Support Recovery for Sparse Signals With Unknown Non-Stationary Modulation

Youye Xie, *Student Member, IEEE*, Michael B. Wakin, *Senior Member, IEEE*, and Gongguo Tang, *Member, IEEE*

***Abstract*—The problem of estimating a sparse signal from low dimensional noisy observations arises in many applications, including super resolution, signal deconvolution, and radar imaging. In this paper, we consider a sparse signal model with non-stationary modulations, in which each dictionary atom contributing to the observations undergoes an unknown, distinct modulation. By applying the lifting technique, under the assumption that the modulating signals live in a common subspace, we recast this sparse recovery and non-stationary blind demodulation problem as the recovery of a column-wise sparse matrix from structured linear observations, and propose to solve it via block ℓ_1-norm regularized quadratic minimization. Due to observation noise, the sparse signal and modulation process cannot be recovered exactly. Instead, we aim to recover the sparse support of the ground truth signal and bound the recovery errors of the signal's non-zero components and the modulation process. In particular, we derive sufficient conditions on the sample complexity and regularization parameter for exact support recovery and bound the recovery error on the support. Numerical simulations verify and support our theoretical findings, and we demonstrate the effectiveness of our model in the application of single molecule imaging.**

***Index Terms*—Support recovery, blind demodulation, sparse matrix recovery, group lasso, compressive sensing.**

I. Introduction

A. Overview

THE problem of recovering a high dimensional sparse signal from its low dimension observations using a fixed sensing mechanism arises naturally in a wide range of applications, including radar autofocus [2], magnetic resonance imaging [3], and video acquisition [4]. Typically, the system receives a low dimensional signal $\boldsymbol{y} = \mathbf{D}\mathbf{A}\boldsymbol{c} \in \mathbf{C}^N$, where $\boldsymbol{c} \in \mathbf{C}^M$ $(M > N)$ is an unknown high dimensional signal, and $\mathbf{D}$ and $\mathbf{A}$ are known sensing matrices. Although the sensing process is underdetermined, one can solve for $\boldsymbol{c}$ by leveraging its sparsity; this sparse recovery problem has been studied extensively by the compressive sensing community [5]–[7].

When $\mathbf{D} \in \mathbf{C}^{N\times N}$ is a diagonal matrix containing a sampled carrier signal along its diagonal, it describes a modulation process, and thus recovery with unknown diagonal $\mathbf{D}$ is sometimes referred to as simultaneous sparse recovery and blind demodulation [8]. Scenarios where $\mathbf{D}$ is unknown arise in certain self-calibration [9] and blind deconvolution problems [10].

In this paper, we further generalize this model, allowing each atom in the dictionary matrix $\mathbf{A}$ to undergo a distinct modulation process, rather than multiplication by the same matrix $\mathbf{D}$. We refer to this generalized scenario as *non-stationary modulation*. Moreover, we suppose that the observation is contaminated with random noise. Although we no longer expect to recover the sparse vector $\boldsymbol{c}$ and modulating signals (which we denote as $\mathbf{D}_j$) exactly due to the existence of noise, we focus on recovering the sparse support of $\boldsymbol{c}$ and on bounding the recovery error of $\boldsymbol{c}$ and $\mathbf{D}_j$. By employing the lifting technique and under the assumption that the modulating signals live in a known, common subspace, we recast our problem as the recovery of a column-wise sparse matrix from structured linear observations. Under this formulation, there are no unknown parameters in the lifted linear operator. We solve the support recovery problem by solving a block ℓ_l-norm ($\ell_{2,1}$-norm) regularized quadratic minimization problem, which is also known as the group lasso in the statistics literature [11], [12]. The generalized model encompasses a wide range of applications, including direction of arrival (DOA) estimation for an antenna array with DOA sensitive channel responses [13], frequency estimation with damping in nuclear magnetic resonance spectroscopy [14], and CDMA communication with a spreading sequence sensitive channel [9]. To give a concrete example, we apply the proposed model to single molecule imaging [15] in Section IV-D.

B. Setup and Notation

Throughout the paper, we represent matrices, vectors, and scalars as bold uppercase, $\mathbf{X}$, bold lower case, $\boldsymbol{x}$, and non-bold letters, x, respectively. We use the symbol C to denote numerical constants that might vary from line to line. Given a support set T, the notation $\mathbf{X}_T$ represents the restriction of $\mathbf{X}$ to the columns indexed by T, and the notation $\boldsymbol{x}_T$ represents the restriction of $\boldsymbol{x}$ to the entries indexed by T. Moreover, we use $||\cdot||$ to denote the spectral norm, which returns the maximum singular value of a matrix, and $||\cdot||_F$ to denote the Frobenius norm. For a matrix $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M] \in \mathbf{C}^{K\times M}$, we define $||\mathbf{X}||_{2,1} = \sum_{j=1}^{M} ||\boldsymbol{x}_j||_2$ and $||\mathbf{X}||_{2,\infty} = \max_j ||\boldsymbol{x}_j||_2$. In addition, later in

Manuscript received October 29, 2019; revised January 29, 2020; accepted February 11, 2020. Date of publication February 24, 2020; date of current version April 24, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark A. Davenport. This work was supported by NSF under Grant CCF-1704204. This paper was presented in part at the 53rd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, Nov. 2019. *(Corresponding author: Gongguo Tang.)*

The authors are with the Department of Electrical Engineering, Colorado School of Mines, Golden, CO 80401 USA (e-mail: youyexie@mines.edu; mwakin@mines.edu; gtang@mines.edu).

Digital Object Identifier 10.1109/TSP.2020.2975935

the paper we will have the vectorized subgradient, $\boldsymbol{s} \in \mathbf{C}^{KM\times 1}$, of a function with respect to its matrix input $\mathbf{X} \in \mathbf{C}^{K\times M}$, and we define $||\boldsymbol{s}||_{2,\infty} = \max_j ||\boldsymbol{s}_j||_2$ where $\boldsymbol{s}_j$ is the subgradient with respect to $\boldsymbol{x}_j$.

C. Problem Formulation

In this paper, we consider the following generalized signal model with an unknown coefficient vector and non-stationary modulation process. Specifically, the observations consist of a contaminated composite signal

$$\boldsymbol{y} = \sum_{j=1}^{M} c_j \mathbf{D}_j \boldsymbol{a}_j + \boldsymbol{n} \in \mathbf{C}^N. \tag{I.1}$$

Here $c_j \in \mathbf{C}$ is an unknown scalar, $\mathbf{D}_j \in \mathbf{C}^{N\times N}$ is an unknown modulation matrix which is non-stationary as it depends on j, $\boldsymbol{a}_j$ is a dictionary atom coming from a dictionary matrix $\mathbf{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M] \in \mathbf{C}^{N\times M}$, and $\boldsymbol{n} \in \mathbf{C}^{N\times 1}$ is additive random Gaussian noise whose real and imaginary entries follow the i.i.d Gaussian distribution with mean 0 and variance σ^2.

Since there are more unknown parameters than the number of observations in the model (I.1), to make the recovery problem well-posed, we assume that at most $J\ (< M)$ of the coefficients c_j are non-zero and that the diagonal modulation matrices, $\mathbf{D}_j$, live in a common K-dimension subspace

$$\mathbf{D}_j = \mathrm{diag}(\mathbf{B}\boldsymbol{h}_j) \tag{I.2}$$

where $\mathbf{B} \in \mathbf{C}^{N\times K}$ $(N > K)$ is a known basis for the subspace with orthonormal columns, and $\boldsymbol{h}_j \in \mathbf{C}^{K\times 1}$ are unknown coefficient vectors. Similar subspace assumptions can be found in the deconvolution and demixing literature [16], [17]. Recovering c_j and $\boldsymbol{h}_j$ from $\boldsymbol{y}$ is a bilinear inverse problem [18], [19].

To combat the difficulties resulting from the bilinearity, we apply the lifting trick [8], [16], [20], which collects the unknown parameters into a matrix $\mathbf{X} = [c_1\boldsymbol{h}_1 \quad c_2\boldsymbol{h}_2 \quad \cdots \quad c_M\boldsymbol{h}_M] \in \mathbf{C}^{K\times M}$. By using Proposition 1 in [21] we can show that, when $\boldsymbol{n} = \mathbf{0}$, the observation model (I.1) takes the following equivalent form:

$$\boldsymbol{y}(n) = \boldsymbol{b}_n'^H \mathbf{X} \boldsymbol{a}_n', n = 1, \ldots, N. \tag{I.3}$$

where $\boldsymbol{b}_n'$ and $\boldsymbol{a}_n'$ are the n-th column of $\mathbf{B}^H$ and $\mathbf{A}^T$ respectively. We write (I.3) succinctly as $\boldsymbol{y} = \mathcal{L}(\mathbf{X})$ with $\mathcal{L}$ being a properly defined linear operator. And the adjoint of the linear operator $\mathcal{L}$ is $\mathcal{L}^*(\boldsymbol{y}) = \sum_{l=1}^N y_l \boldsymbol{b}_l' \boldsymbol{a}_l'^H$. The matrix $\mathbf{X}$ incorporates the unknown sparse signal and modulation process with at most $J\ (< M)$ non-zero columns. The support recovery problem we study in this paper aims to determine the indices, j, of the non-zero columns in $\mathbf{X}$ from the observation vector $\boldsymbol{y}$. We also aim to bound the recovery error of $\mathbf{X}$ in terms of the $\ell_{2,\infty}$-norm. If we assume there is no trivial null modulation, namely all $\mathbf{D}_j \neq 0$, finding the indices of the non-zero columns of $\mathbf{X}$ is equivalent to recovering the support of $\boldsymbol{c}$. Moreover, note that due to the scaling ambiguity between c_j and $\boldsymbol{h}_j$, the recovery error bound is expressed with respect to their multiplication $c_j\boldsymbol{h}_j$.

A natural way to recover the ground truth $\mathbf{X}_0$ from $\boldsymbol{y}$ is to exploit its sparse property and solve the following $\ell_{2,1}$-norm regularized quadratic minimization problem

$$\underset{\mathbf{X}\in\mathbf{C}^{K\times M}}{\mathrm{minimize}} \ \frac{1}{2}||\boldsymbol{y} - \mathcal{L}(\mathbf{X})||_2^2 + \lambda||\mathbf{X}||_{2,1}. \tag{I.4}$$

Alternatively, we can write (I.4) equivalently as

$$\underset{\mathbf{X}\in\mathbf{C}^{K\times M}}{\mathrm{minimize}} \ \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\Phi}\cdot \mathrm{vec}(\mathbf{X})||_2^2 + \lambda\sum_{i=1}^{M}||\boldsymbol{x}_i||_2. \tag{I.5}$$

Here $\mathcal{L}(\mathbf{X}) = \boldsymbol{\Phi}\cdot\mathrm{vec}(\mathbf{X})$ with

$$\begin{aligned}\boldsymbol{\Phi} = [&\boldsymbol{\phi}_{1,1} \quad \cdots \quad \boldsymbol{\phi}_{K,1} \quad \cdots \quad \boldsymbol{\phi}_{1,M} \quad \cdots \quad \boldsymbol{\phi}_{K,M}] \\ &\in \mathbf{C}^{N\times KM}\end{aligned} \tag{I.6}$$

and $\boldsymbol{\phi}_{i,j} = \mathrm{diag}(\boldsymbol{b}_i)\boldsymbol{a}_j \in \mathbf{C}^{N\times 1}$, where $\boldsymbol{b}_i$ is the i-th column of $\mathbf{B}$. Moreover, we denote the set containing the indices of the non-zero columns of the ground-truth matrix $\mathbf{X}_0$ as $T := T(\mathbf{X}_0)$ with $|T| = J$ and its complement as T^C. Due to the special block structure of $\boldsymbol{\Phi}$, when using the subscript notation $\boldsymbol{\Phi}_T$ we refer to the $N\times KJ$ sub-matrix of $\boldsymbol{\Phi}$ containing the $K(j-1)+1$ to $K(j-1)+K$-th columns for all $j \in T$.

D. Main Contributions

Our contributions are twofold. First, we propose to apply $\ell_{2,1}$-norm regularized quadratic minimization to recover the support of the generalized signal model in (I.1). Second, we derive sufficient conditions under which, with overwhelming probability, the support of the recovered signal is a subset of the support of the ground truth. More precisely, we show that the required number of observations, N, is proportional to the number of degrees of freedom, $O(JK)$, up to logarithmic factors. Moreover, the regularization parameter, λ, should be chosen to be proportional to the σ of the noise. We also bound the error in recovering the non-zero columns of the ground truth as measured in the $\ell_{2,\infty}$-norm. With an additional assumption on the ground truth signal, all conditions lead to exact support recovery.

E. Related Work

The $\ell_{2,1}$-norm constrained quadratic minimization problem, also known as the group lasso in statistics literature [11], [12], [22], has been widely studied. However, under our particular signal model (I.1), the linear operator $\boldsymbol{\Phi}$ contains randomness and has a special block structure as presented in (I.6), which distinguishes our work from other group lasso research. For example, [12] assumes each block of $\boldsymbol{\Phi}$, $[\phi_{1,j}, \ldots, \phi_{K,j}]$, to be orthonormal. [23] considers the adaptive group lasso and derives sufficient support recovery conditions using the block coherence of a deterministic $\boldsymbol{\Phi}$. [24] allows varying block sizes but still assumes a deterministic $\boldsymbol{\Phi}$. [25] assumes that $\boldsymbol{\Phi}$ has independent sub-exponential rows which is not consistent with our formulation, and they bound the recovery error in terms of ℓ_2-norm instead of $\ell_{2,\infty}$-norm as in our theorem. Moreover, [26], [27] provide a general recovery analysis for regression problems regularized with partly smooth functions relative to a manifold defined in [26], which encompasses the $\ell_{2,1}$-norm. However, the precise bounds on the regularization parameter and sample complexity for exact support recovery with $\mathbf{\Phi}$ defined in (I.6) are not derived, and that work bounds the error in terms of ℓ_2-norm instead of the $\ell_{2,\infty}$-norm.

As for the signal model itself, the model we study is closely related to certain works in self-calibration and blind deconvolution [9], [10]. The work in [17] considers a similar model except that the dictionary therein consists of all sampled sinusoids over a continuous frequency range, and its modulating waveforms, $\mathbf{D}_j$, are all the same. As an extension, [14] allows non-stationary modulating waveforms but still concerns the sinusoid dictionary. The fact that [14] considers a more general signal model than [17] actually facilitates the derivation of a near optimal result on the sufficient sample complexity. Our work in this paper similarly benefits from expanding the signal model of [9]. Specifically, our model fits into the self-calibration problem [9] when all $\mathbf{D}_j$ are the same. However, in the noisy case, [9] does not aim to recover the support and only bounds the error in terms of the ℓ_2-norm. [16] generalizes the model in [9] and can be interpreted as the self-calibration with multiple sensors, while allowing varying calibration parameters. However, [16] studies a constrained nuclear norm minimization problem with bounded noise and requires knowing the number of sensors. Additional related models for different applications, all requiring the same modulation matrix, are available in [10], [28]–[30].

We have also previously studied the sparse recovery and blind demodulation problem [8], [21] and numerically compared the support recovery performance of the SparseLift method [9] and the $\ell_{2,1}$-norm minimization method for direction of arrival estimation in [8]. In those works, however, we assume either zero or bounded additive noise, whereas we consider random Gaussian noise in this paper. Moreover, in [8], [21] we solve a constrained $\ell_{2,1}$-norm minimization problem due to the consideration of bounded noise. The regularized formulation used in this paper is a natural choice when considering unbounded noise [31] and is more convenient for support recovery analysis. Finally, in those papers, we derive the recovery error bound in terms of the ℓ_2-norm and do not study the question of exact support recovery when noise is involved.

The rest of the paper is organized as follows. In Section II, we present our main theorem regarding the support recovery problem. The detailed proof of the main theorem is shown in Section III. Several simulations and an experiment are conducted in Section IV to demonstrate the important scaling relationships and the effectiveness of our model in practical application. Finally, we conclude this paper in Section V.

II. Main Result

In this section, we present our main theorem, which presents the support recovery conditions and recovery error bound for solving (I.4) (or equivalently (I.5)). In this result, we assume that the dictionary matrix $\mathbf{A}$ is a random Gaussian matrix, by which we mean a matrix whose entries follow the i.i.d standard normal distribution.

Theorem II.1: Consider the observation model in equation (I.1), assume that $\mathbf{A} \in \mathbf{R}^{N\times M}$ $(N < M)$ is a random Gaussian matrix, at most J $(< M)$ coefficients c_j are nonzero, and the real and imaginary parts of each entry of the noise vector $\boldsymbol{n} \in \mathbf{C}^{N\times 1}$ follow the i.i.d Gaussian distribution with 0 mean and σ^2 variance. Suppose also that each modulation matrix $\mathbf{D}_j$ satisfies the subspace constraint (I.2), where $\mathbf{B}^H\mathbf{B} = \mathbf{I}_K$. If the number of observations

$$N \geq C_{\alpha,1}\mu_{\max}^2 JK\left[\log(M-J) + \log^2(N)\right] \tag{II.1}$$

and the regularization parameter

$$\lambda \geq \sqrt{C_{\alpha,2}\sigma^2\mu_{\max}^2 K\left[\log(M-J) + \log(N)\right]} \tag{II.2}$$

where $C_{\alpha,1}$ and $C_{\alpha,2}$ are constants that grow linearly with $\alpha > 1$ and the coherence parameter

$$\mu_{\max} = \max_{i,j}\sqrt{N}|\mathbf{B}_{ij}|,$$

then the following properties hold with probability at least $1 - O(N^{-\alpha+1})$:

1) Problem (I.5) has a unique solution $\hat{\mathbf{X}} \in \mathbf{C}^{K\times M}$ with its support, the set of indices of the non-zero columns in $\hat{\mathbf{X}}$, contained within the support T of the ground truth solution, $\mathbf{X}_0$.
2) The recovery error between the solution, $\hat{\mathbf{X}}$, and the ground truth, $\mathbf{X}_0$, satisfies

$$\begin{aligned}||\hat{\mathbf{X}} - \mathbf{X}_0||_{2,\infty} \leq \sqrt{C_\alpha\sigma^2\mu_{\max}^2 JK\left[\log(J) + \log(N)\right]}\\ + 4\sqrt{J}\lambda\end{aligned} \tag{II.3}$$

where C_α is a constant that grows linearly with α. If in addition the non-zero columns of $\mathbf{X}_0$ are bounded below

$$\begin{aligned}\min_{j\in T}||\boldsymbol{x}_{0,j}||_2 > \sqrt{C_\alpha\sigma^2\mu_{\max}^2 JK\left[\log(J) + \log(N)\right]}\\ + 4\sqrt{J}\lambda,\end{aligned} \tag{II.4}$$

then $\hat{\mathbf{X}}$ and $\mathbf{X}_0$ have exactly the same support which implies exact support recovery.

According to (II.3), we can derive that for any $\hat{\boldsymbol{x}}_j = \hat{c}_j\hat{\boldsymbol{h}}_j$ and $\boldsymbol{x}_{0,j} = c_{0,j}\boldsymbol{h}_{0,j}$ which are the j-th columns of the solution $\hat{\mathbf{X}}$ and the ground truth $\mathbf{X}_0$ respectively, $||\hat{c}_j\hat{\mathbf{D}}_j - c_{0,j}\mathbf{D}_{0,j}||_F = ||\hat{c}_j\hat{\boldsymbol{h}}_j - c_{0,j}\boldsymbol{h}_{0,j}||_2 \leq \sqrt{C_\alpha\sigma^2\mu_{\max}^2 JK[\log(J) + \log(N)]} + 4\sqrt{J}\lambda$. Moreover, since the columns of $\mathbf{B}$ are orthonormal, $\mu_{\max} \in [1,\sqrt{N}]$. Given the system parameters and a large enough N, (II.1) is satisfied when $1 \leq \mu_{\max} \leq \sqrt{\frac{N}{C_{\alpha,1}KJ[\log(M-J)+\log^2(N)]}}$. In addition, since we solve the column-wise sparse matrix support recovery problem via the group lasso and bound the recovery error in terms of $\ell_{2,\infty}$-norm, Theorem II.1 may be of interest outside the support recovery problem and shed light on the performance of the group lasso with random block structured linear operators.

III. Proof of Theorem II.1

We present proof of the main theorem in this section. We first derive the optimality and uniqueness conditions of the solution to (I.5) and then apply the primal-dual witness method [32] to construct a solution and find the conditions regarding the regularization parameter λ and number of observations N such that the optimality and uniqueness conditions are satisfied.

A. Optimality and Uniqueness Conditions

Lemma III.1:

1) A matrix $\hat{\mathbf{X}} \in \mathbf{C}^{K\times M}$ is an optimal solution to (I.5) if and only if there exists a subgradient vector $\boldsymbol{s} = \begin{bmatrix} \boldsymbol{s}_1 \\ \vdots \\ \boldsymbol{s}_M \end{bmatrix} \in$ vec$(\partial||\hat{\mathbf{X}}||_{2,1})$, such that

$$\boldsymbol{\Phi}^H\boldsymbol{\Phi} \cdot \text{vec}(\hat{\mathbf{X}}) - \boldsymbol{\Phi}^H\boldsymbol{y} + \lambda \cdot \begin{bmatrix} \boldsymbol{s}_1 \\ \vdots \\ \boldsymbol{s}_M \end{bmatrix} = 0$$

which is equivalent to

$$\boldsymbol{\Phi}^H\boldsymbol{\Phi} \cdot \left(\text{vec}(\hat{\mathbf{X}}) - \text{vec}(\mathbf{X}_0)\right) - \boldsymbol{\Phi}^H\boldsymbol{n} + \lambda\boldsymbol{s} = 0 \tag{III.1}$$

where $\boldsymbol{s}_i \in \mathbf{C}^K$ is the subgradient of $||\cdot||_2$ at $\hat{\boldsymbol{x}}_i$ defined as

$$\boldsymbol{s}_i = \begin{cases} \frac{\hat{\boldsymbol{x}}_i}{||\hat{\boldsymbol{x}}_i||_2} & \text{if } ||\hat{\boldsymbol{x}}_i||_2 \neq 0; \\ \{\boldsymbol{z} : ||\boldsymbol{z}||_2 \leq 1\} & \text{if } ||\hat{\boldsymbol{x}}_i||_2 = 0. \end{cases} \tag{III.2}$$

2) If the subgradient vectors of the optimal solution $\hat{\mathbf{X}}$ satisfy $||\boldsymbol{s}_i||_2 < 1$ for all $i \notin T(\hat{\mathbf{X}})$, then any optimal solution, $\check{\mathbf{X}}$, to (I.5) satisfies $\check{\boldsymbol{x}}_i = 0$ for all $i \notin T(\hat{\mathbf{X}})$.
3) When conditions in (2) are satisfied, if in addition $\boldsymbol{\Phi}^H_{T(\hat{\mathbf{X}})}\boldsymbol{\Phi}_{T(\hat{\mathbf{X}})} \in \mathbf{C}^{KJ\times KJ}$ is invertible, then $\hat{\mathbf{X}}$ is the unique solution to (I.5).

Proof:

1) Since problem (I.5) is convex, any optimal solution, $\hat{\mathbf{X}}$, must satisfy the first-order condition (III.1).
2) We first argue that when λ is fixed, for two arbitrary different optimal solutions $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$ to (I.5), we have $\boldsymbol{\Phi} \cdot \text{vec}(\hat{\mathbf{X}}_1) = \boldsymbol{\Phi} \cdot \text{vec}(\hat{\mathbf{X}}_2)$. This can be proved by contradiction as follows.
Assume $\boldsymbol{\Phi} \cdot \text{vec}(\hat{\mathbf{X}}_1) \neq \boldsymbol{\Phi} \cdot \text{vec}(\hat{\mathbf{X}}_2)$ for two arbitrary optimal solutions $\hat{\mathbf{X}}_1 \neq \hat{\mathbf{X}}_2$ to (I.5). By constructing $\hat{\mathbf{X}}_3 = \frac{1}{2}(\hat{\mathbf{X}}_1 + \hat{\mathbf{X}}_2)$, a little linear algebra yields

$$\frac{1}{2}||\boldsymbol{y} - \mathcal{L}(\hat{\mathbf{X}}_3)||_2^2 + \lambda||\hat{\mathbf{X}}_3||_{2,1} < \frac{1}{2}||\boldsymbol{y} - \mathcal{L}(\hat{\mathbf{X}}_k)||_2^2 + \lambda||\hat{\mathbf{X}}_k||_{2,1}$$

for $k \in \{1, 2\}$, due to the strict convexity of the function $f(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}||_2^2$ and the optimality of $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$. Thus, $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$ are not optimal. By contradiction, $\boldsymbol{\Phi} \cdot \text{vec}(\hat{\mathbf{X}}_1) = \boldsymbol{\Phi} \cdot \text{vec}(\hat{\mathbf{X}}_2)$. Then from (III.1), we can derive that $\boldsymbol{s}$ for different optimal solutions are the same. Therefore, assume we have an optimal solution $\hat{\mathbf{X}}$ such that $||\boldsymbol{s}_i||_2 < 1$ for all $i \notin T(\hat{\mathbf{X}})$, any other optimal solution, $\check{\mathbf{X}}$, would have subgradient vectors $||\check{\boldsymbol{s}}_i||_2 = ||\boldsymbol{s}_i||_2 < 1$ for all $i \notin T(\hat{\mathbf{X}})$ which implies $\check{\boldsymbol{x}}_i = 0$ according to (III.2).
3) If conditions in (2) are satisfied and $\boldsymbol{\Phi}^H_{T(\hat{\mathbf{X}})}\boldsymbol{\Phi}_{T(\hat{\mathbf{X}})} \in \mathbf{C}^{KJ\times KJ}$ is invertible, the solution of the support restricted problem $\frac{1}{2}||\boldsymbol{y} - \boldsymbol{\Phi}_{T(\hat{\mathbf{X}})} \cdot \text{vec}(\mathbf{X})||_2^2 + \lambda||\mathbf{X}||_{2,1}$ is unique by solving the restricted first order condition. ■

B. Primal-Dual Witness Construction

The method we apply to find the conditions regarding the regularization parameter λ and number of observations N for satisfying optimality and uniqueness conditions is the primal-dual witness method [32] which constructs the solution matrix, $\hat{\mathbf{X}}$, and subgradient vector, $\boldsymbol{s}$, through the following steps.

1) Conditioned on $\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T \in \mathbf{C}^{KJ\times KJ}$ is invertible, we first obtain $\hat{\mathbf{X}}_T \in \mathbf{C}^{K\times J}$ by solving the support restricted problem

$$\hat{\mathbf{X}}_T = \arg\min_{\mathbf{X}\in\mathbf{C}^{K\times J}} \left\{\frac{1}{2}||\boldsymbol{y} - \boldsymbol{\Phi}_T \cdot \text{vec}\ (\mathbf{X})||_2^2 + \lambda||\mathbf{X}||_{2,1}\right\}. \tag{III.3}$$

The solution $\hat{\mathbf{X}}_T$ is unique under the invertibility condition on $\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T$. And we set $\hat{\mathbf{X}}_{T^C} \in \mathbf{C}^{K\times(M-J)} = \mathbf{0}$. Thus, $\hat{\mathbf{X}}$ has support contained within the support T of the ground truth solution $\mathbf{X}_0$.
2) We calculate the subgradient vector $\boldsymbol{s}_T \in \mathbf{C}^{JK}$ based on $\hat{\mathbf{X}}_T$, where $\boldsymbol{s}_T$ is a sub-vector of $\boldsymbol{s}$ consisting of $\boldsymbol{s}_j$ for all $j \in T$.
3) We solve for a vector $\boldsymbol{s}_{T^C} \in \mathbf{C}^{(M-J)K}$ satisfying (III.1) and check whether $||\boldsymbol{s}_i||_2 < 1$ for all $i \notin T$.

If $\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T$ is invertible and $||\boldsymbol{s}_i||_2 < 1$ for all $i \in T^C$, $\hat{\mathbf{X}}$ constructed via the primal-dual witness method is the unique optimal solution to (I.5) with its support contained within the support of the ground truth solution $\mathbf{X}_0$. And note that the primal-dual witness construction succeeds only if the problem (I.5) has a unique solution whose support is contained within the support of the ground truth. The challenges of the construction lie in characterizing the regularization parameter λ and the number of observations N such that $||\boldsymbol{s}_i||_2 < 1$ for all $i \in T^C$.

To simplify the notation, without loss of generality, we assume the support of $\mathbf{X}_0$ is the first J columns and $T = \{1, 2, \ldots, J\}$ throughout the proof. Therefore, rewriting (III.1) into matrix multiplication form results in

$$\begin{bmatrix} \boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T & \boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_{T^C} \\ \boldsymbol{\Phi}_{T^C}^H\boldsymbol{\Phi}_T & \boldsymbol{\Phi}_{T^C}^H\boldsymbol{\Phi}_{T^C} \end{bmatrix} \begin{bmatrix} \text{vec}(\hat{\mathbf{X}}_T) - \text{vec}(\mathbf{X}_{0,T}) \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Phi}_T^H \\ \boldsymbol{\Phi}_{T^C}^H \end{bmatrix} \boldsymbol{n} + \lambda \begin{bmatrix} \boldsymbol{s}_T \\ \boldsymbol{s}_{T^C} \end{bmatrix} = 0. \tag{III.4}$$

When $\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T$ is invertible, from (III.4) we can derive that

$$\Delta(\mathbf{X}) = \text{vec}(\hat{\mathbf{X}}_T) - \text{vec}(\mathbf{X}_{0,T}) = (\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\left(\boldsymbol{\Phi}_T^H\boldsymbol{n} - \lambda\boldsymbol{s}_T\right)$$

and

$$\boldsymbol{s}_{T^C} = \frac{1}{\lambda}\left(\boldsymbol{\Phi}_{T^C}^H\boldsymbol{n} - \boldsymbol{\Phi}_{T^C}^H\boldsymbol{\Phi}_T\Delta(\mathbf{X})\right). \tag{III.5}$$

Substituting the full expression of $\Delta(\mathbf{X})$ into (III.5) results in

$$\begin{aligned}\boldsymbol{s}_{T^C} &= \boldsymbol{\Phi}_{T^C}^H \left(\mathbf{I}_N - \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{\Phi}_T^H\right)\frac{\boldsymbol{n}}{\lambda}\\ &\quad + \boldsymbol{\Phi}_{T^C}^H\boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{s}_T. \qquad \text{(III.6)}\end{aligned}$$

C. Important Lemmas

In this section, we introduce some important lemmas and propositions that will be applied during the proof of Theorem II.1. First is the isometry bound for the linear operator $\mathcal{L}$ defined in (I.3) (and $\boldsymbol{\Phi}$ defined in (I.6)) which can be found in Lemma 4.3 in [9].

Lemma III.2: [9, Lemma 4.3] (Isometry) For the linear operator $\mathcal{L}$ defined in (I.3) with $\mathbf{B}^H\mathbf{B} = \mathbf{I}_K$ and $\delta > 0$,

$$||\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T - \mathbf{I}_T|| = ||\mathcal{L}_T^*\mathcal{L}_T - \mathbf{I}_T|| \leq \delta$$

with probability at least $1 - N^{-\alpha+1}$ where $\mathbf{I}_T$ is the identity operator on the support T such that $\mathbf{I}_T(\mathbf{X}) = \mathbf{X}_T$, if $\mathbf{A}$ is a random Gaussian matrix and $N \geq C_\alpha\mu_{\max}^2 KJ \max\{\log(N)/\delta^2, \log^2(N)/\delta\}$. Here C_α is a constant that grows linearly with $\alpha > 1$.

According to Lemma A.12 in [33], if $||\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T - \mathbf{I}_T|| \leq \delta < 1$, $\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T$ is invertible and $||(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}|| \leq (1-\delta)^{-1}$. In addition, we have the following quadratic Gaussian tail bound proposition, developed from Theorem 1 in [34].

Proposition 1: Let $\mathbf{H} \in \mathbf{C}^{K\times N}$ and $\boldsymbol{\Sigma} = \mathbf{H}^H\mathbf{H}$. Let $\boldsymbol{a} \in \mathbf{C}^N$ whose real and imaginary entries follow the i.i.d normal distribution with 0 mean and σ^2 variance. For all $\alpha > 0$,

$$\begin{aligned}\Pr\left(||\mathbf{H}\boldsymbol{a}||_2^2 > \sigma^2\left[2\,\mathrm{Tr}\,(\boldsymbol{\Sigma}) + 2\sqrt{2\,\mathrm{Tr}\,(\boldsymbol{\Sigma}^2)\alpha} + 2||\boldsymbol{\Sigma}||\alpha\right]\right)\\ \leq e^{-\alpha}.\end{aligned}$$

If $\boldsymbol{a} \in \mathbf{R}^N$ only contains the real part, for all $\alpha > 0$,

$$\begin{aligned}\Pr\left(||\mathbf{H}\boldsymbol{a}||_2^2 > \sigma^2\left[\mathrm{Tr}\,(\boldsymbol{\Sigma}) + 2\sqrt{\mathrm{Tr}\,(\boldsymbol{\Sigma}^2)\alpha} + 2||\boldsymbol{\Sigma}||\alpha\right]\right)\\ \leq e^{-\alpha}.\end{aligned}$$

Proof: When $\mathbf{H}$ and $\boldsymbol{a}$ are a complex matrix and vector, we can write $\mathbf{H} = \mathbf{H}_R + i\mathbf{H}_I$ and $\boldsymbol{a} = \boldsymbol{a}_R + i\boldsymbol{a}_I$ where $\mathbf{H}_R$, $\mathbf{H}_I$, $\boldsymbol{a}_R$ and $\boldsymbol{a}_I$ are all real and the entries of $\boldsymbol{a}_R$ and $\boldsymbol{a}_I$ are i.i.d Gaussian random variables with 0 mean and σ^2 variance. We then have

$$\begin{aligned}||\mathbf{H}\boldsymbol{a}||_2^2 &= ||(\mathbf{H}_R + i\mathbf{H}_I)(\boldsymbol{a}_R + i\boldsymbol{a}_I)||_2^2\\ &= ||(\mathbf{H}_R\boldsymbol{a}_R - \mathbf{H}_I\boldsymbol{a}_I) + i(\mathbf{H}_R\boldsymbol{a}_I + \mathbf{H}_I\boldsymbol{a}_R)||_2^2\\ &= ||\mathbf{H}_R\boldsymbol{a}_R - \mathbf{H}_I\boldsymbol{a}_I||_2^2 + ||\mathbf{H}_R\boldsymbol{a}_I + \mathbf{H}_I\boldsymbol{a}_R||_2^2\\ &= \left|\left|\begin{bmatrix}\mathbf{H}_R & -\mathbf{H}_I\\ \mathbf{H}_I & \mathbf{H}_R\end{bmatrix}\begin{bmatrix}\boldsymbol{a}_R\\ \boldsymbol{a}_I\end{bmatrix}\right|\right|_2^2.\end{aligned}$$

Define $\mathbf{H}_o = \begin{bmatrix}\mathbf{H}_R & -\mathbf{H}_I\\ \mathbf{H}_I & \mathbf{H}_R\end{bmatrix}$ and $\boldsymbol{\Sigma}_o = \mathbf{H}_o^T\mathbf{H}_o$. $\boldsymbol{\Sigma}_o$ has the form

$$\begin{aligned}\boldsymbol{\Sigma}_o &= \begin{bmatrix}\mathbf{H}_R^T & \mathbf{H}_I^T\\ -\mathbf{H}_I^T & \mathbf{H}_R^T\end{bmatrix}\begin{bmatrix}\mathbf{H}_R & -\mathbf{H}_I\\ \mathbf{H}_I & \mathbf{H}_R\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{H}_R^T\mathbf{H}_R + \mathbf{H}_I^T\mathbf{H}_I & -\mathbf{H}_R^T\mathbf{H}_I + \mathbf{H}_I^T\mathbf{H}_R\\ -\mathbf{H}_I^T\mathbf{H}_R + \mathbf{H}_R^T\mathbf{H}_I & \mathbf{H}_I^T\mathbf{H}_I + \mathbf{H}_R^T\mathbf{H}_R\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{H}_1 & -\mathbf{H}_2\\ \mathbf{H}_2 & \mathbf{H}_1\end{bmatrix} \qquad \text{(III.7)}\end{aligned}$$

where we define $\mathbf{H}_1 = \mathbf{H}_R^T\mathbf{H}_R + \mathbf{H}_I^T\mathbf{H}_I$ and $\mathbf{H}_2 = -\mathbf{H}_I^T\mathbf{H}_R + \mathbf{H}_R^T\mathbf{H}_I$. Applying Theorem 1 in [34], we get

$$\begin{aligned}\Pr\left(||\mathbf{H}\boldsymbol{a}||_2^2 > \sigma^2\left[\mathrm{Tr}\,(\boldsymbol{\Sigma}_o) + 2\sqrt{\mathrm{Tr}\,(\boldsymbol{\Sigma}_o^2)\alpha} + 2||\boldsymbol{\Sigma}_o||\alpha\right]\right)\\ \leq e^{-\alpha}.\end{aligned}$$

If we further define

$$\begin{aligned}\boldsymbol{\Sigma} = \mathbf{H}^H\mathbf{H} &= (\mathbf{H}_R + i\mathbf{H}_I)^H(\mathbf{H}_R + i\mathbf{H}_I)\\ &= (\mathbf{H}_R^T - i\mathbf{H}_I^T)(\mathbf{H}_R + i\mathbf{H}_I)\\ &= (\mathbf{H}_R^T\mathbf{H}_R + \mathbf{H}_I^T\mathbf{H}_I) + i(-\mathbf{H}_I^T\mathbf{H}_R + \mathbf{H}_R^T\mathbf{H}_I)\\ &= \mathbf{H}_1 + i\mathbf{H}_2, \qquad \text{(III.8)}\end{aligned}$$

by comparing (III.7) and (III.8), one can check that $\mathrm{Tr}\,(\boldsymbol{\Sigma}_o) = 2\,\mathrm{Tr}\,(\mathbf{H}_1) = 2\,\mathrm{Tr}\,(\boldsymbol{\Sigma})$ since $\mathrm{Tr}\,(\mathbf{H}_2) = 0$, $\mathrm{Tr}\,(\boldsymbol{\Sigma}_o^2) = ||\boldsymbol{\Sigma}_o||_F^2 = 2(||\mathbf{H}_1||_F^2 + ||\mathbf{H}_2||_F^2) = 2||\boldsymbol{\Sigma}||_F^2 = 2\,\mathrm{Tr}\,(\boldsymbol{\Sigma}^2)$, and

$$\begin{aligned}||\boldsymbol{\Sigma}_o|| &= \left|\left|\begin{bmatrix}\mathbf{H}_1 & -\mathbf{H}_2\\ \mathbf{H}_2 & \mathbf{H}_1\end{bmatrix}\right|\right|\\ &= \max_{\left|\left|\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}\right|\right|_2 = 1}\left|\left|\begin{bmatrix}\mathbf{H}_1 & -\mathbf{H}_2\\ \mathbf{H}_2 & \mathbf{H}_1\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}\right|\right|_2\\ &= \max_{\left|\left|\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\end{bmatrix}\right|\right|_2 = 1}\sqrt{||\mathbf{H}_1\boldsymbol{x}_1 - \mathbf{H}_2\boldsymbol{x}_2||_2^2 + ||\mathbf{H}_2\boldsymbol{x}_1 + \mathbf{H}_1 x_2||_2^2}\\ &= \max_{||\boldsymbol{x}_1||_2^2 + ||\boldsymbol{x}_2||_2^2 = 1}\sqrt{||(\mathbf{H}_1\boldsymbol{x}_1 - \mathbf{H}_2\boldsymbol{x}_2) + i(\mathbf{H}_2\boldsymbol{x}_1 + \mathbf{H}_1\boldsymbol{x}_2)||_2^2}\\ &= \max_{||\boldsymbol{x}_1||_2^2 + ||\boldsymbol{x}_2||_2^2 = 1}\sqrt{||(\mathbf{H}_1 + i\mathbf{H}_2)(\boldsymbol{x}_1 + i\boldsymbol{x}_2)||_2^2}\\ &= \max_{||\boldsymbol{x}_1||_2^2 + ||\boldsymbol{x}_2||_2^2 = 1}||\boldsymbol{\Sigma}(\boldsymbol{x}_1 + i\boldsymbol{x}_2)||_2 = ||\boldsymbol{\Sigma}||\end{aligned}$$

where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2 \in \mathbf{R}^N$ since $\boldsymbol{\Sigma}_o$ is a real matrix, so that the vector corresponding to its largest singular value is also real. Therefore, we have

$$\begin{aligned}\Pr\left(||\mathbf{H}\boldsymbol{a}||_2^2 > \sigma^2\left[2\,\mathrm{Tr}\,(\boldsymbol{\Sigma}) + 2\sqrt{2\,\mathrm{Tr}\,(\boldsymbol{\Sigma}^2)\alpha} + 2||\boldsymbol{\Sigma}||\alpha\right]\right)\\ \leq e^{-\alpha}.\end{aligned}$$

Similarly, when $\boldsymbol{a}$ only contains the real part

$$||\mathbf{H}\boldsymbol{a}||_2^2 = \left|\left|\begin{bmatrix}\mathbf{H}_R \\ \mathbf{H}_I\end{bmatrix} \boldsymbol{a}_R\right|\right|_2^2,$$

$\boldsymbol{\Sigma}$ still follows (III.8) and

$$\boldsymbol{\Sigma}_o = [\mathbf{H}_R^T\mathbf{H}_R + \mathbf{H}_I^T\mathbf{H}_I] = \mathbf{H}_1.$$

In this case, Tr $(\boldsymbol{\Sigma}_o) =$ Tr $(\boldsymbol{\Sigma})$, Tr $(\boldsymbol{\Sigma}_o^2) \leq$ Tr $(\boldsymbol{\Sigma}^2)$ and since $\boldsymbol{\Sigma}_o$ is real, we have

$$\begin{aligned}||\boldsymbol{\Sigma}_o|| &= \max_{||\boldsymbol{x}||_2=1, \boldsymbol{x}\in\mathbf{R}^N} \sqrt{||\mathbf{H}_1\boldsymbol{x}||_2^2} \\ &\leq \max_{||\boldsymbol{x}||_2=1, \boldsymbol{x}\in\mathbf{R}^N} \sqrt{||\mathbf{H}_1\boldsymbol{x}||_2^2 + ||\mathbf{H}_2\boldsymbol{x}||_2^2} \\ &= \max_{||\boldsymbol{x}||_2=1, \boldsymbol{x}\in\mathbf{R}^N} \sqrt{||\mathbf{H}_1\boldsymbol{x} + i\mathbf{H}_2\boldsymbol{x}||_2^2} \\ &\leq \max_{||\boldsymbol{x}||_2=1, \boldsymbol{x}\in\mathbf{C}^N} \sqrt{||(\mathbf{H}_1 + i\mathbf{H}_2)\boldsymbol{x}||_2^2} \\ &= \max_{||\boldsymbol{x}||_2=1, \boldsymbol{x}\in\mathbf{C}^N} ||\boldsymbol{\Sigma}\boldsymbol{x}||_2 = ||\boldsymbol{\Sigma}||.\end{aligned}$$

So we have, for $\alpha > 0$,

$$\begin{aligned}\text{Tr }(\boldsymbol{\Sigma}) + 2\sqrt{\text{Tr }(\boldsymbol{\Sigma}^2)\alpha} + 2||\boldsymbol{\Sigma}||\alpha &\geq \text{Tr }(\boldsymbol{\Sigma}_o) \\ &+ 2\sqrt{\text{Tr }(\boldsymbol{\Sigma}_o^2)\alpha} + 2||\boldsymbol{\Sigma}_o||\alpha\end{aligned}$$

which results in

$$\begin{aligned}\Pr\left(||\mathbf{H}\boldsymbol{a}||_2^2 > \sigma^2\left[\text{Tr }(\boldsymbol{\Sigma}) + 2\sqrt{\text{Tr }(\boldsymbol{\Sigma}^2)\alpha} + 2||\boldsymbol{\Sigma}||\alpha\right]\right) \\ \leq e^{-\alpha}.\end{aligned}$$

■

Proposition 2: Let $\mathbf{H} \in \mathbf{C}^{K\times N}$ and $\boldsymbol{\Sigma} = \mathbf{H}^H\mathbf{H}$. Let $\boldsymbol{a} \in \mathbf{C}^N$ whose real and imaginary entries follow the i.i.d normal distribution with 0 mean and σ^2 variance. For all $\alpha > 1$,

$$\Pr\left(||\mathbf{H}\boldsymbol{a}||_2^2 > \sigma^2\left[2 + (2\sqrt{2}+2)\alpha\right] \text{Tr }(\boldsymbol{\Sigma})\right) \leq e^{-\alpha}.$$

If $\boldsymbol{a} \in \mathbf{R}^N$ only contains the real part, for all $\alpha > 1$,

$$\Pr\left(||\mathbf{H}\boldsymbol{a}||_2^2 > \sigma^2 (1+4\alpha) \text{Tr }(\boldsymbol{\Sigma})\right) \leq e^{-\alpha}.$$

Proof: Since $\boldsymbol{\Sigma}$ is a positive semi-definite and hermitian matrix, all its eigenvalues, λ_i, are non-negative. Thus, Tr $(\boldsymbol{\Sigma}^2) = \sum_{i=1}^N \lambda_i^2 \leq (\sum_{i=1}^N \lambda_i)^2 =$ Tr $(\boldsymbol{\Sigma})^2$ and $||\boldsymbol{\Sigma}|| = \lambda_{\max} \leq \sum_{i=1}^N \lambda_i =$ Tr $(\boldsymbol{\Sigma})$. As a result, for $\alpha > 1$,

$$\begin{aligned}&\sigma^2\left[2 + (2\sqrt{2}+2)\alpha\right] \text{Tr }(\boldsymbol{\Sigma}) \\ &\quad\geq \sigma^2\left[2\,\text{Tr }(\boldsymbol{\Sigma}) + 2\sqrt{2\,\text{Tr }(\boldsymbol{\Sigma}^2)\alpha} + 2||\boldsymbol{\Sigma}||\alpha\right]\end{aligned}$$

and

$$\begin{aligned}&\sigma^2 (1+4\alpha) \text{Tr }(\boldsymbol{\Sigma}) \\ &\quad\geq \sigma^2\left[\text{Tr }(\boldsymbol{\Sigma}) + 2\sqrt{\text{Tr }(\boldsymbol{\Sigma}^2)\alpha} + 2||\boldsymbol{\Sigma}||\alpha\right].\end{aligned}$$

Then applying Proposition 1 yields Proposition 2. ■

D. Bounding $||\boldsymbol{s}_{T^C}||_{2,\infty}$

Recalling (III.6), to prove that $||\boldsymbol{s}_i||_2 < 1$ for all $i \in T^C$ which is equivalent to $||\boldsymbol{s}_{T^C}||_{2,\infty} < 1$, where the $\ell_{2,\infty}$-norm of the subgradient vector is defined in Section I-B, we only need to show that for a constant $\gamma \in (0,1)$,

$$||\boldsymbol{\Phi}_{T^C}^H \left(\mathbf{I}_N - \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{\Phi}_T^H\right)\frac{\boldsymbol{n}}{\lambda}||_{2,\infty} \leq \frac{\gamma}{2}$$

and

$$||\boldsymbol{\Phi}_{T^C}^H\boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{s}_T||_{2,\infty} \leq \frac{\gamma}{2}.$$

Then by the triangle inequality, $||\boldsymbol{s}_{T^C}||_{2,\infty} \leq \gamma < 1$.

Lemma III.3: Conditioned on $\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T$ being invertible, we have

$$||\boldsymbol{\Phi}_{T^C}^H \left(\mathbf{I}_N - \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{\Phi}_T^H\right)\frac{\boldsymbol{n}}{\lambda}||_{2,\infty} \leq \frac{\gamma}{2}$$

for $\gamma \in (0,1)$ with probability at least $1 - N^{-\alpha+1}$ when

$$\lambda \geq \sqrt{\frac{C_\alpha\sigma^2\mu_{\max}^2K[\log(M-J) + \log(N)]}{\gamma^2}}$$

and

$$N \geq 10\log(M-J) + 10\alpha\log(N)$$

where C_α is a constant that grows linearly with $\alpha > 1$.

Proof: $||\boldsymbol{\Phi}_{T^C}^H(\mathbf{I}_N - \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{\Phi}_T^H)\frac{\boldsymbol{n}}{\lambda}||_{2,\infty} \leq \frac{\gamma}{2}$ for $\gamma \in (0,1)$ is equivalent to

$$\max_{i\in T^C} ||\boldsymbol{\Phi}_i^H\left(\mathbf{I}_N - \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{\Phi}_T^H\right)\boldsymbol{n}||_2^2 \leq \frac{\lambda^2\gamma^2}{4}$$

where $\boldsymbol{\Phi}_i$ $(i \in T^C)$ is the sub-matrix containing the $[K(i-1)+1]$ to $[K(i-1)+K]$-th columns of $\boldsymbol{\Phi}$. If we define $\mathbf{H}_i = \boldsymbol{\Phi}_i^H(\mathbf{I}_N - \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{\Phi}_T^H)$, the projection matrix $\mathbf{P} = (\mathbf{I}_N - \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{\Phi}_T^H)$, and $\boldsymbol{\Sigma} = \mathbf{H}_i^H\mathbf{H}_i$, we get

$$\begin{aligned}\text{Tr }(\boldsymbol{\Sigma}) &= ||\mathbf{H}_i||_F^2 = ||\boldsymbol{\Phi}_i^H \cdot \mathbf{P}||_F^2 = ||\mathbf{B}^H \operatorname{diag}(\boldsymbol{a}_i)^H \cdot \mathbf{P}||_F^2 \\ &= ||\mathbf{P}\operatorname{diag}(\boldsymbol{a}_i)\mathbf{B}||_F^2 \\ &= ||\mathbf{P}\left[\operatorname{diag}(\boldsymbol{b}_1)\boldsymbol{a}_i, \operatorname{diag}(\boldsymbol{b}_2)\boldsymbol{a}_i, \ldots, \operatorname{diag}(\boldsymbol{b}_K)\boldsymbol{a}_i\right]||_F^2 \\ &= \sum_{k=1}^K ||\mathbf{P}\operatorname{diag}(\boldsymbol{b}_k)\boldsymbol{a}_i||_2^2 \leq \sum_{k=1}^K ||\mathbf{P}||^2\frac{\mu_{\max}^2}{N}||\boldsymbol{a}_i||_2^2 \\ &\leq \frac{\mu_{\max}^2K}{N}||\boldsymbol{a}_i||_2^2.\end{aligned}$$

Since $\boldsymbol{n}$ is the additive Gaussian noise vector, applying Proposition 2 gives us, for $\alpha_1 > 1$

$$\begin{aligned}\Pr\left(||\mathbf{H}_i\boldsymbol{n}||_2^2 > \frac{\lambda^2\gamma^2}{4} \geq \sigma^2\left[2 + (2\sqrt{2}+2)\alpha_1\right]\text{Tr }(\boldsymbol{\Sigma})\right) \\ \leq e^{-\alpha_1},\end{aligned} \tag{III.9}$$

in which we need

$$\begin{aligned}\lambda &\geq \sqrt{\frac{\sigma^2\left[8 + (8\sqrt{2}+8)\alpha_1\right]\mu_{\max}^2K||\boldsymbol{a}_i||_2^2}{\gamma^2N}} \\ &\geq \sqrt{\frac{\sigma^2\left[8 + (8\sqrt{2}+8)\alpha_1\right]\text{Tr }(\boldsymbol{\Sigma})}{\gamma^2}}.\end{aligned} \tag{III.10}$$

To control the term $||\boldsymbol{a}_i||_2^2$, we define an event $E = \{\max_{i\in T^C} ||\boldsymbol{a}_i||_2^2 < 2N\}$. Because each entry of $\boldsymbol{a}_i \in \mathbf{R}^N$ follows the standard normal distribution, $||\boldsymbol{a}_i||_2^2$ is a χ_N^2 random variable. According to Lemma 1 in [35], for $\alpha_2 > 0$

$$\Pr(||\boldsymbol{a}_i||_2^2 \geq 2\sqrt{\alpha_2 N} + 2\alpha_2 + N) \leq e^{-\alpha_2}.$$

By solving $2N \geq 2\sqrt{\alpha_2 N} + 2\alpha_2 + N$, we require $\alpha_2 \leq (\frac{2\sqrt{3}-2}{4})^2 N \approx 0.1340N$. So for $0 < \alpha_2 \leq \frac{N}{10}$, we have

$$\Pr(||\boldsymbol{a}_i||_2^2 \geq 2N) \leq e^{-\alpha_2}.$$

Taking the union over all $i \in T^C$ gives us

$$\Pr(E^C) \leq (M-J)e^{-\alpha_2}$$

which is meaningful when $\log(M-J) \leq \alpha_2 \leq \frac{N}{10}$.

In addition, if we define another event $F = \{\max_{i\in T^C} ||\mathbf{H}_i \boldsymbol{n}||_2^2 > \frac{\lambda^2\gamma^2}{4}\}$, conditioned on E and with

$$\lambda \geq \sqrt{\frac{\sigma^2\left[16 + (16\sqrt{2}+16)\alpha_1\right]\mu_{\max}^2 K}{\gamma^2}}, \tag{III.11}$$

by taking the union of (III.9) over all $i \in T^C$, we obtain

$$\Pr(F \mid E) \leq (M-J)e^{-\alpha_1}.$$

Therefore,

$$\begin{aligned}\Pr(F \mid E) + \Pr(E^C) &\leq (M-J)e^{-\alpha_1} + (M-J)e^{-\alpha_2}\\ &= 2N^{-\alpha} \leq N^{-\alpha+1}\end{aligned}$$

for $\alpha > 1$ by setting $\alpha_1 = \alpha_2 = \log(M-J) + \alpha\log(N)$. Substituting α_1 into (III.11) and some simplification yields

$$\lambda \geq \sqrt{\frac{C_\alpha \sigma^2 \mu_{\max}^2 K[\log(M-J) + \log(N)]}{\gamma^2}}$$

where $C_\alpha = (16\sqrt{2}+16)\alpha + 16$ is a constant that grows linearly with $\alpha > 1$. Moreover, $\log(M-J) \leq \alpha_2 = \log(M-J) + \alpha\log(N) \leq \frac{N}{10}$ requires $N \geq 10\log(M-J) + 10\alpha\log(N)$. Finally, the law of probability implies

$$\begin{aligned}&\Pr(\max_{i\in T^C} ||\boldsymbol{\Phi}_i^H\left(\mathbf{I}_N - \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{\Phi}_T^H\right)\boldsymbol{n}||_2^2 \leq \frac{\lambda^2\gamma^2}{4})\\ &= \Pr(F^C) \geq \Pr(F^C \cap E) = 1 - [\Pr(E^C) + \Pr(F \cap E)]\\ &\geq 1 - [\Pr(E^C) + \Pr(F|E)] \geq 1 - N^{-\alpha+1}.\end{aligned}$$

■

Lemma III.4: Conditioned on $\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T$ being invertible and $||(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}|| \leq 2$, we have

$$||\boldsymbol{\Phi}_{T^C}^H\boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{s}_T||_{2,\infty} \leq \frac{\gamma}{2}$$

for $\gamma \in (0,1)$ with probability at least $1 - N^{-\alpha}$ when

$$N \geq C_\alpha \frac{\mu_{\max}^2 KJ}{\gamma^2}[\log(M-J) + \log(N)],$$

where C_α is a constant that grows linearly with $\alpha > 1$.

Proof: $||\boldsymbol{\Phi}_{T^C}^H\boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{s}_T||_{2,\infty} \leq \frac{\gamma}{2}$ for $\gamma \in (0,1)$ can be reformulated as

$$\begin{aligned}&\max_{i\in T^C} ||\boldsymbol{\Phi}_i^H\boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{s}_T||_2^2\\ &= \max_{i\in T^C}\left|\left|\begin{bmatrix}\boldsymbol{a}_i^H\operatorname{diag}(\boldsymbol{b}_1)^H\\ \boldsymbol{a}_i^H\operatorname{diag}(\boldsymbol{b}_2)^H\\ \vdots\\ \boldsymbol{a}_i^H\operatorname{diag}(\boldsymbol{b}_K)^H\end{bmatrix}\cdot\boldsymbol{v}\right|\right|_2^2\\ &= \max_{i\in T^C} ||\boldsymbol{v}^H[\operatorname{diag}(\boldsymbol{b}_1)\boldsymbol{a}_i, \operatorname{diag}(\boldsymbol{b}_2)\boldsymbol{a}_i, \ldots, \operatorname{diag}(\boldsymbol{b}_K)\boldsymbol{a}_i]||_2^2\\ &= \max_{i\in T^C}\left|\left|\begin{bmatrix}\boldsymbol{v}^H\operatorname{diag}(\boldsymbol{b}_1)\\ \boldsymbol{v}^H\operatorname{diag}(\boldsymbol{b}_2)\\ \vdots\\ \boldsymbol{v}^H\operatorname{diag}(\boldsymbol{b}_K)\end{bmatrix}\boldsymbol{a}_i\right|\right|_2^2 = \max_{i\in T^C}||\mathbf{H}\boldsymbol{a}_i||_2^2 \leq \frac{\gamma^2}{4}\end{aligned}$$

where we define $\boldsymbol{v} = \boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{s}_T \in \mathbf{C}^N$ and

$$\mathbf{H} = \begin{bmatrix}\boldsymbol{v}^H\operatorname{diag}(\boldsymbol{b}_1)\\ \boldsymbol{v}^H\operatorname{diag}(\boldsymbol{b}_2)\\ \vdots\\ \boldsymbol{v}^H\operatorname{diag}(\boldsymbol{b}_K)\end{bmatrix} \in \mathbf{C}^{K\times N}.$$

Therefore, for $\boldsymbol{\Sigma} = \mathbf{H}^H\mathbf{H}$, we have

$$\begin{aligned}\operatorname{Tr}(\boldsymbol{\Sigma}) &= \left|\left|\begin{bmatrix}\boldsymbol{v}^H\operatorname{diag}(\boldsymbol{b}_1)\\ \boldsymbol{v}^H\operatorname{diag}(\boldsymbol{b}_2)\\ \vdots\\ \boldsymbol{v}^H\operatorname{diag}(\boldsymbol{b}_K)\end{bmatrix}\right|\right|_F^2 \leq \frac{\mu_{\max}^2 K}{N}||\boldsymbol{v}||_2^2\\ &\leq \frac{2\mu_{\max}^2 KJ}{N}\end{aligned} \tag{III.12}$$

since

$$\begin{aligned}||\boldsymbol{v}||_2^2 &= |\boldsymbol{v}^H\boldsymbol{v}| = |\boldsymbol{s}_T^H(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{s}_T|\\ &= |\boldsymbol{s}_T^H(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{s}_T| \leq ||(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}|| \cdot ||\boldsymbol{s}_T||_2^2 \leq 2J.\end{aligned}$$

Because $\boldsymbol{a}_i \in \mathbf{R}^N$ for $i \in T^C$ is independent of $\boldsymbol{\Phi}_T$ and $\boldsymbol{a}_i$'s entries follow the i.i.d standard normal distribution, Proposition 2 implies, for $\alpha_1 > 1$

$$\Pr\left(||\mathbf{H}\boldsymbol{a}_i||_2^2 > (1+4\alpha_1)\operatorname{Tr}(\boldsymbol{\Sigma})\right) \leq e^{-\alpha_1}.$$

To ensure that $\frac{\gamma^2}{4} \geq (1+4\alpha_1)\operatorname{Tr}(\boldsymbol{\Sigma})$, we need

$$N \geq \frac{(8+32\alpha_1)\mu_{\max}^2 KJ}{\gamma^2}. \tag{III.13}$$

By taking the union over all $i \in T^C$, we get

$$\begin{aligned}\Pr\left(\max_{i\in T^C}||\boldsymbol{\Phi}_i^H\boldsymbol{\Phi}_T(\boldsymbol{\Phi}_T^H\boldsymbol{\Phi}_T)^{-1}\boldsymbol{s}_T||_2^2\frac{\gamma^2}{4}\right) &\leq (M-J)e^{-\alpha_1}\\ &= N^{-\alpha}\end{aligned}$$

if we set $\alpha_1 = \log(M-J) + \alpha \log(N)$ for $\alpha > 1$. Substituting the full expression of α_1 into (III.13) and some simplification yields

$$N \geq C_\alpha \frac{\mu_{\max}^2 KJ}{\gamma^2}[\log(M-J) + \log(N)]$$

where $C_\alpha = 32\alpha + 8$ is a constant that grows linearly with $\alpha > 1$. ■

E. Bounding $||\hat{\mathbf{X}}_T - \mathbf{X}_{0,T}||_{2,\infty}$

When the support of the unique optimal solution $\hat{\mathbf{X}}$ is contained within the support of the ground truth solution $\mathbf{X}_0$, the recovery error $||\hat{\mathbf{X}} - \mathbf{X}_0||_{2,\infty} = ||\hat{\mathbf{X}}_T - \mathbf{X}_{0,T}||_{2,\infty}$. And because the optimal solution on the support, $\hat{\mathbf{X}}_T \in \mathbf{C}^{K\times J}$ (we assume $\hat{\mathbf{X}}_T \neq \mathbf{X}_{0,T}$, otherwise $||\mathbf{X}_{0,T} - \hat{\mathbf{X}}_T||_{2,\infty} = 0$) is attained by solving the support-restricted regularized least square problem (III.3) whose objective function $f(\mathrm{vec}(\mathbf{X})) = \frac{1}{2}||\boldsymbol{y} - \mathbf{\Phi}_T \cdot \mathrm{vec}(\mathbf{X})||_2^2 + \lambda||\mathbf{X}||_{2,1}$ is strongly convex, since $\frac{1}{2}||\boldsymbol{y} - \mathbf{\Phi}_T \cdot \mathrm{vec}(\mathbf{X})||_2^2$ is strongly convex conditioned on $\mathbf{\Phi}_T^H\mathbf{\Phi}_T$ being positive definite and $\lambda||\mathbf{X}||_{2,1}$ is convex, by the property of the strongly convex function, we have

$$f(\mathrm{vec}(\hat{\mathbf{X}}_T)) \geq f(\mathrm{vec}(\mathbf{X}_{0,T})) + \mathrm{Re}\left\{g(\mathrm{vec}(\mathbf{X}_{0,T}))^H \cdot \left[\mathrm{vec}(\hat{\mathbf{X}}_T) - \mathrm{vec}(\mathbf{X}_{0,T})\right]\right\} + \frac{m}{2}||\hat{\mathbf{X}}_T - \mathbf{X}_{0,T}||_F^2$$

where $g(\mathrm{vec}(\mathbf{X}_{0,T}))$ is the subgradient of $f(\mathrm{vec}(\mathbf{X}_{0,T}))$. In addition, if we set $\delta = \frac{1}{2}$ in Lemma III.2, we have $||(\mathbf{\Phi}_T^H\mathbf{\Phi}_T)^{-1}|| \leq 2$ according to Lemma A.12 in [33], which implies $\mathbf{\Phi}_T^H\mathbf{\Phi}_T \succeq \frac{1}{2}\mathbf{I}$. As a result, $m = \frac{1}{2}$. Then by the Hölder inequality,

$$\begin{aligned} f(\mathrm{vec}(\hat{\mathbf{X}}_T)) &\geq f(\mathrm{vec}(\mathbf{X}_{0,T})) + \mathrm{Re}\left\{g(\mathrm{vec}(\mathbf{X}_{0,T}))^H \cdot \right. \\ &\quad \left. \left[\mathrm{vec}(\hat{\mathbf{X}}_T) - \mathrm{vec}(\mathbf{X}_{0,T})\right]\right\} + \frac{1}{4}||\hat{\mathbf{X}}_T - \mathbf{X}_{0,T}||_F^2 \\ &\geq f(\mathrm{vec}(\mathbf{X}_{0,T})) - ||g(\mathrm{vec}(\mathbf{X}_{0,T}))||_{2,\infty} \cdot \\ &\quad ||\hat{\mathbf{X}}_T - \mathbf{X}_{0,T}||_{2,1} + \frac{1}{4}||\hat{\mathbf{X}}_T - \mathbf{X}_{0,T}||_F^2 \\ &\geq f(\mathrm{vec}(\mathbf{X}_{0,T})) - ||g(\mathrm{vec}(\mathbf{X}_{0,T}))||_{2,\infty} \cdot \\ &\quad ||\hat{\mathbf{X}}_T - \mathbf{X}_{0,T}||_{2,1} + \frac{1}{4\sqrt{J}}||\hat{\mathbf{X}}_T \\ &\quad - \mathbf{X}_{0,T}||_{2,\infty}||\hat{\mathbf{X}}_T - \mathbf{X}_{0,T}||_{2,1} \end{aligned} \tag{III.14}$$

where the $\ell_{2,\infty}$-norm of the subgradient vector is defined in Section I-B, and the third inequality comes from the fact that $\frac{||\mathbf{X}||_F^2}{||\mathbf{X}||_{2,1}||\mathbf{X}||_{2,\infty}} \geq \frac{1}{\sqrt{J}}$. Because if $||\mathbf{X}||_F^2 = L$, one can check that $||\mathbf{X}||_{2,\infty} \leq \sqrt{L}$ and $||\mathbf{X}||_{2,1} \leq \sqrt{LJ}$ where the equality is achieved when the 2-norm of all J columns are the same.

Therefore, since $\hat{\mathbf{X}}_T \neq \mathbf{X}_{0,T}$ and $\hat{\mathbf{X}}_T$ is the optimal solution, $f(\mathrm{vec}(\hat{\mathbf{X}}_T)) \leq f(\mathrm{vec}(\mathbf{X}_{0,T}))$, (III.14) yields

$$\begin{aligned} ||\hat{\mathbf{X}}_T - \mathbf{X}_{0,T}||_{2,\infty} &\leq 4\sqrt{J}||g(\mathrm{vec}(\mathbf{X}_{0,T}))||_{2,\infty} \\ &= 4\sqrt{J}||\mathbf{\Phi}_T^H\left[\mathbf{\Phi}_T\mathrm{vec}(\mathbf{X}_{0,T}) - \boldsymbol{y}\right] + \lambda\boldsymbol{s}_{0,T}||_{2,\infty} \\ &= 4\sqrt{J}|| - \mathbf{\Phi}_T^H\boldsymbol{n} + \lambda\boldsymbol{s}_{0,T}||_{2,\infty} \\ &\leq 4\sqrt{J}\left(||\mathbf{\Phi}_T^H\boldsymbol{n}||_{2,\infty} + ||\lambda\boldsymbol{s}_{0,T}||_{2,\infty}\right) \\ &= 4\sqrt{J}\left(||\mathbf{\Phi}_T^H\boldsymbol{n}||_{2,\infty} + \lambda\right) \end{aligned} \tag{III.15}$$

where we have used $\boldsymbol{y} = \mathbf{\Phi}_T\mathrm{vec}(\mathbf{X}_{0,T}) + \boldsymbol{n}$ and $||\boldsymbol{s}_{0,T}||_{2,\infty} = 1$. Now we turn to bound the term $||\mathbf{\Phi}_T^H\boldsymbol{n}||_{2,\infty}$ applying the following lemma.

Lemma III.5: Conditioned on $\mathbf{\Phi}_T^H\mathbf{\Phi}_T$ being invertible, we have

$$||\mathbf{\Phi}_T^H\boldsymbol{n}||_{2,\infty} \leq \sqrt{C_\alpha\sigma^2\mu_{\max}^2K\left[\log(J) + \log(N)\right]} \tag{III.16}$$

with probability at least $1 - N^{-\alpha+1}$ when

$$N \geq 10\log(J) + 10\alpha\log(N)$$

where C_α is a constant that grows linearly with $\alpha > 1$.

Proof: If we define $\mathbf{\Phi}_j$ $(j \in T)$ to be the $[K(j-1)+1]$ to $[K(j-1)+K]$-th columns of $\mathbf{\Phi}$, we have $||\mathbf{\Phi}_T^H\boldsymbol{n}||_{2,\infty} = \max_{j\in T}||\mathbf{\Phi}_j^H\boldsymbol{n}||_2$. For an arbitrary $j \in T$, let $\mathbf{\Sigma} = \mathbf{\Phi}_j\mathbf{\Phi}_j^H$,

$$\begin{aligned} \mathrm{Tr}\ (\mathbf{\Sigma}) &= \mathrm{Tr}\ (\mathbf{\Phi}_j\mathbf{\Phi}_j^H) = \mathrm{Tr}\ (\mathbf{\Phi}_j^H\mathbf{\Phi}_j) \\ &= ||\mathbf{\Phi}_j||_F^2 = ||\left[\mathrm{diag}(\boldsymbol{b}_1)\boldsymbol{a}_j, \ldots, \mathrm{diag}(\boldsymbol{b}_K)\boldsymbol{a}_j\right]||_F^2 \\ &= \sum_{k=1}^{K}||\,\mathrm{diag}(\boldsymbol{b}_k)\boldsymbol{a}_j||_2^2 \leq \frac{\mu_{\max}^2K}{N}||\boldsymbol{a}_j||_2^2. \end{aligned}$$

If we define an event $E = \{\max_{j\in T}||\boldsymbol{a}_j||_2^2 < 2N\}$, in the proof of Lemma III.3 we have shown that for $0 < \alpha_1 \leq \frac{N}{10}$

$$\mathrm{Pr}(||\boldsymbol{a}_i||_2^2 \geq 2N) \leq e^{-\alpha_1}.$$

Taking the union over all $j \in T$ results in

$$\mathrm{Pr}(E^C) \leq Je^{-\alpha_1}$$

which is meaningful when $\log(J) \leq \alpha_1 \leq \frac{N}{10}$. Therefore, conditioned on E, $\mathrm{Tr}\ (\mathbf{\Sigma}) < 2\mu_{\max}^2K$. Applying Proposition 2 gives us, for $\alpha_2 > 1$

$$\mathrm{Pr}(||\mathbf{\Phi}_j^H\boldsymbol{n}||_2^2 > \left[4 + (4\sqrt{2}+4)\alpha_2\right]\sigma^2\mu_{\max}^2K \mid E) \leq e^{-\alpha_2}.$$

Taking the union over all $j \in T$ yields

$$\begin{aligned} &\mathrm{Pr}(\max_{j\in T}||\mathbf{\Phi}_j^H\boldsymbol{n}||_2^2 > \left[4 + (4\sqrt{2}+4)\alpha_2\right]\sigma^2\mu_{\max}^2K \mid E) \\ &\quad \leq Je^{-\alpha_2}. \end{aligned} \tag{III.17}$$

Therefore,

$$\begin{aligned} &\mathrm{Pr}(\max_{j\in T}||\mathbf{\Phi}_j^H\boldsymbol{n}||_2^2 > \left[4 + (4\sqrt{2}+4)\alpha_2\right]\sigma^2\mu_{\max}^2K \mid E) \\ &\quad + \mathrm{Pr}(E^C) \leq Je^{-\alpha_2} + Je^{-\alpha_1} = 2N^{-\alpha} \leq N^{-\alpha+1} \end{aligned}$$

if we set $\alpha_1 = \alpha_2 = \log(J) + \alpha\log(N)$ for $\alpha > 1$. Moreover, $\log(J) \leq \alpha_1 = \log(J) + \alpha\log(N) \leq \frac{N}{10}$ requires that $N \geq 10\log(J) + 10\alpha\log(N)$. Substituting $\alpha_2 = \log(J) + \alpha\log(N)$ into (III.17) and some simplification yields that, for an event

$$F = \left\{\max_{j\in T}||\mathbf{\Phi}_j^H\boldsymbol{n}||_2 > \sqrt{C_\alpha\sigma^2\mu_{\max}^2K\left[\log(J) + \log(N)\right]}\right\}$$

where $C_\alpha = (4\sqrt{2}+4)\alpha + 4$, we have $\Pr(F \mid E) + \Pr(E^C) \le N^{-\alpha+1}$. Therefore,

$$\begin{aligned}
&\Pr\left(||\boldsymbol{\Phi}_T^H \boldsymbol{n}||_{2,\infty} \le \sqrt{C_\alpha \sigma^2 \mu_{\max}^2 K\left[\log(J)+\log(N)\right]}\right)\\
&\quad = \Pr(F^C) \ge \Pr(F^C \cap E) = 1 - [\Pr(E^C) + \Pr(F \cap E)]\\
&\quad \ge 1 - [\Pr(E^C) + \Pr(F|E)] \ge 1 - N^{-\alpha+1}.
\end{aligned}$$

■

F. Proof of Theorem II.1

We now sum up the related lemmas to derive the final results in Theorem II.1. By setting $\delta = \frac{1}{2}$, Lemma III.2 shows that $\boldsymbol{\Phi}_T^H \boldsymbol{\Phi}_T$ is invertible and $||(\boldsymbol{\Phi}_T^H \boldsymbol{\Phi}_T)^{-1}|| \le 2$ with probability at least $1 - N^{-\alpha+1}$ when $N \ge C_{\alpha,0}\mu_{\max}^2 KJ\log^2(N)$ for $\alpha > 1$.

By applying the same α to Lemma III.3 and III.4 and setting $\gamma = \frac{1}{2}$, we can get that, conditioned on $\boldsymbol{\Phi}_T^H \boldsymbol{\Phi}_T$ being invertible and $||(\boldsymbol{\Phi}_T^H \boldsymbol{\Phi}_T)^{-1}|| \le 2$, $||\boldsymbol{s}_{T^C}||_{2,\infty} \le \frac{1}{2}$ which implies that the support of the unique optimal solution $\hat{\mathbf{X}}$ to (I.5) is contained within the support of the ground truth solution $\mathbf{X}_0$, with probability at least $1 - 2N^{-\alpha+1}$, when

$$\lambda \ge \sqrt{C_{\alpha,2}\sigma^2\mu_{\max}^2 K[\log(M-J)+\log(N)]} \qquad \text{(III.18)}$$

and

$$N \ge C_{\alpha,3}\mu_{\max}^2 JK[\log(M-J)+\log(N)].$$

As for the recovery error, we apply the same α to Lemma III.5 and substitute (III.16) into (III.15). As a result, conditioned on $\boldsymbol{\Phi}_T^H \boldsymbol{\Phi}_T$ being invertible and the support of the unique optimal solution $\hat{\mathbf{X}}$ being contained within the support of $\mathbf{X}_0$,

$$\begin{aligned}
||\hat{\mathbf{X}} - \mathbf{X}_0||_{2,\infty} &\le 4\sqrt{J}\left(||\boldsymbol{\Phi}_T^H \boldsymbol{n}||_{2,\infty} + \lambda\right)\\
&\le \sqrt{C_\alpha \sigma^2 \mu_{\max}^2 JK\left[\log(J)+\log(N)\right]} + 4\sqrt{J}\lambda \qquad \text{(III.19)}
\end{aligned}$$

where C_α is a constant that grows linearly with α, with probability at least $1 - N^{-\alpha+1}$ when $N \ge 10\log(J) + 10\alpha\log(N)$.

Therefore, after combining the probability and the requirement on N and λ, we can conclude that, with probability at least $1 - 4N^{-\alpha+1}$, (I.5) has a unique optimal solution $\hat{\mathbf{X}}$ with its support contained within the support of the ground truth solution $\mathbf{X}_0$ and the recovery error in terms of $\ell_{2,\infty}$-norm satisfies (III.19) when λ satisfies (III.18) and $N \ge C_{\alpha,1}\mu_{\max}^2 JK[\log(M-J) + \log^2(N)]$ where $C_{\alpha,1} = \max\{C_{\alpha,0}, C_{\alpha,3}\}$ for $\alpha > 1$.

IV. Numerical Simulations

In this section, we present several numerical simulations to demonstrate and support the theoretical results in Theorem II.1. In these simulations, each entry of the dictionary $\mathbf{A} \in \mathbf{R}^{N\times M}$ is sampled independently from the standard normal distribution and $\mathbf{B} \in \mathbf{C}^{N\times K}$ contains the first K columns of the normalized $N \times N$ DFT matrix. The real and imaginary components of $c_j \in \mathbf{C}$ and $\boldsymbol{h}_j \in \mathbf{C}^{K\times 1}$ follow the i.i.d standard normal distribution and the support, T with $|T| = J$, of the ground truth solution $\mathbf{X}_0 = [c_1\boldsymbol{h}_1, \ldots, \boldsymbol{c}_M\boldsymbol{h}_M] \in \mathbf{C}^{K\times M}$ is selected uniformly at random. Problem (I.5) is solved via CVX [36].

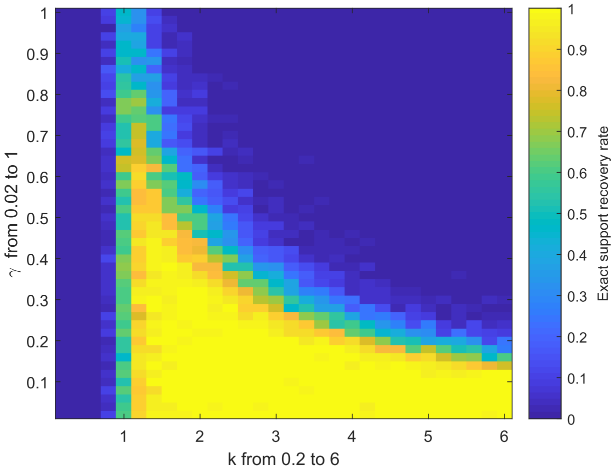


Fig. 1. The relation between k and γ in terms of the exact support recovery rate where $\lambda = k\gamma_0$ and $\gamma = \frac{\gamma_0}{\min_{j\in T}||\boldsymbol{x}_{0,j}||_2}$.

A. Range of λ for Exact Support Recovery

In the first simulation, we determine the effective range of λ for exact support recovery. Theoretically, (II.2) provides a lower bound for λ such that Theorem II.1 holds and (II.4) gives an upper bound to achieve exact support recovery. To verify the bounds of λ, we define $\gamma_0 = \sqrt{\sigma^2\mu_{\max}^2 K[\log(M-J)+\log(N)]}$ and $\gamma = \frac{\gamma_0}{\min_{j\in T}||\boldsymbol{x}_{0,j}||_2}$. (II.2) implies that we could set $\lambda = k\gamma_0$ for some $k > 0$. In addition, according to (II.2) and (II.4) in Theorem II.1, when all system parameters except λ are fixed, to achieve exact support recovery, λ should satisfy

$$C_1\gamma_0 \le \lambda = k\gamma_0 < \frac{\min_{j\in T}||\boldsymbol{x}_{0,j}||_2 - C_2}{C_3} \qquad \text{(IV.1)}$$

which is equivalent to

$$C_1 \le k < \frac{C_4}{\gamma} - C_5$$

where $C_4 = \frac{1}{C_3}$ and $C_5 = \frac{C_2}{C_3\gamma_0}$. To examine this relation, we fix $\sigma = 0.1$, $J = K = 3$, $N = 100$, and $M = 150$, and we vary k and γ. 50 trials are run for each (k,γ) pair and we record the exact support recovery rate in Fig. 1, from which we do observe that k should be larger than a constant which is approximately 1.2 under this setting and that k has a reciprocal relation with γ.

B. Number of Observations N for Exact Support Recovery

Equation (II.1) in Theorem II.1 indicates that the sufficient number of observations, N, scales nearly linearly with respect to the subspace dimension K and the sparsity J. To verify that, in the second simulation, we set $M = 150$, $k = 3$, and $\gamma = 0.02$ to make sure that λ is in an appropriate range for exact support recovery. We vary N and K (with fixed $J = 3$) and record the exact support recovery rate in Fig. 2. The result of a similar simulation but varying N and J (with fixed $K = 3$) is shown in Fig. 3. 50 simulations are run for each setting. We observe that linear scaling of the number of observations N with the

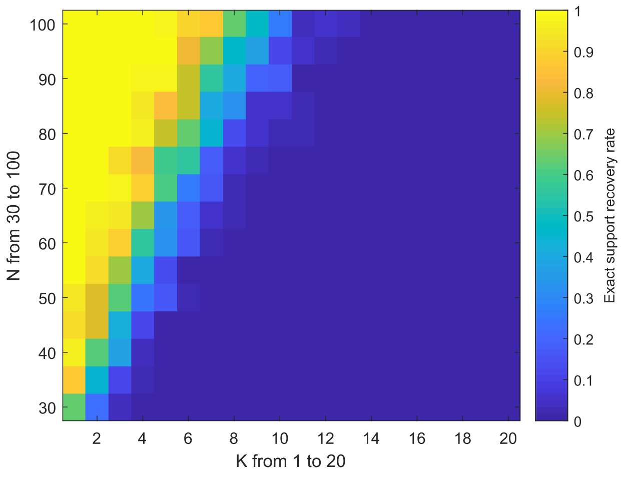


Fig. 2. The nearly linear relation between the number of observations, N, and the subspace dimension K, to achieve exact support recovery.

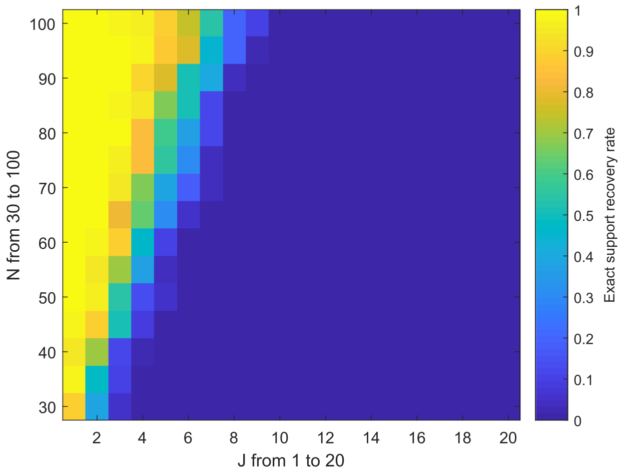


Fig. 3. The nearly linear relation between the number of observations, N, and the sparsity J, to achieve exact support recovery.

subspace dimension K and the sparsity J is sufficient for exact support recovery.

C. Recovery Error Bound

Next we turn to verify the recovery error bound in (II.3), which scales linearly with respect to λ and nearly linearly with respect to $\sqrt{J}$. We set $K = 3$, $N = 100$, $M = 150$, and $\gamma = 0.02$. In Fig. 4, we use $\lambda = k\gamma_0$ (with fixed $J = 3$) and vary k within the proper range for exact support recovery based on Fig. 1. 100 trials are run for each k and we calculate the mean and standard deviation of the recovery error $||\hat{\mathbf{X}} - \mathbf{X}_0||_{2,\infty}$. Note that the recovery error is counted only when the exact support recovery is achieved. In this figure, we do observe linear scaling of the error with λ.

Similarly, we vary J (with fixed $\lambda = 3\gamma_0$) within the proper range for exact support recovery based on Fig. 3 and record the squared recovery error $||\hat{\mathbf{X}} - \mathbf{X}_0||_{2,\infty}^2$ in Fig. 5. Again, the squared recovery error is counted only when the exact support recovery is achieved. In this figure, we can observe nearly linear scaling of the squared error with J.

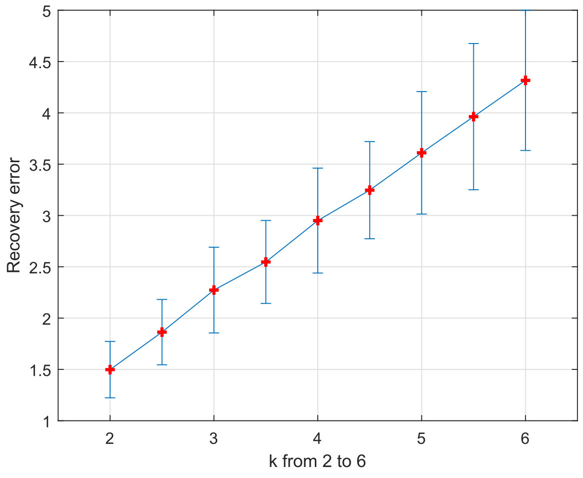


Fig. 4. The linear relation between the recovery error, $||\hat{\mathbf{X}} - \mathbf{X}_0||_{2,\infty}$, and the regularization parameter $\lambda = k\gamma_0$. The red plus signs and the blue horizontal sticks indicates the mean and standard deviation of the recovery error.

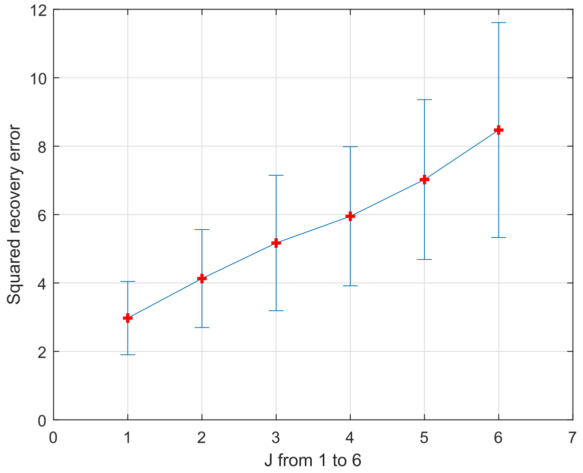
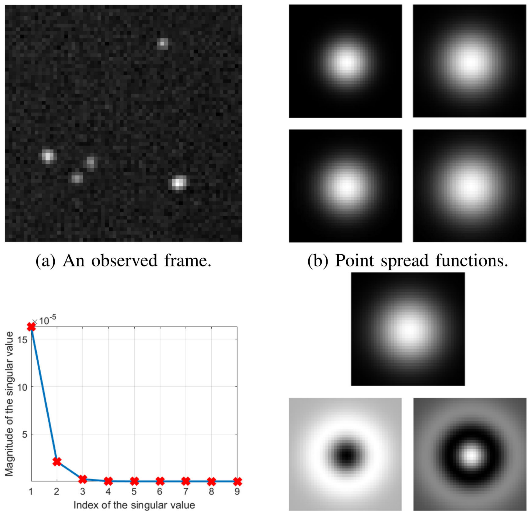


Fig. 5. The nearly linear relation between the squared recovery error, $||\hat{\mathbf{X}} - \mathbf{X}_0||_{2,\infty}^2$, and the sparsity J. The red plus signs and the blue horizontal sticks indicates the mean and standard deviation of the squared recovery error.

D. Single Molecule Imaging

In this experiment, we apply our signal model (I.1) to the single molecule imaging problem and achieve super-resolution by solving (I.5). In molecule imaging via stochastic optical reconstruction microscopy (STORM) [37], the sub-cellular structures are dyed using fluorophores, and during each observation only a small portion of the fluorophores are activated. Moreover, fluorophores at different depths will undergo different degrees of blurring.

Consequently, each observed image frame consists of a few activated fluorophores convolved with the non-stationary Gaussian point spread functions of the microscope as shown in Fig. 6(a). Specifically, the observed low resolution frame is of size 64×64 pixels and each pixel corresponds to a region of



(a) An observed frame.

(b) Point spread functions.

(c) The singular values.

(d) The singular vectors.

Fig. 6. The analysis of point spread functions. (a) A typical observed frame is of size 64×64 pixels with each pixel corresponding to a region of size 100×100 nm. (b) Four examples of the non-stationary point spread functions. (c) The singular values of the point spread functions. (d) The singular vectors corresponding to the three largest singular values.

size 100×100 nm. The goal is to construct a target image with 320×320 pixels with each pixel corresponding to a region of size 20×20 nm.

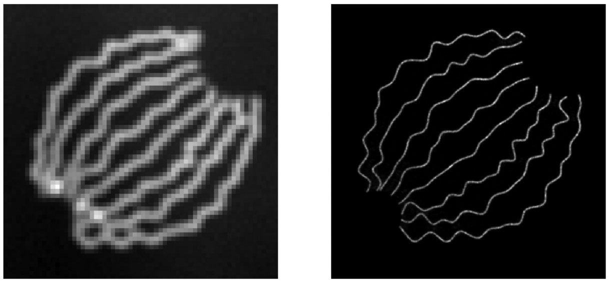
If we vectorize the frames, each observed low resolution frame can be represented as

$$\boldsymbol{y} = Sample\left[\sum_{j=1}^{M} c_j(\mathbf{B}'\boldsymbol{h}_j) \circledast \boldsymbol{e}_j + \boldsymbol{n}'\right] \in \mathbf{R}^{N\times 1} \qquad \text{(IV.2)}$$

where $Sample[\cdot]$ indicates the sub-sampling operator, $N = 64 \times 64 = 4096$, and $M = 320 \times 320 = 102400$. Moreover, c_j is the unknown fluorophore intensity at the j-th position, $\mathbf{B}'$ models the subspace containing the non-stationary Gaussian point spread functions (with unknown coefficient vector $\boldsymbol{h}_j$ for the j-th position), $\boldsymbol{e}_j \in \mathbf{R}^M$ is the j-th column of the identity matrix, and $\boldsymbol{n}'$ is the unknown additive noise. All observed frames, $\boldsymbol{y}$, come from the Single-Molecule Localization Microscopy grand challenge organized by ISBI [38]. The dataset contains 12000 low resolution frames, and the maximum number of activated fluorophores in each frame is 18 which implies that at most $J = 18$ coefficients c_j are non-zero for each $\boldsymbol{y}$.

To apply our model, we must construct the subspace, $\mathbf{B}'$, to capture the non-stationary point spread functions. By changing the variances (widths), we generate nine different Gaussian point spread functions; four examples are shown in Fig. 6(b). We then apply the singular value decomposition (SVD) to a matrix of the vectorized point spread functions and record their singular values in Fig. 6(c). From this we see that the point spread functions approximately live in a 3 dimensional subspace. Therefore, we set $K = 3$ and let $\mathbf{B}'$ contain the singular vectors corresponding to the 3 largest singular values. We display the corresponding singular vectors in Fig. 6(d).

To better illustrate the connection between the single molecule imaging problem and the signal model we study, (IV.2) can be equivalently represented as



(a) Low resolution input.

(b) Super-resolution result.

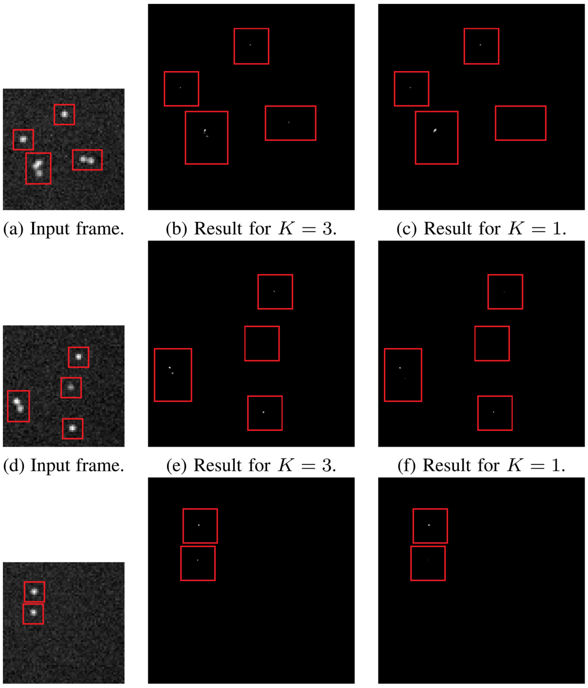
Fig. 7. The single molecule imaging experiment. The image in (a) is of size 64×64 pixels with each pixel corresponding to a region of size 100×100 nm. (b) shows the super-resolution result, which has size 320×320 pixels with each pixel corresponding to a region of size 20×20 nm.

$$\boldsymbol{y} = Sample\left\{IDFT\left[\sum_{j=1}^{M} c_j \mathbf{D}_j \boldsymbol{a}_j + \boldsymbol{n}\right]\right\} \in \mathbf{R}^{N\times 1}$$

where $IDFT[\cdot]$ denotes the inverse discrete Fourier transform operator, $\mathbf{D}_j = \mathrm{diag}(\mathbf{B}\boldsymbol{h}_j)$ where $\mathbf{B} = DFT[\mathbf{B}']$, $\boldsymbol{a}_j$s are the DFTs of spikes at all possible spatial locations, and $\boldsymbol{n} = DFT[\boldsymbol{n}']$. In this case, if we represent $\boldsymbol{y} = \mathcal{L}(\mathbf{X})$ with $\mathbf{X} = [c_1\boldsymbol{h}_1, \ldots, c_M\boldsymbol{h}_M]$, the linear operator $\mathcal{L}$ incorporates additional inverse Fourier transform and sub-sample operators, and $\mathbf{A}$ is a Fourier dictionary instead of random Gaussian. The noise $\boldsymbol{n}$, $\boldsymbol{h}_j$, and c_j for all j are unknown, and the indices of the non-zero columns in $\mathbf{X}$ indicate the locations of the activated fluorophores in the high resolution image.

We pre-process each low resolution frame by subtracting the average intensity of the data set, and superimposing all the frame results in the low resolution image in Fig. 7(a). Moreover, we solve (I.5) for each observed low resolution frame via SpaRSA [39]. By superimposing all the high resolution images that we get, we obtain the super-resolution result in Fig. 7(b). Although the dictionary is not Gaussian in this application, the superior super-resolution result verifies the effectiveness of the proposed signal model and minimization problem.

Finally, when $K = 1$, (I.5) degenerates to the classical ℓ_1-norm constrained lasso problem which has been comprehensively studied. However, by choosing $K = 1$, the model sacrifices its ability to capture non-stationary modulation, which is significant in this problem when the point spread functions have several comparable singular values. Although in our case, we happen to have one dominant singular value as shown in Fig. 6(c), which implies that super-resolution can be attempted with $K = 1$, we see that a larger K still benefits the super-resolution process. To demonstrate this, we run the single molecule imaging experiments again using $K = 3$ and $K = 1$. Three super-resolution examples are shown in Fig. 8, from which we can find that although $K = 3$ and $K = 1$ achieve similar performance, some activated fluorophores can be more accurately represented using the 3-dimensional subspace ($K = 3$), and that leads to a more clear and accurate super-resolution result.



(a) Input frame. (b) Result for $K = 3$. (c) Result for $K = 1$.

(d) Input frame. (e) Result for $K = 3$. (f) Result for $K = 1$.

(g) Input frame. (h) Result for $K = 3$. (i) Result for $K = 1$.

Fig. 8. Comparison between the super-resolution results for $K = 3$ and $K = 1$. (a), (d), and (g) are three low-resolution input frames. (b), (e), and (h) show the super-resolution results for $K = 3$. (c), (f), and (i) show the super-resolution results for $K = 1$. The area of interest is highlighted using the red rectangle. The input frames are of size 64 × 64 pixels and the outputs are 320 × 320 pixels.

V. Conclusion

In this paper, we consider the problem of recovering a sparse signal with unbounded noise and non-stationary blind modulation. Using the lifting technique and with a subspace assumption on the modulating signals, we recast this problem as the recovery of a column-wise sparse matrix from structured linear observations. We apply $\ell_{2,1}$-norm regularized quadratic minimization, also known as the group lasso, to solve this problem and derive sufficient conditions on the sample complexity and regularization parameter for exact support recovery. We also bound the recovery error in terms of the $\ell_{2,\infty}$-norm. Numerical simulations are consistent with our predictions and support the theoretical results. Moreover, we apply our model to single molecule imaging and achieve promising super-resolution results. One useful generalization of the results in this paper would be to consider the random Fourier dictionary. Allowing $\mathbf{D}_j$ to be non-diagonal but live in a low-dimensional matrix subspace is another important generalization, which could open up other potential applications.

References

[1] Y. Xie, M. B. Wakin, and G. Tang, "Support recovery for sparse recovery and non-stationary blind demodulation," in *Proc. IEEE 53rd Asilomar Conf. Signals, Syst. Comput.*, 2019, pp. 5566–5570.

[2] H. Mansour, D. Liu, P. T. Boufounos, and U. S. Kamilov, "Radar autofocus using sparse blind deconvolution," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1623–1627.

[3] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid mr imaging," *Magn. Reson. Medicine: An Official J. Int. Soc. Magn. Reson. Medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.

[4] H. Mansour and Ö. Yilmaz, "Adaptive compressed sensing for video acquisition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 3465–3468.

[5] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[6] E. J. Candes and Y. Plan, "A probabilistic and ripless theory of compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7235–7254, Nov. 2011.

[7] Y. Xie, S. Li, G. Tang, and M. B. Wakin, "Radar signal demixing via convex optimization," in *Proc. IEEE 22nd Int. Conf. Digit. Signal Process.*, 2017, pp. 1–5.

[8] Y. Xie, M. B. Wakin, and G. Tang, "Simultaneous sparse recovery and blind demodulation," *IEEE Trans. Signal Process.*, vol. 67, no. 19, pp. 5184–5199, Oct. 2019.

[9] S. Ling and T. Strohmer, "Self-calibration and biconvex compressive sensing," *Inverse Problems*, vol. 31, no. 11, 2015, Art. no. 115002.

[10] A. Ahmed, B. Recht, and J. Romberg, "Blind deconvolution using convex programming," *IEEE Trans. Inf. Theory*, vol. 60, no. 3, pp. 1711–1732, Mar. 2014.

[11] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statistical Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[12] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Soc.: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[13] B. Friedlander and T. Strohmer, "Bilinear compressed sensing for array self-calibration," in *Proc. IEEE 48th Asilomar Conf. Signals, Syst. Comput.*, 2014, pp. 363–367.

[14] D. Yang, G. Tang, and M. B. Wakin, "Super-resolution of complex exponentials from modulations with unknown waveforms," *IEEE Trans. Inf. Theory*, vol. 62, no. 10, pp. 5809–5830, Oct. 2016.

[15] G. Tang, B. N. Bhaskar, and B. Recht, "Sparse recovery over continuous dictionaries-just discretize," in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, 2013, pp. 1043–1047.

[16] S. Ling and T. Strohmer, "Blind deconvolution meets blind demixing: Algorithms and performance bounds," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4497–4520, Jul. 2017.

[17] Y. Chi, "Guaranteed blind sparse spikes deconvolution via lifting and convex optimization," *J. Sel. Topics Signal Process.*, vol. 10, no. 4, pp. 782–794, 2016.

[18] S. Ling and T. Strohmer, "Self-calibration and bilinear inverse problems via linear least squares," *SIAM J. Imag. Sci.*, vol. 11, no. 1, pp. 252–292, 2018.

[19] A. Aghasi, A. Ahmed, P. Hand, and B. Joshi, "A convex program for bilinear inversion of sparse vectors," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 8548–8558.

[20] M. S. Asif, W. Mantzel, and J. Romberg, "Random channel coding and blind deconvolution," in *Proc. IEEE 47th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, 2009, pp. 1021–1025.

[21] Y. Xie, M. B. Wakin, and G. Tang, "Sparse recovery and non-stationary blind demodulation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5566–5570.

[22] Y. Nardi *et al.*, "On the asymptotic properties of the group lasso estimator for linear models," *Electron. J. Statist.*, vol. 2, pp. 605–633, 2008.

[23] X. Lv, G. Bi, and C. Wan, "The group lasso for stable recovery of block-sparse signal representations," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1371–1382, Apr. 2011.

[24] J. Huang *et al.*, "The benefit of group sparsity," *Ann. Statist.*, vol. 38, no. 4, pp. 1978–2004, 2010.

[25] V. Sivakumar, A. Banerjee, and P. K. Ravikumar, "Beyond sub-gaussian measurements: High-dimensional structured estimation with sub-exponential designs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2206–2214.

[26] S. Vaiter, M. Golbabaee, J. Fadili, and G. Peyré, "Model selection with low complexity priors," *Inf. Inference: A J. IMA*, vol. 4, no. 3, pp. 230–287, 2015.

[27] S. Vaiter, G. Peyré, and J. Fadili, "Model consistency of partly smooth regularizers," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1725–1737, Mar. 2018.

[28] C.-Y. Hung and M. Kaveh, "Low rank matrix recovery for joint array self-calibration and sparse model doa estimation," in *Proc. IEEE 7th Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, 2017, pp. 1–5.

[29] Y. C. Eldar, W. Liao, and S. Tang, "Sensor calibration for off-the-grid spectral estimation," *Appl. Comput. Harmon. Anal.*, vol. 48, no. 2, pp. 570–598, 2020.

[30] A. Flinth, "Sparse blind deconvolution and demixing through $\ell_{1,2}$-minimization," *Adv. Comput. Math.*, vol. 44, no. 1, pp. 1–21, 2018.

[31] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning With Sparsity: The Lasso and Generalizations*. London, U.K.: Chapman & Hall, 2015.

[32] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[33] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Cambridge, MA, USA: Birkhäuser, 2013.

[34] D. Hsu, S. M. Kakade, and T. Zhang, "A tail inequality for quadratic forms of subgaussian random vectors," *Electron. Commun. Probability*, vol. 17, pp. 1–6, 2012.

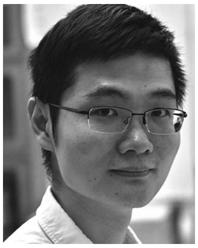
[35] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1302–1338, 2000.

[36] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," Mar. 2014. [Online]. Available: http://cvxr.com/cvx

[37] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm)," *Nature Methods*, vol. 3, no. 10, 2006, Art. no. 793.

[38] EPFL-Biomedical-Imaging-Group, "Single-molecule localization microscopy," 2013. [Online]. Available: http://bigwww.epfl.ch/smlm/

[39] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.



Youye Xie (Student Member, IEEE) received the B.Eng. (Hons.) degree in electronic engineering from The Hong Kong Polytechnic University, Hong Kong and the B.Eng. degree in microelectronics from Sun Yat-sen University, Guangzhou, China, in 2015. He is currently working toward the Ph.D. degree with the Colorado School of Mines, Golden, CO, USA. His research interests include signal processing, machine learning, and computer vision.

Michael B. Wakin (Senior Member, IEEE) received the B.S. degree in electrical engineering and the B.A. (*summa cum laude*) degree in mathematics, in 2000, the M.S. degree in electrical engineering, in 2002, and the Ph.D. degree in electrical engineering, in 2007, all from Rice University, Houston, TX, USA. He was an National Science Foundation (NSF) Mathematical Sciences Postdoctoral Research Fellow with Caltech from 2006 to 2007 and an Assistant Professor with the University of Michigan from 2007 to 2008. He is currently a Professor with the Department of Electrical Engineering, the Colorado School of Mines, Golden, CO, USA. His research interests include sparse, geometric, and manifold-based models for signal processing and compressive sensing. He is an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and was previously an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS. He was the recipient of the Hershel M. Rich Invention Award, in 2007, from Rice University for the design of a single-pixel camera based on compressive sensing, the DARPA Young Faculty Award, in 2008, for his research in compressive multisignal processing for environments, such as sensor and camera networks, the NSF CAREER Award, in 2012, for research into dimensionality reduction techniques for structured data sets, the Excellence in Research Award, in 2014, for his research as a junior faculty member at CSM, and the Best Paper Award from the IEEE Signal Processing Society, in 2015.

Gongguo Tang (Member, IEEE) received the Ph.D. degree in electrical engineering from Washington University, St. Louis, MO, USA, in 2011. He is an Assistant Professor with the Colorado School of Mines, Golden, CO, USA. From 2011 to 2013, he was a Postdoctoral Researcher with the University of Wisconsin-Madison and a Visiting Scholar to the Big Data program with the Simons Institute, University of California, Berkeley, CA, USA. His research revolves around modeling and optimization to extract information from data through computation. He is especially interested in the design of learning models, optimization formulations, and numerical procedures that come with theoretical performance guarantees and are scalable to large datasets, with target applications in signal processing, machine learning, and imaging.