathcal{G}}\boldsymbol{S}\| \\
\le& 8\sqrt{2}\sqrt{\frac{r\log(n)}{m}}n_cn\|\widetilde{\mathcal{G}}\boldsymbol{S}\|_\infty+2\sqrt{2r}n_cn\|\widetilde{\mathcal{G}}\boldsymbol{S}\|_\infty \\
\le& \frac{1}{16\sqrt{2}}\sqrt{\frac{r\log(n)}{m}}\frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{\kappa^2 r^{1/2}}+\frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{512\sqrt{2}\kappa^2 r^{1/2}} \\
\le& \frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{256\sqrt{2}\kappa^2 r^{1/2}},
\end{aligned} \tag{67}$$

where the second last inequality comes from Lemma 12. Following $\nu=2\varepsilon_0\le 1/2$ and (66), with probability at least $1-3n_cn^{-2}$, we can bound $\|\widetilde{\boldsymbol{L}}_{l+1}-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F$ by

$$\frac{1}{2}\|\widetilde{\boldsymbol{L}}_l-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F+\frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{256\sqrt{2}\kappa^2 r^{1/2}} \le \frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{128\sqrt{2}\kappa^2 r^{1/2}}.$$

Hence, (66) also holds when $k=l+1$.
Base Case: Since $\widetilde{\boldsymbol{L}}_0=\mathcal{Q}_r\big(\widehat{p}^{-1}\widetilde{\mathcal{H}}\mathcal{P}_{\Omega_0}(\boldsymbol{X}+\boldsymbol{N})\big)$, then with probability at least $1-n_cn^{-2}$,

$$\begin{aligned}
&\|\widetilde{\boldsymbol{L}}_0-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le \sqrt{2r}\|\widetilde{\boldsymbol{L}}_0-\widetilde{\mathcal{G}}\boldsymbol{Y}\| \\
\le& \sqrt{2r}\left\|p^{-1}\widetilde{\mathcal{G}}\mathcal{P}_\Omega(\boldsymbol{Y}+\boldsymbol{S})-\widetilde{\boldsymbol{L}}_0\right\|+\sqrt{2r}\left\|p^{-1}\widetilde{\mathcal{G}}\mathcal{P}_\Omega(\boldsymbol{Y}+\boldsymbol{S})-\widetilde{\mathcal{G}}\boldsymbol{Y}\right\| \\
\le& 2\sqrt{2r}\left\|p^{-1}\widetilde{\mathcal{G}}\mathcal{P}_\Omega(\boldsymbol{Y})-\widetilde{\mathcal{G}}\boldsymbol{Y}\right\|+2\sqrt{2r}\left\|p^{-1}\widetilde{\mathcal{G}}\mathcal{P}_\Omega(\boldsymbol{S})\right\| \\
\le& \sqrt{\frac{512\mu c_s r^2\log(n)}{m}}\|\widetilde{\mathcal{G}}\boldsymbol{Y}\|+\frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{256\sqrt{2}\kappa^2\sqrt{r}}.
\end{aligned}$$

where the last inequality comes from (67) and Lemma 6. To guarantee that (66) holds with $k=0$, we need

$$\sqrt{\frac{512\mu c_s r^2\log(n)}{m}}\|\widetilde{\mathcal{G}}\boldsymbol{Y}\| \le \frac{\varepsilon_0\sigma_{\min}(\widetilde{\mathcal{G}}\boldsymbol{Y})}{256\sqrt{2}\kappa^2\sqrt{r}}. \tag{68}$$

That is $\widehat{m}\ge C_7\varepsilon_0^{-2}\mu c_s\kappa^6 r^3\log(n)$ for some constant C_7.

Let $C_3=\max\{C_6,C_7\}$, if $\widehat{m}\ge C_3\varepsilon_0^{-2}\mu c_s\kappa^6 r^3\log(n)$, for each $l\ge 0$, with probability $1-2n_cn^{-2}$, we have

$$\|\widetilde{\boldsymbol{L}}_{l+1}-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le 2\varepsilon_0\|\widetilde{\boldsymbol{L}}_l-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F+\Delta. \tag{69}$$

where $\Delta=32\sqrt{2}n_cn\|\widetilde{\mathcal{G}}\boldsymbol{S}\|_\infty+2\sqrt{2}r^{1/2}\|\widetilde{\mathcal{G}}\boldsymbol{S}\|$. Then

$$\|\widetilde{\boldsymbol{L}}_{l+1}-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F-\frac{\Delta}{1-\nu} \le \nu\Big(\|\widetilde{\boldsymbol{L}}_l-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F-\frac{\Delta}{1-\nu}\Big).$$

Therefore, with probability $1-(3L+1)n_cn^{-2}$,

$$\|\widetilde{\boldsymbol{L}}_L-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le \nu^L\|\widetilde{\boldsymbol{L}}_0-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F+\frac{\Delta}{1-\nu}. \tag{70}$$

Similar to (64), take $L=\left\lceil\varepsilon_0^{-1}\log\left(\frac{\sigma_{\max}(\mathcal{H}\boldsymbol{X})}{128\kappa^3\varepsilon}\right)\right\rceil$ with an arbitrarily small positive constant ε, since $\nu\le 1/2$,

$$\begin{aligned}
\|\widetilde{\boldsymbol{L}}_L-\widetilde{\mathcal{G}}\boldsymbol{Y}\|_F \le& n_c^{1/2}\varepsilon+64\sqrt{2}n_cn\|\widetilde{\mathcal{G}}\boldsymbol{S}\|_\infty+4\sqrt{2}r^{1/2}\|\widetilde{\mathcal{G}}\boldsymbol{S}\| \\
\le& n_c^{1/2}\varepsilon+128n_cn\|\widetilde{\mathcal{G}}\boldsymbol{S}\|_\infty+8r^{1/2}\|\widetilde{\mathcal{G}}\boldsymbol{S}\|.
\end{aligned}$$

which completes the proof of Theorem 4. □

G. Proof of Theorem 5

We first introduce some useful lemmas.

Lemma 13 ([23], Corollary 7.7.4(a)). *If $\boldsymbol{A},\boldsymbol{B}\in\mathbb{C}^n$ are positive-definite, then $\boldsymbol{A}\succeq\boldsymbol{B}$ if and only if $\boldsymbol{B}^{-1}\succeq\boldsymbol{A}^{-1}$*

Lemma 14 ([32], Theorem 7). *Let $\lambda_1\ge\cdots\ge\lambda_n$ be eigenvalues of $\boldsymbol{A}$, denoted by $\lambda_i(\boldsymbol{A})=\lambda_i$. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be Hermitian positive semi-definite $n\times n$ matrices. If $1\le k\le i\le n$ and $1\le l\le n-i+1$, then*

$$\lambda_{i+l-1}(\boldsymbol{A})\lambda_{n-l+1}(\boldsymbol{B}) \le \lambda_i(\boldsymbol{AB}) \le \lambda_{i-k+1}(\boldsymbol{A})\lambda_k(\boldsymbol{B}).$$

In particular,

$$\lambda_n(\boldsymbol{A})\lambda_n(\boldsymbol{B}) \le \lambda_n(\boldsymbol{AB}),\quad \lambda_1(\boldsymbol{AB}) \le \lambda_1(\boldsymbol{A})\lambda_1(\boldsymbol{B}).$$

Proof of Theorem 5. Consider $n_c=1$, the definition of $\mathcal{H}$ can be extended to a row vector, which corresponds with one

channel data. Let $\mathcal{H}\boldsymbol{X}_{k*} = \boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k$ and $\mathcal{H}\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$ be the SVD of $\mathcal{H}\boldsymbol{X}_{k*}$ and $\mathcal{H}\boldsymbol{X}$. Then, μ_0 is defined as

$$\max_{k_1} \left\|\boldsymbol{e}_{k_1}^*\boldsymbol{U}_k\right\|^2 \le \frac{\mu_0 r}{n_1}, \quad \max_{k_2} \left\|\boldsymbol{e}_{k_2}^*\boldsymbol{V}_k\right\|^2 \le \frac{\mu_0 r}{n_2}.$$

Notice that all $\mathcal{H}\boldsymbol{X}_{k*}$ share the same column space and row space, they have the same incoherence μ_0. It is trivial $\mu = \mu_0$ if we consider the incoherence of row spaces. Hence, we only focus on the incoherence of column space.

By (6), $\mathcal{H}\boldsymbol{X} = \boldsymbol{P}_L\boldsymbol{\Gamma}\boldsymbol{P}_R^T$. Define a series of diagonal matrices $\boldsymbol{D}_k$ as $\boldsymbol{D}_k = \text{diag}(\boldsymbol{d}_k)$ with $1 \le k \le n_c$, and $\boldsymbol{d}_k = [d_{k,1}, \cdots, d_{k,r}]$, where $d_{k,i} = \boldsymbol{r}_i^*\boldsymbol{s}_1\boldsymbol{C}_{k*}\boldsymbol{l}_i$. We need one mild assumption that $d_{k,i} \neq 0$. It guarantees that each $\boldsymbol{D}_k$ is full rank. Thus, $\mathcal{H}\boldsymbol{X}_{k*} = \boldsymbol{E}_L\boldsymbol{D}_k\boldsymbol{P}_R$, where $\boldsymbol{E}_L = \boldsymbol{P}_L$ with $n_c = 1$. There exists a row-switching matrix $\boldsymbol{Q}_1$ satisfying

$$\boldsymbol{Q}_1(\mathcal{H}\boldsymbol{X}) = \begin{bmatrix} \mathcal{H}\boldsymbol{X}_{1*} \\ \mathcal{H}\boldsymbol{X}_{2*} \\ \vdots \\ \mathcal{H}\boldsymbol{X}_{n_c*} \end{bmatrix} = \begin{bmatrix} \boldsymbol{E}_L\boldsymbol{D}_1 \\ \boldsymbol{E}_L\boldsymbol{D}_2 \\ \vdots \\ \boldsymbol{E}_L\boldsymbol{D}_{n_c} \end{bmatrix} \boldsymbol{P}_R := \widetilde{\boldsymbol{E}}_L\boldsymbol{P}_R.$$

Define a mapping $f : \{1, 2, \cdots, n_c n_1\} \mapsto \{1, 2, \cdots, n_c n_1\}$, $f(z) = w$ with $\boldsymbol{e}_z = \boldsymbol{Q}_1\boldsymbol{e}_w$, then f is a bijective mapping. Hence, we have

$$\begin{aligned} \max_{l_1} \|\boldsymbol{e}_{l_1}^*\boldsymbol{U}\|^2 &= \max_{k_1} (\boldsymbol{Q}_1\boldsymbol{e}_{k_1})^* \widetilde{\boldsymbol{E}}_L (\widetilde{\boldsymbol{E}}_L^*\widetilde{\boldsymbol{E}}_L)^{-1} \widetilde{\boldsymbol{E}}_L^* (\boldsymbol{Q}_1\boldsymbol{e}_{k_1}) \\ &= \max_{k_1} \boldsymbol{e}_{k_1}^*\widetilde{\boldsymbol{E}}_L \left(\sum_{k=1}^{n_c} \boldsymbol{D}_k^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_k\right)^{-1} \widetilde{\boldsymbol{E}}_L^*\boldsymbol{e}_{k_1}. \end{aligned} \tag{71}$$

Consider $1 \le k_1 \le n_1$, we know that $\boldsymbol{e}_{k_1}^*\widetilde{\boldsymbol{E}}_L = \widehat{\boldsymbol{e}}_{k_1}^*\boldsymbol{E}_L\boldsymbol{D}_1$, where $\boldsymbol{e}_{k_1} \in \mathbb{C}^{n_c n_1}$ and $\widehat{\boldsymbol{e}}_{k_1} \in \mathbb{C}^{n_1}$ are both coordinate vectors. Additionally, it is easy to show that symmetric matrices $\{\boldsymbol{D}_k^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_k\}_{k=1}^{n_c}$ are positive definite since $\{\boldsymbol{E}_L\boldsymbol{D}_k\}_{k=1}^{n_c}$ are full rank. Also, following Lemma 13, we have

$$\begin{aligned} \sum_{k=1}^{n_c} \boldsymbol{D}_k^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_k &\succ \boldsymbol{D}_1^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_1 \succ 0, \\ \Big(\sum_{k=1}^{n_c} \boldsymbol{D}_k^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_k\Big)^{-1} &\prec (\boldsymbol{D}_1^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_1)^{-1}. \end{aligned} \tag{72}$$

Then,

$$\begin{aligned} \|\boldsymbol{e}_{k_1}^*\boldsymbol{U}\|^2 &= \widehat{\boldsymbol{e}}_{k_1}^*\boldsymbol{E}_L\boldsymbol{D}_1\Big(\sum_{k=1}^{n_c} \boldsymbol{D}_k^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_k\Big)^{-1} \boldsymbol{D}_1^*\boldsymbol{E}_L^*\widehat{\boldsymbol{e}}_{k_1} \\ &< \widehat{\boldsymbol{e}}_{k_1}^*\boldsymbol{E}_L\boldsymbol{D}_1(\boldsymbol{D}_1^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_1)^{-1}\boldsymbol{D}_1^*\boldsymbol{E}_L^*\widehat{\boldsymbol{e}}_{k_1} \\ &= \widehat{\boldsymbol{e}}_{k_1}^*\boldsymbol{E}_L(\boldsymbol{E}_L^*\boldsymbol{E}_L)^{-1}\boldsymbol{E}_L^*\widehat{\boldsymbol{e}}_{k_1} \le \frac{u_0 r}{n_1} = \frac{(n_c u_0) r}{n_c n_1}. \end{aligned}$$

Similarly, we can prove $\left\|\boldsymbol{e}_{k_1}^*\boldsymbol{U}\right\|^2 < \dfrac{(n_c u_0) r}{n_c n_1}$ for all i satisfying $1 \le i \le n_c n_1$, which leads to (25).

Moreover, we can provide a tighter bound on μ with a stronger assumption. Suppose there exists a $\widehat{d} \in \mathbb{C}$ and a real number $\delta \in (0, 1)$ satisfying $(1-\delta)|\widehat{d}| \le |d_{k,i}| \le (1+\delta)|\widehat{d}|$. By Lemma 14, define $\kappa_L = \frac{\sigma_{\max}(\boldsymbol{E}_L)}{\sigma_{\min}(\boldsymbol{E}_L)}$, then:

$$\begin{aligned} &\lambda_{\max}(\boldsymbol{D}_1^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_1) = \lambda_{\max}(\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_1\boldsymbol{D}_1^*) \\ &\le \lambda_{\max}(\boldsymbol{E}_L^*\boldsymbol{E}_L)\lambda_{\max}(\boldsymbol{D}_1\boldsymbol{D}_1^*) \\ &\le \frac{\kappa_L^2(1+\delta)^2}{(1-\delta)^2}\lambda_{\min}(\boldsymbol{E}_L^*\boldsymbol{E}_L)\lambda_{\min}(\boldsymbol{D}_k\boldsymbol{D}_k^*) \\ &\le \frac{\kappa_L^2(1+\delta)^2}{(1-\delta)^2}\lambda_{\min}(\boldsymbol{D}_k^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_k). \end{aligned}$$

since even the minimum eigenvalue of $\boldsymbol{D}_k^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_k$ is larger than the maximum one of $\frac{(1-\delta)^2}{\kappa_L^2(1+\delta)^2}\boldsymbol{D}_1^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_1$, we have

$$\sum_{k=1}^{n_c n} \boldsymbol{D}_k^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_k \succeq [1 + (n_c - 1)\frac{(1-\delta)^2}{\kappa_L^2(1+\delta)^2}]\boldsymbol{D}_1^*\boldsymbol{E}_L^*\boldsymbol{E}_L\boldsymbol{D}_1.$$

Similarly, we can establish the following relation between μ and μ_0,

$$\mu \le \frac{n_c\mu_0}{1 + (n_c - 1)\frac{(1-\delta)^2}{\kappa_L^2(1+\delta)^2}}.$$

□

Shuai Zhang received the B.E. degree from University of Science and Technology of China, Hefei, China, in 2016.

He is pursuing the Ph.D. degree in electrical engineering at Rensselaer Polytechnic Institute, Troy, NY. His research interests include signal processing and high dimensional data analysis.

Yingshuai Hao (S'14) received the B.E. degree from Shandong University, Jinan, China, in 2011 and the M.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014.

He is pursuing the Ph.D. degree in electrical engineering at Rensselaer Polytechnic Institute, Troy, NY. His research interests include cyber security of power systems, and PMU data quality improvement.

Meng Wang (M'12) received the Ph.D. degree from Cornell University, Ithaca, NY, USA, in 2012.

She is an Assistant Professor in the department of Electrical, Computer, and Systems Engineering at Rensselaer Polytechnic Institute. Her research interests include high dimensional data analysis and their applications in power systems monitoring and network inference.

Joe H. Chow (F'92) received the M.S. and Ph.D. degrees from the University of Illinois, Urbana-Champaign, Urbana, IL, USA.

After working in the General Electric power system business in Schenectady, NY, USA, he joined Rensselaer Polytechnic Institute, Troy, NY, USA, in 1987, where he is a Professor of Electrical, Computer, and Systems Engineering. His research interests include multivariable control, power system dynamics and control, FACTS controllers, and synchronized phasor data.