Exact Joint Sparse Frequency Recovery via Optimization Methods

Zai Yang, *Member, IEEE*, and Lihua Xie, *Fellow, IEEE*

arXiv:1405.6585v2 [cs.IT] 30 May 2016

***Abstract*—Frequency recovery/estimation from discrete samples of superimposed sinusoidal signals is a classic yet important problem in statistical signal processing. Its research has recently been advanced by atomic norm techniques which exploit signal sparsity, work directly on continuous frequencies, and completely resolve the grid mismatch problem of previous compressed sensing methods. In this work we investigate the frequency recovery problem in the presence of multiple measurement vectors (MMVs) which share the same frequency components, termed as joint sparse frequency recovery and arising naturally from array processing applications. To study the advantage of MMVs, we first propose an $\ell_{2,0}$ norm like approach by exploiting joint sparsity and show that the number of recoverable frequencies can be increased except in a trivial case. While the resulting optimization problem is shown to be rank minimization that cannot be practically solved, we then propose an MMV atomic norm approach that is a convex relaxation and can be viewed as a continuous counterpart of the $\ell_{2,1}$ norm method. We show that this MMV atomic norm approach can be solved by semidefinite programming. We also provide theoretical results showing that the frequencies can be exactly recovered under appropriate conditions. The above results either extend the MMV compressed sensing results from the discrete to the continuous setting or extend the recent super-resolution and continuous compressed sensing framework from the single to the multiple measurement vectors case. Extensive simulation results are provided to validate our theoretical findings and they also imply that the proposed MMV atomic norm approach can improve the performance in terms of reduced number of required measurements and/or relaxed frequency separation condition.**

***Index Terms*—Atomic norm, compressed sensing, direction of arrival (DOA) estimation, joint sparse frequency recovery, multiple measurement vectors (MMVs).**

I. INTRODUCTION

Suppose that we observe uniform samples (with the Nyquist sampling rate) of a number of L sinusoidal signals:

$$y_{jt}^o = \sum_{k=1}^{K} s_{kt} e^{i2\pi j f_k}, \quad (j,t) \in \boldsymbol{J} \times [L], \qquad (1)$$

which form an $N \times L$ matrix $\boldsymbol{Y}^o = \left[y_{jt}^o \right]$, on the index set $\boldsymbol{\Omega} \times [L]$, where $\boldsymbol{\Omega} \subset \boldsymbol{J} := \{0, 1, \ldots, N-1\}$, $[L] := \{1, 2, \ldots, L\}$, and N is the number of uniform samples per sinusoidal signal.

This means that each sinusoidal signal corresponds to one measurement vector. Here (j, t) indexes the entries of $\boldsymbol{Y}^o$, $i = \sqrt{-1}$, $f_k \in \mathbb{T} := [0, 1]$ denotes the kth normalized frequency (note that the starting point 0 and the ending point 1 of the unit circle $\mathbb{T}$ are identical), $s_{kt} \in \mathbb{C}$ is the (complex) amplitude of the kth frequency component composing the tth sinusoidal signal, and K is the number of the components which is small but unknown. Moreover, let $M = |\boldsymbol{\Omega}| \leq N$ be the sample size of each measurement vector. The observed $M \times L$ data matrix $\boldsymbol{Y}_{\boldsymbol{\Omega}}^o := \left\{ y_{jt}^o \right\}_{(j,t) \in \boldsymbol{\Omega} \times [L]}$ are referred to as full data, if $M = N$ (i.e., $\boldsymbol{\Omega} = \boldsymbol{J}$ and $\boldsymbol{Y}_{\boldsymbol{\Omega}}^o = \boldsymbol{Y}^o$), and otherwise, compressive data. Let $\mathcal{T} = \{f_1, \ldots, f_K\}$ denote the set of the frequencies. The problem concerned in this paper is to recover $\mathcal{T}$ given $\boldsymbol{Y}_{\boldsymbol{\Omega}}^o$, which is referred to as joint sparse frequency recovery (JSFR) in the sense that the multiple measurement vectors (MMVs) (i.e., the L columns of $\boldsymbol{Y}_{\boldsymbol{\Omega}}^o$) share the same K frequencies. Once $\mathcal{T}$ is obtained, the amplitudes $\{s_{kt}\}$ and the full data $\boldsymbol{Y}^o$ can be easily obtained by a simple least-squares method.

An application of the JSFR problem is direction of arrival (DOA) estimation in array processing [2], [3]. In particular, suppose that K farfield, narrowband sources impinge on a linear array of sensors and one wants to know their directions. The output of the sensor array can be modeled by (1) under appropriate conditions, where each frequency corresponds to one source's direction. The sampling index set $\boldsymbol{\Omega}$ therein represents the geometry of the sensor array. To be specific, $\boldsymbol{\Omega} = \boldsymbol{J}$ in the full data case corresponds to an N-element uniform linear array (ULA) with adjacent sensors spaced by half a wavelength, while $\boldsymbol{\Omega} \subsetneq \boldsymbol{J}$ corresponds to a sparse linear array (SLA) that can be obtained by retaining only the sensors of the above ULA indexed by $\boldsymbol{\Omega}$. Each measurement vector consists of the outputs of the sensor array at one snapshot. The L MMVs are obtained by taking L snapshots under the assumption of static sources (during a time window). Note that, since the array size can be limited in practice due to physical constraints and/or cost considerations, it is crucial in DOA estimation to exploit the temporal redundancy (a.k.a., the joint sparsity that we refer to) contained in the MMVs.

In conventional methods for JSFR one usually assumes that the source signals (or the rows of $[s_{kt}]$) have zero mean and are spatially uncorrelated. It follows that the covariance matrix of the full data snapshot (or the columns of $\boldsymbol{Y}^o$) is positive semidefinite (PSD), Toeplitz and low rank (of rank K). Exploiting these structures for frequency recovery was firstly proposed by Pisarenko who rediscovered the classical Vandermonde decomposition lemma that states that the frequencies

can be exactly retrieved from the data covariance matrix [4], [5]. A prominent class of methods was then proposed and designated as subspace methods such as MUSIC and ESPRIT [6], [7]. While these methods estimate the data covariance using the sample covariance, the Toeplitz structure cannot be exploited in general, a sufficient number of snapshots is required, and their performance can be degraded in the presence of source correlations.

With the development of sparse signal representation and later the compressed sensing (CS) concept [8], [9], sparse (for $L = 1$) and joint sparse (for $L > 1$) methods for frequency recovery have been popular in the past decade. In these methods, however, the frequencies of interest are usually assumed to lie on a fixed grid on $\mathbb{T}$ because the development of CS so far has been focused on signals that can be sparsely represented under a finite discrete dictionary. Under the on-grid assumption, the observation model in (1) can be written into an underdetermined system of linear equations and CS methods are applied to solve an involved sparse signal whose support is finally identified as the frequency set $\mathcal{T}$. Typical sparse methods include combinatorial optimization or ℓ_0 (pseudo-)norm minimization, its convex relaxation or ℓ_1 norm minimization, and greedy methods such as orthogonal matching pursuit (OMP) as well as their joint sparse counterparts [10]–[15]. While the ℓ_0 minimization can exploit sparsity to the greatest extent possible, it is NP-hard and cannot be practically solved. The maximal K allowed in ℓ_1 minimization and OMP for guaranteed exact recovery is inversely proportional to a metric called coherence which, however, increases dramatically as the grid gets fine. Moveover, grid mismatches become a major problem of CS-based methods though several modifications have been proposed to alleviate this drawback (see, e.g., [16]–[19]).

Breakthroughs came out recently. In the single measurement vector (SMV) case when $L = 1$, Candès and Fernandez-Granda [20] dealt directly with continuous frequencies and completely resolved the grid mismatch problem. In particular, they considered the full data case and showed that the frequencies can be exactly recovered by exploiting signal sparsity if all the frequencies are mutually separated by at least $\frac{4}{N}$. This means that up to $K = \frac{N}{4}$ frequencies can be recovered. Their method is based on the total variation norm or the atomic norm that extends the ℓ_1 norm from the discrete to the continuous frequency case and can be computed using semidefinite programming (SDP) [21], [22]. Following from [20], Tang *et al.* [23] studied the same problem in the case of compressive data using atomic norm minimization (ANM). Under the same frequency separation condition, they showed that a number of $M \geq O(K \log K \log N)$ randomly selected samples is sufficient to guarantee exact recovery with high probability. Several subsequent papers on this topic include [24]–[29]. However, similar *gridless sparse* methods are rare for JSFR in the MMV case concerned in this paper. A gridless method designated as the sparse and parametric approach (SPA) was proposed in our previous work [30] based on weighted covariance fitting by exploiting the structures of the data covariance matrix. In the main context of this paper we will show that this method is closely related to the MMV

atomic norm method that we will introduce in the present paper. Another related work is [31]; however, in this paper the MMV problem was reformulated as an SMV one, with the joint sparsity missing, and solved within the framework in [20]. Therefore, the frequency recovery performance can be degraded. As an example, in the noiseless case the frequencies cannot be exactly recovered using the method in [31] due to some new 'noise' term introduced.

In this paper, we first study the advantage of exploiting joint sparsity in the MMVs and then propose a practical approach to utilize this information. In particular, following from the literature on CS we propose an ℓ_0 norm like sparse metric that is referred to as the MMV atomic ℓ_0 norm and is a continuous counterpart of the $\ell_{2,0}$ norm used for joint sparse recovery [13]. We theoretically show that the MMVs can help improve the frequency recovery performance in terms of the number of recoverable frequencies except in a trivial case. But unfortunately (in fact, not surprisingly), this atomic ℓ_0 norm approach is proven to be a rank minimization problem that cannot be practically solved. We then propose a convex relaxation approach in which the MMV atomic norm is adopted that is a continuous counterpart of the $\ell_{2,1}$ norm. We show that this atomic norm approach can be efficiently solved via semidefinite programming. Theoretical results are also provided to show that the frequencies can be exactly recovered under similar conditions as in [20], [23]. Extensive simulation results are provided to validate our theoretical results and they also imply that the proposed MMV atomic norm approach can result in improved frequency recovery performance in terms of reduced number of required measurements and/or relaxed frequency separation condition.

It is interesting to note that the proposed MMV atomic ℓ_0 norm and atomic norm approaches somehow exploit the structures of the "data covariance matrix" and are related to the aforementioned subspace methods. In particular, a PSD Toeplitz matrix is involved in both the proposed methods that can be interpreted as the data covariance matrix (as if certain statistical assumptions were satisfied) from the Vandermonde decomposition of which the true frequencies are finally obtained, while the low rank structure is exploited by matrix rank minimization in the atomic ℓ_0 norm method and by matrix trace norm (or nuclear norm) minimization in the atomic norm method. As compared to the subspace methods, the proposed methods exploit the matrix structures to a greater extent. Moreover, the proposed methods do not require the assumption of uncorrelated sources and can be applied to the case of limited measurement vectors.

The results of this work were published online in the technical report [32] and were presented in part in the conference paper [1]. When preparing this paper we found that the same MMV atomic norm approach was also independently proposed in [33], [34]. This paper is different form [33], [34] in the following aspects. First, in this paper, the advantage of MMVs is theoretically proven in terms of the number of recoverable frequencies based on the proposed MMV atomic ℓ_0 norm approach, while no such theoretical results are provided in [33], [34]. Second, in this paper, the SDP formulation of the MMV atomic norm is proven inspired by our previous work

[30], while the proof in [33], [34] is given following [23] on the SMV case. Finally, as pointed out in [34], the theoretical guarantee of the MMV atomic norm approach provided in [34, Theorem 2] is weaker than ours (see Theorem 5; note that the technical report [32] appeared online earlier than [34]).

Notations used in this paper are as follows. $\mathbb{R}$ and $\mathbb{C}$ denote the sets of real and complex numbers respectively. $\mathbb{T}$ denotes the unit circle $[0, 1]$ by identifying the starting and ending points. Boldface letters are reserved for vectors and matrices. For an integer L, $[L] := \{1, \cdots, L\}$. $|\cdot|$ denotes the amplitude of a scalar or the cardinality of a set. $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the ℓ_1, ℓ_2 and Frobenius norms respectively. $\boldsymbol{A}^T$ and $\boldsymbol{A}^H$ are the matrix transpose and conjugate transpose of $\boldsymbol{A}$ respectively. x_j is the jth entry of a vector $\boldsymbol{x}$, and $\boldsymbol{A}_j$ denotes the jth row of a matrix $\boldsymbol{A}$. Unless otherwise stated, $\boldsymbol{x}_\Omega$ and $\boldsymbol{A}_\Omega$ are subvector and submatrix of $\boldsymbol{x}$ and $\boldsymbol{A}$ respectively by retaining the entries of $\boldsymbol{x}$ and the rows of $\boldsymbol{A}$ indexed by the set Ω. For a vector $\boldsymbol{x}$, $\mathrm{diag}(\boldsymbol{x})$ is a diagonal matrix with $\boldsymbol{x}$ on the diagonal. $\boldsymbol{x} \succeq \boldsymbol{0}$ means $x_j \geq 0$ for all j. $\mathrm{rank}(\boldsymbol{A})$ denotes the rank of a matrix $\boldsymbol{A}$ and $\mathrm{tr}(\boldsymbol{A})$ the trace. For positive semidefinite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \geq \boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite. $\mathbb{E}[\cdot]$ denotes the expectation and $\mathbb{P}(\cdot)$ the probability of an event.

The rest of the paper is organized as follows. Section II presents the main results of this paper. Section III discusses connections between the proposed methods and prior art. Section IV presents proofs of the main results in Section II. Section V provides numerical simulations and Section VI concludes this paper.

II. MAIN RESULTS

This section presents the main results of this paper whose proofs will be given in Section IV.

A. Preliminary: Vandermonde Decomposition

The Vandermonde decomposition of Toeplitz matrices can date back to 1910s and has been important in the signal processing society since its rediscovery and use for frequency estimation in 1970s [4], [5] (see also [3]). In particular, it states that any PSD, rank-$K \leq N$, Toeplitz matrix $T(\boldsymbol{u}) \in \mathbb{C}^{N \times N}$, which is parameterized by $\boldsymbol{u} \in \mathbb{C}^N$ and given by

$$T(\boldsymbol{u}) = \begin{bmatrix} u_1 & u_2 & \cdots & u_N \\ u_2^H & u_1 & \cdots & u_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ u_N^H & u_{N-1}^H & \cdots & u_1 \end{bmatrix}, \quad (2)$$

can be decomposed as

$$T(\boldsymbol{u}) = \sum_{k=1}^{K} p_k \boldsymbol{a}(f_k) \boldsymbol{a}^H(f_k) = \boldsymbol{A}(\boldsymbol{f}) \boldsymbol{P} \boldsymbol{A}^H(\boldsymbol{f}), \quad (3)$$

where $\boldsymbol{A}(\boldsymbol{f}) = [\boldsymbol{a}(f_1), \ldots, \boldsymbol{a}(f_K)] \in \mathbb{C}^{N \times K}$ with $\boldsymbol{a}(f) = [1, e^{i2\pi f}, \ldots, e^{i2\pi(N-1)f}]^T \in \mathbb{C}^N$, $\boldsymbol{P} = \mathrm{diag}(p_1, \ldots, p_K)$ with $p_k > 0$, $k = 1, \ldots, K$ and $\{f_k\}$ are distinct points in $\mathbb{T}$. Moreover, the decomposition in (3) is unique if $K < N$. Note that the name 'Vandermonde' comes from the fact that $\boldsymbol{A}(\boldsymbol{f})$ is a Vandermonde matrix.

It is well known that under the assumption of uncorrelated sources the data covariance matrix (i.e., the covariance matrix of each column of $\boldsymbol{Y}^o$) is a rank-K, PSD, Toeplitz matrix. Therefore, the Vandermonde decomposition actually says that the frequencies can be uniquely obtained from the data covariance matrix given $K < N$ [5]. Note that a subspace method such as ESPRIT can be used to compute the decomposition in (3).

B. Frequency Recovery Using Joint Sparsity

To exploit the joint sparsity in the MMVs, we let $\boldsymbol{s}_k = [s_{k1}, \cdots, s_{kL}] \in \mathbb{C}^{1 \times L}$. It follows that (1) can be written as

$$\boldsymbol{Y}^o = \sum_{k=1}^{K} \boldsymbol{a}(f_k) \boldsymbol{s}_k = \sum_{k=1}^{K} c_k \boldsymbol{a}(f_k) \boldsymbol{\phi}_k, \quad (4)$$

where $\boldsymbol{a}(f)$ is as defined in (3), $c_k = \|\boldsymbol{s}_k\|_2 > 0$ and $\boldsymbol{\phi}_k = c_k^{-1} \boldsymbol{s}_k$ with $\|\boldsymbol{\phi}_k\|_2 = 1$. Let $\mathbb{S}^{2L-1} = \{\boldsymbol{\phi} \in \mathbb{C}^{1 \times L} : \|\boldsymbol{\phi}\|_2 = 1\}$ denote the unit complex $(L-1)$-sphere (or real $(2L-1)$-sphere). Define the set of atoms

$$\mathcal{A} := \{\boldsymbol{a}(f, \boldsymbol{\phi}) = \boldsymbol{a}(f) \boldsymbol{\phi} : f \in \mathbb{T}, \boldsymbol{\phi} \in \mathbb{S}^{2L-1}\}. \quad (5)$$

It follows from (4) that $\boldsymbol{Y}^o$ is a linear combination of K atoms in $\mathcal{A}$. In particular, we say that a decomposition of $\boldsymbol{Y}^o$ as in (4) is an atomic decomposition of order K if $c_k > 0$ and the frequencies f_k are distinct.

Following from the literature on CS, we first propose an (MMV) atomic ℓ_0 norm approach to signal and frequency recovery that exploits sparsity to the greatest extent possible. In particular, the atomic ℓ_0 norm of $\boldsymbol{Y} \in \mathbb{C}^{N \times L}$ is defined as the smallest number of atoms in $\mathcal{A}$ that can express $\boldsymbol{Y}$:

$$\|\boldsymbol{Y}\|_{\mathcal{A},0} = \inf \left\{ \mathcal{K} : \boldsymbol{Y} = \sum_{k=1}^{\mathcal{K}} c_k \boldsymbol{a}_k, \boldsymbol{a}_k \in \mathcal{A}, c_k > 0 \right\}. \quad (6)$$

The following optimization method is proposed for signal recovery that generalizes a method in [23] from the SMV to the MMV case:

$$\min_{\boldsymbol{Y}} \|\boldsymbol{Y}\|_{\mathcal{A},0}, \quad \text{subject to } \boldsymbol{Y}_\Omega = \boldsymbol{Y}_\Omega^o. \quad (7)$$

The frequencies composing the solution of $\boldsymbol{Y}$ are the frequency estimates.

To show the advantage of MMVs, we define the continuous dictionary

$$\mathcal{A}_\Omega^1 := \{\boldsymbol{a}_\Omega(f) : f \in \mathbb{T}\} \quad (8)$$

and then define the spark of $\mathcal{A}_\Omega^1$, denoted by $\mathrm{spark}(\mathcal{A}_\Omega^1)$, as the smallest number of atoms in $\mathcal{A}_\Omega^1$ that are linearly dependent. Note that this definition of spark generalizes that in [35] from the discrete to the continuous dictionary case. We have the following theoretical guarantee for (7).

Theorem 1. $\boldsymbol{Y}^o = \sum_{j=1}^{K} c_j \boldsymbol{a}(f_j, \boldsymbol{\phi}_j)$ is the unique optimizer to (7) if

$$K < \frac{\mathrm{spark}(\mathcal{A}_\Omega^1) - 1 + \mathrm{rank}(\boldsymbol{Y}_\Omega^o)}{2}. \quad (9)$$

Moreover, the atomic decomposition above is the unique one satisfying that $K = \|\boldsymbol{Y}^o\|_{\mathcal{A},0}$.

By Theorem 1 the frequencies can be exactly recovered using the atomic ℓ_0 norm approach if the sparsity K is sufficiently small with respect to the sampling index set $\boldsymbol{\Omega}$ and the observed data $\boldsymbol{Y}_{\boldsymbol{\Omega}}^o$. Note that the number of recoverable frequencies can be increased, as compared to the SMV case, if $\text{rank}\left(\boldsymbol{Y}_{\boldsymbol{\Omega}}^o\right) > 1$, which happens except in a trivial case when the MMVs in $\boldsymbol{Y}_{\boldsymbol{\Omega}}^o$ are identical up to scaling factors.

But unfortunately, the following result shows that $\|\boldsymbol{Y}\|_{\mathcal{A},0}$ is substantially a rank minimization problem that cannot be practically solved.

Theorem 2. $\|\boldsymbol{Y}\|_{\mathcal{A},0}$ *defined in (6) equals the optimal value of the following rank minimization problem:*

$$\min_{\boldsymbol{W},\boldsymbol{u}} \text{rank}\left(T\left(\boldsymbol{u}\right)\right), \text{ subject to } \begin{bmatrix} \boldsymbol{W} & \boldsymbol{Y}^H \\ \boldsymbol{Y} & T\left(\boldsymbol{u}\right) \end{bmatrix} \geq \boldsymbol{0}. \quad (10)$$

It immediately follows from (10) that (7) can be cast as the following rank minimization problem:

$$\min_{\boldsymbol{Y},\boldsymbol{W},\boldsymbol{u}} \text{rank}\left(T\left(\boldsymbol{u}\right)\right),$$
$$\text{subject to } \begin{bmatrix} \boldsymbol{W} & \boldsymbol{Y}^H \\ \boldsymbol{Y} & T\left(\boldsymbol{u}\right) \end{bmatrix} \geq \boldsymbol{0} \text{ and } \boldsymbol{Y}_{\boldsymbol{\Omega}} = \boldsymbol{Y}_{\boldsymbol{\Omega}}^o. \quad (11)$$

Suppose that (11) can be globally solved and let $\boldsymbol{u}^*$ and $\boldsymbol{Y}^*$ denote the solutions of $\boldsymbol{u}$ and $\boldsymbol{Y}$, respectively. If the condition of Theorem 1 is satisfied, then $\boldsymbol{Y}^o = \boldsymbol{Y}^*$ and the frequencies as well as the atomic decomposition of $\boldsymbol{Y}^o$ in Theorem 1 can be computed accordingly. In particular, it is guaranteed that $\text{rank}\left(T\left(\boldsymbol{u}^*\right)\right) = K < N$ (see the proof in Section IV-A). It follows that the true frequencies can be uniquely obtained from the Vandermonde decomposition of $T\left(\boldsymbol{u}^*\right)$. After that, the atomic decomposition of $\boldsymbol{Y}^o$ can be obtained by the fact that $\boldsymbol{Y}^*$ lies in the range space of $T\left(\boldsymbol{u}^*\right)$. Moreover, it is worth noting that, although (7) has a trivial solution in the full data case, the problem in (11) still makes sense and the frequency retrieval process also applies.

C. Frequency Recovery via Convex Relaxation

While the rank minimization problem in (11) is nonconvex and cannot be globally solved with a practical algorithm, it motivates the (MMV) atomic norm method—a convex relaxation. In particular, the atomic norm of $\boldsymbol{Y} \in \mathbb{C}^{N \times L}$ is defined as the gauge function of $\text{conv}\left(\mathcal{A}\right)$, the convex hull of $\mathcal{A}$ [22]:

$$\|\boldsymbol{Y}\|_{\mathcal{A}} := \inf \left\{t > 0 : \boldsymbol{Y} \in t\text{conv}\left(\mathcal{A}\right)\right\}$$
$$= \inf \left\{\sum_k c_k : \boldsymbol{Y} = \sum_k c_k \boldsymbol{a}_k, c_k > 0, \boldsymbol{a}_k \in \mathcal{A}\right\}, \quad (12)$$

in which the joint sparsity is exploited in a different manner. Indeed, $\|\cdot\|_{\mathcal{A}}$ is a norm by the property of the gauge function and thus it is convex. Corresponding to (7), we propose the following convex optimization problem:

$$\min_{\boldsymbol{Y}} \|\boldsymbol{Y}\|_{\mathcal{A}}, \text{ subject to } \boldsymbol{Y}_{\boldsymbol{\Omega}} = \boldsymbol{Y}_{\boldsymbol{\Omega}}^o. \quad (13)$$

Though we know that (13) is convex, (13) still cannot be practically solved since by (12) it is a semi-infinite program with an infinite number of variables. To practically solve (13), an SDP formulation of $\|\boldsymbol{Y}\|_{\mathcal{A}}$ is provided in the following theorem.

Theorem 3. $\|\boldsymbol{Y}\|_{\mathcal{A}}$ *defined in (12) equals the optimal value of the following SDP:*

$$\min_{\boldsymbol{W},\boldsymbol{u}} \frac{1}{2\sqrt{N}} \left[tr\left(\boldsymbol{W}\right) + tr\left(T\left(\boldsymbol{u}\right)\right)\right],$$
$$\text{subject to } \begin{bmatrix} \boldsymbol{W} & \boldsymbol{Y}^H \\ \boldsymbol{Y} & T\left(\boldsymbol{u}\right) \end{bmatrix} \geq \boldsymbol{0}. \quad (14)$$

By Theorem 3, (13) can be cast as the following SDP which can be solved using an off-the-shelf SDP solver:

$$\min_{\boldsymbol{Y},\boldsymbol{W},\boldsymbol{u}} \text{tr}\left(\boldsymbol{W}\right) + \text{tr}\left(T\left(\boldsymbol{u}\right)\right),$$
$$\text{subject to } \begin{bmatrix} \boldsymbol{W} & \boldsymbol{Y}^H \\ \boldsymbol{Y} & T\left(\boldsymbol{u}\right) \end{bmatrix} \geq \boldsymbol{0} \text{ and } \boldsymbol{Y}_{\boldsymbol{\Omega}} = \boldsymbol{Y}_{\boldsymbol{\Omega}}^o. \quad (15)$$

Given the optimal solution $\boldsymbol{u}^*$ to (15), the frequencies and the atomic decomposition of $\boldsymbol{Y}^o$ can be computed as previously based on the Vandermonde decomposition of $T\left(\boldsymbol{u}^*\right)$.

Finally, we analyze the theoretical performance of the atomic norm approach. To do so, we define the minimum separation of a finite subset $\mathcal{T} \subset \mathbb{T}$ as the closest wrap-around distance between any two elements,

$$\Delta_{\mathcal{T}} = \inf_{a,b \in \mathcal{T}:a \neq b} \min \left\{|a - b|, 1 - |a - b|\right\}.$$

We first study the full data case that, as we will see, forms the basis of the compressive data case. Note that (15) can be solved for frequency recovery though (13) admits a trivial solution. We have the following theoretical guarantee.

Theorem 4. $\boldsymbol{Y}^o = \sum_{j=1}^{K} c_j \boldsymbol{a}\left(f_j, \phi_j\right)$ *is the unique atomic decomposition satisfying that* $\|\boldsymbol{Y}^o\|_{\mathcal{A}} = \sum_{j=1}^{K} c_j$ *if* $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor(N-1)/4\rfloor}$ *and* $N \geq 257$.[1]

In the compressive data case, the following result holds.

Theorem 5. *Suppose we observe* $\boldsymbol{Y}^o = \sum_{j=1}^{K} c_j \boldsymbol{a}\left(f_j, \phi_j\right)$ *on the index set* $\boldsymbol{\Omega} \times [L]$, *where* $\boldsymbol{\Omega} \subset \boldsymbol{J}$ *is of size* M *and selected uniformly at random. Assume that* $\left\{\phi_j\right\}_{j=1}^{K} \subset \mathbb{S}^{2L-1}$ *are independent random variables with* $\mathbb{E}\phi_j = \boldsymbol{0}$. *If* $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor(N-1)/4\rfloor}$, *then there exists a numerical constant* C *such that*

$$M \geq C \max \left\{\log^2 \frac{\sqrt{L}N}{\delta}, K \log \frac{K}{\delta} \log \frac{\sqrt{L}N}{\delta}\right\} \quad (16)$$

is sufficient to guarantee that, with probability at least $1 - \delta$, $\boldsymbol{Y}^o$ *is the unique optimizer to (13) and* $\boldsymbol{Y}^o = \sum_{j=1}^{K} c_j \boldsymbol{a}\left(f_j, \phi_j\right)$ *is the unique atomic decomposition satisfying that* $\|\boldsymbol{Y}^o\|_{\mathcal{A}} = \sum_{j=1}^{K} c_j$.

---

[1] The condition $N \geq 257$ is more like a technical requirement but not an obstacle in practice (see numerical simulations in Section V).

D. Discussions

We have proposed two optimization approaches to JSFR by exploiting the joint sparsity in the MMVs. Based on the atomic ℓ_0 norm approach, we theoretically show that the MMVs help improve the frequency recovery performance in terms of the number of recoverable frequencies. But unfortunately, the resulting optimization problem is NP-hard to solve. We therefore turn to the atomic norm approach and show that this convex relaxation approach can be cast as SDP and solved in a polynomial time. We also provide theoretical results showing that the atomic norm approach can successfully recover the frequencies under similar technical conditions as in [20], [23].

At a first glance, both the methods can be viewed as covariance-based by exploiting the structures of the data covariance matrix (obtained as if certain statistical assumptions for the source signals were satisfied). In particular, in both (11) and (15), the PSD Toeplitz matrix $T(\boldsymbol{u})$, which can be written as in (3), can be viewed as the covariance matrix of the full data candidate $\boldsymbol{Y}$ that is consistent with the observed data $\boldsymbol{Y}_{\boldsymbol{\Omega}}^o$ (see more details in the proofs of Theorems 2 and 3 in Section IV). The Toeplitz structure is explicitly given, the PSD is imposed by the first constraint, and the low rank is exploited in the objective function. The essential difference between the two methods lies in the way to exploit the low rank. To be specific, the atomic ℓ_0 norm method utilizes this structure to the greatest extent possible by directly minimizing the rank, leading to a nonconvex optimization problem. In contrast, the atomic norm method uses convex relaxation and minimizes the nuclear norm or the trace norm of the matrix (note that the additional term $\text{tr}(\boldsymbol{W})$ in (15) helps control the magnitude of $\boldsymbol{u}$ and avoids a trivial solution). As a result, the theoretical guarantees that we provide actually state that the full data covariance matrix can be exactly recovered using the proposed methods given full or compressive data under certain conditions. Finally, note that source correlations in $[s_{kt}]$, if present, will be removed in the covariance estimate $T(\boldsymbol{u})$ in both (11) and (15), whereas they will be retained in the sample covariance used in conventional subspace methods.

The theoretical results presented above extend several existing results from the SMV to the MMV case or from the discrete to the continuous setting. To be specific, Theorem 1 is a continuous counterpart of [13, Theorem 2.4] which deals with the conventional discrete setting. Theorem 1 shows that the number of recoverable frequencies can be increased in general as we take MMVs. This is practically relevant in array processing applications. But in a trivial case where all the sources are coherent, i.e., all the rows of $[s_{kt}]$ (and thus all the columns of $\boldsymbol{Y}_{\boldsymbol{\Omega}}^o$) are identical up to scaling factors, it holds that $\text{rank}(\boldsymbol{Y}_{\boldsymbol{\Omega}}^o) = 1$ as in the SMV case and hence, as expected, MMVs do not help improve the performance. Note also that it is generally difficult to compute $\text{spark}(\mathcal{A}_{\boldsymbol{\Omega}}^1)$, except in the full data case where we have $\text{spark}(\mathcal{A}_{\boldsymbol{\Omega}}^1) = N + 1$ by the fact that any N atoms in $\mathcal{A}_{\boldsymbol{\Omega}}^1$ are linear independent. An interesting topic in future studies will be the selection of $\boldsymbol{\Omega}$, which in array processing corresponds to geometry design of the sensor array, such that $\text{spark}(\mathcal{A}_{\boldsymbol{\Omega}}^1)$ is maximized.

Theorem 4 generalizes [20, Theorem 1.2] from the SMV

to the MMV case. Since Theorem 4 applies to all kinds of source signals, including the aforementioned trivial case, one cannot expect that the theoretical guarantee improves in the MMV case.

Theorem 5 generalizes [23, Theorem I.1] from the SMV to the MMV case. Note that in (16) the dependence of M on L is for controlling the probability of successful recovery. To make it clear, we consider the case when we seek to recover the columns of $\boldsymbol{Y}^o$ independently via the SMV method in [23]. When M satisfies (16) with $L = 1$, each column of $\boldsymbol{Y}^o$ can be recovered with probability $1 - \delta$. It follows that $\boldsymbol{Y}^o$ can be exactly recovered with probability at least $1 - L\delta$. In contrast, if we recover $\boldsymbol{Y}^o$ via a single convex optimization problem that we propose, then with the same number of measurements the success probability is improved to $1 - \sqrt{L}\delta$ (to see this, replace δ in (16) by $\sqrt{L}\delta$).

We note that in Theorem 5 the assumption on the phases ϕ_j is relaxed as compared to that in [23, Theorem I.1] (note that ϕ_j's are assumed in the latter drawn i.i.d. from a uniform distribution). This relaxation is significant in array processing since each ϕ_j corresponds to one source and therefore they do not necessarily obey an identical distribution. Note also that this assumption is weak in the sense that the sources can be coherent, resulting in the aforementioned trivial case. To see this, suppose that the rows of $[s_{kt}]$ are i.i.d. Gaussian with zero mean and covariance of rank one. Then the sources are certain to be independent and coherent. This explains why the theoretical guarantee given in Theorem 5 does not improve in the presence of MMVs. In this sense, therefore, the results of Theorems 4 and 5 are referred to as *worst case* analysis.

Our contribution by Theorems 4 and 5 is showing that in the presence of MMVs we can confidently recover the frequencies via a single convex optimization problem by exploiting the joint sparsity therein. Although the worst case analysis we provide cannot shed light on the advantage of MMVs, numerical simulations provided in Section V indeed imply that the proposed atomic norm approach significantly improves the recovery performance when the source signals are at general positions. We pose such *average case* analysis as a future work.

III. CONNECTIONS TO PRIOR ART

A. Grid-based Joint Sparse Recovery

The JSFR problem concerned in this paper has been widely studied within the CS framework, typically under the topic of DOA estimation. It has been popular in the past decade to assume that the true frequencies lie on a fixed grid since, according to conventional wisdom on CS, the signal needs to be sparsely represented under a finite discrete dictionary. Now recall the atomic ℓ_p norm in (6) and (12) with $p = 0$ and 1, respectively, that can be written collectively as:

$$\|\boldsymbol{Y}\|_{\mathcal{A},p} = \inf\left\{\sum_k \|\boldsymbol{s}_k\|_2^p : \boldsymbol{Y} = \sum_k \boldsymbol{a}(f_k)\boldsymbol{s}_k, f_k \in \mathbb{T}\right\}, \tag{17}$$

where $\boldsymbol{s}_k \in \mathbb{C}^{1 \times L}$. Consequently, the atomic ℓ_0 norm and the atomic norm can be viewed, respectively, as the continuous

counterparts of the $\ell_{2,0}$ norm and the $\ell_{2,1}$ norm in grid-based joint sparse recovery methods (see, e.g., [12], [14]). It is worth noting that for the existing grid-based methods one cannot expect exact frequency recovery since in practice the true frequencies typically do not lie on the grid. Moreover, even if this on-grid assumption is satisfied, the existing coherence or RIP-based analysis in the discrete setting is very conservative, as compared to the results in this paper, due to high coherence in the case of a dense grid. Readers are referred to [20] for detailed discussions on the SMV case.

B. Gridless Joint Sparse Recovery

To the best of our knowledge, the only discretization-free/gridless technique for JSFR was introduced in [30] prior to this work, termed as the sparse and parametric approach (SPA). Different from the atomic norm technique proposed in this paper, SPA is from a statistical perspective and based on a weighted covariance fitting criterion [36]. But we show next that the two methods are strongly connected. Consider the full data case as an example. In the limiting noiseless case, SPA solves the following problem:

$$\min_{\boldsymbol{u}\in\mathbb{C}^{N},T(\boldsymbol{u})\geq\boldsymbol{0}} \operatorname{tr}\left(\widehat{\boldsymbol{R}}\left[T\left(\boldsymbol{u}\right)\right]^{-1}\widehat{\boldsymbol{R}}\right) + \operatorname{tr}\left(T\left(\boldsymbol{u}\right)\right), \quad (18)$$

where $\widehat{\boldsymbol{R}} = \frac{1}{L}\boldsymbol{Y}^{o}\boldsymbol{Y}^{oH}$ denotes the sample covariance matrix. Let $\boldsymbol{V} = \frac{1}{L}\left(\boldsymbol{Y}^{oH}\boldsymbol{Y}^{o}\right)^{\frac{1}{2}} \in \mathbb{C}^{L\times L}$. Then we have the following equalities/equivalences:

$$
\begin{aligned}
(18) &= \min_{\boldsymbol{u},T(\boldsymbol{u})\geq\boldsymbol{0}} \operatorname{tr}\left(\left(\boldsymbol{Y}^{o}\boldsymbol{V}\right)^{H}\left[T\left(\boldsymbol{u}\right)\right]^{-1}\boldsymbol{Y}^{o}\boldsymbol{V}\right) + \operatorname{tr}\left(T\left(\boldsymbol{u}\right)\right)\\
&= \min_{\boldsymbol{W},\boldsymbol{u}} \operatorname{tr}\left(\boldsymbol{W}\right) + \operatorname{tr}\left(T\left(\boldsymbol{u}\right)\right),\\
&\quad \text{subject to } \begin{bmatrix} \boldsymbol{W} & \left(\boldsymbol{Y}^{o}\boldsymbol{V}\right)^{H}\\ \boldsymbol{Y}^{o}\boldsymbol{V} & T\left(\boldsymbol{u}\right)\end{bmatrix} \geq \boldsymbol{0}\\
&= 2\sqrt{N}\left\|\boldsymbol{Y}^{o}\boldsymbol{V}\right\|_{\mathcal{A}},
\end{aligned}
$$

where the last equality follows from Theorem 3. This means that SPA actually computes the atomic norm of

$$\boldsymbol{Y}^{o}\boldsymbol{V} = \sum_{k=1}^{K}\boldsymbol{a}\left(f_{k}\right)\left(\boldsymbol{s}_{k}\boldsymbol{V}\right). \quad (19)$$

Therefore, SPA can be interpreted as an atomic norm approach with modification of the source signals. In the SMV case where $\boldsymbol{V}$ is a positive scalar, the two techniques are exactly equivalent, which has been shown in [29]. While details are omitted, note that a similar result holds in the compressive data case.

IV. PROOFS

The proofs of Theorems 1-5 are provided in this section. While our proofs generalize several results in the literature either from the SMV to the MMV case or from the discrete to the continuous setting, note that they are not straightforward. For example, the proof of Theorem 3 does not follow from [23] in the SMV case but is motivated by [29], [30]. The main challenge of the proofs of Theorems 4 and 5 lie in how to

construct and deal with vector-valued dual polynomials instead of the scalar-valued ones in [20] and [23]. Moreover, the proof of Theorem 4 forms the basis of the proof of Theorem 5. Some inaccuracy in [23] is also pointed out and corrected.

A. Proof of Theorem 2

Let $K = \|\boldsymbol{Y}\|_{\mathcal{A},0}$ and $K^* = \operatorname{rank}\left(T\left(\boldsymbol{u}^*\right)\right)$, where $\boldsymbol{u}^*$ denotes an optimal solution of $\boldsymbol{u}$ in (10). It suffices to show that $K = K^*$. On one hand, using the Vandermonde decomposition, we have that $T\left(\boldsymbol{u}^*\right) = \sum_{j=1}^{K^*} p_j \boldsymbol{a}\left(f_j\right)\boldsymbol{a}^{H}\left(f_j\right)$. Moreover, the fact that $\boldsymbol{Y}$ lies in the range space of $T\left(\boldsymbol{u}^*\right)$ implies that there exist $\boldsymbol{s}_j \in \mathbb{C}^{1\times L}$, $j \in \left[K^*\right]$ such that $\boldsymbol{Y} = \sum_{j=1}^{K^*} \boldsymbol{a}\left(f_j\right)\boldsymbol{s}_j$. It follows from the definition of $\|\boldsymbol{Y}\|_{\mathcal{A},0}$ that $K \leq K^*$.

On the other hand, let $\boldsymbol{Y} = \sum_{j=1}^{K}\boldsymbol{a}\left(f_j\right)\boldsymbol{s}_j$ be an atomic decomposition of $\boldsymbol{Y}$. Let $T\left(\boldsymbol{u}\right) = \sum_{j=1}^{K} p_j \boldsymbol{a}\left(f_j\right)\boldsymbol{a}^{H}\left(f_j\right)$ and $\boldsymbol{W} = \sum_{j=1}^{K} p_j^{-1}\boldsymbol{s}_j^{H}\boldsymbol{s}_j$ for arbitrary $p_j > 0$, $j \in [K]$. Then,

$$\begin{bmatrix} \boldsymbol{W} & \boldsymbol{Y}^{H}\\ \boldsymbol{Y} & T\left(\boldsymbol{u}\right)\end{bmatrix} = \sum_{j=1}^{K} p_j \begin{bmatrix} p_j^{-1}\boldsymbol{s}_j^{H}\\ \boldsymbol{a}\left(f_j\right)\end{bmatrix}\begin{bmatrix} p_j^{-1}\boldsymbol{s}_j & \boldsymbol{a}\left(f_j\right)^{H}\end{bmatrix} \geq \boldsymbol{0}.$$

This means that $(\boldsymbol{W}, \boldsymbol{u})$ defines a feasible solution of (10). Consequently, $K^* \leq \operatorname{rank}\left(T\left(\boldsymbol{u}\right)\right) = K$.

B. Proof of Theorem 3

We use the following identity whenever $\boldsymbol{R} \geq \boldsymbol{0}$:

$$\boldsymbol{y}^{H}\boldsymbol{R}^{-1}\boldsymbol{y} = \min t, \text{ subject to } \begin{bmatrix} t & \boldsymbol{y}^{H}\\ \boldsymbol{y} & \boldsymbol{R}\end{bmatrix} \geq 0. \quad (20)$$

In fact, (20) is equivalent to defining $\boldsymbol{y}^{H}\boldsymbol{R}^{-1}\boldsymbol{y} := \lim_{\sigma\to 0_+}\boldsymbol{y}^{H}\left(\boldsymbol{R} + \sigma\boldsymbol{I}\right)^{-1}\boldsymbol{y}$ when $\boldsymbol{R}$ loses rank. We also use the following lemma.

Lemma 1 ([29]). *Given* $\boldsymbol{R} = \boldsymbol{A}\boldsymbol{A}^{H} \geq \boldsymbol{0}$, *it holds that* $\boldsymbol{y}^{H}\boldsymbol{R}^{-1}\boldsymbol{y} = \min\|\boldsymbol{s}\|_2^2$, *subject to* $\boldsymbol{A}\boldsymbol{s} = \boldsymbol{y}$.

Now we prove Theorem 3. It follows from the constraint in (14) that $T\left(\boldsymbol{u}\right) \geq \boldsymbol{0}$ and $\boldsymbol{W} \geq \boldsymbol{Y}^{H}\left[T\left(\boldsymbol{u}\right)\right]^{-1}\boldsymbol{Y}$. So, it suffices to show that

$$\|\boldsymbol{Y}\|_{\mathcal{A}} = \min_{\boldsymbol{u}} \frac{\sqrt{N}}{2}u_1 + \frac{1}{2\sqrt{N}}\operatorname{tr}\left(\boldsymbol{Y}^{H}\left[T\left(\boldsymbol{u}\right)\right]^{-1}\boldsymbol{Y}\right), \\ \text{subject to } T\left(\boldsymbol{u}\right) \geq \boldsymbol{0}, \quad (21)$$

where u_1 is the first entry of $\boldsymbol{u}$. Let $T\left(\boldsymbol{u}\right) = \boldsymbol{A}\boldsymbol{P}\boldsymbol{A}^{H} = \left[\boldsymbol{A}\boldsymbol{P}^{\frac{1}{2}}\right]\left[\boldsymbol{A}\boldsymbol{P}^{\frac{1}{2}}\right]^{H}$ be any feasible Vandermonde decomposition, where $\boldsymbol{A} = \boldsymbol{A}\left(\boldsymbol{f}\right) = \left[\ldots, \boldsymbol{a}\left(f_j\right), \ldots\right]$ and $\boldsymbol{P} = \operatorname{diag}\left(\ldots, p_j, \ldots\right)$ with $p_j > 0$. It follows that $u_1 = \sum p_j$. For the tth column of $\boldsymbol{Y}$, say $\boldsymbol{y}_{:t}$, it holds by Lemma 1 that

$$
\begin{aligned}
\boldsymbol{y}_{:t}^{H}\left[T\left(\boldsymbol{u}\right)\right]^{-1}\boldsymbol{y}_{:t} &= \min_{\boldsymbol{v}}\|\boldsymbol{v}\|_2^2, \text{ subject to } \boldsymbol{A}\boldsymbol{P}^{\frac{1}{2}}\boldsymbol{v} = \boldsymbol{y}_{:t}\\
&= \min_{\boldsymbol{s}}\left\|\boldsymbol{P}^{-\frac{1}{2}}\boldsymbol{s}\right\|_2^2, \text{ subject to } \boldsymbol{A}\boldsymbol{s} = \boldsymbol{y}_{:t}\\
&= \min_{\boldsymbol{s}}\boldsymbol{s}^{H}\boldsymbol{P}^{-1}\boldsymbol{s}, \text{ subject to } \boldsymbol{A}\boldsymbol{s} = \boldsymbol{y}_{:t}.
\end{aligned}
$$

It follows that

$$\text{tr}\left(\boldsymbol{Y}^H \left[T\left(\boldsymbol{u}\right)\right]^{-1} \boldsymbol{Y}\right) = \sum_{t=1}^{N} \boldsymbol{y}_{:t}^H \left[T\left(\boldsymbol{u}\right)\right]^{-1} \boldsymbol{y}_{:t}$$

$$= \min_{\boldsymbol{S}, \boldsymbol{A}(\boldsymbol{f})\boldsymbol{S}=\boldsymbol{Y}} \text{tr}\left(\boldsymbol{S}^H \boldsymbol{P}^{-1} \boldsymbol{S}\right).$$

We complete the proof via the following equalities:

$$\min_{\boldsymbol{u}} \frac{\sqrt{N}}{2} u_1 + \frac{1}{2\sqrt{N}} \text{tr}\left(\boldsymbol{Y}^H \left[T\left(\boldsymbol{u}\right)\right]^{-1} \boldsymbol{Y}\right)$$

$$= \min_{\substack{\boldsymbol{f}, \boldsymbol{p} \succeq 0, \boldsymbol{S} \\ \boldsymbol{A}(\boldsymbol{f})\boldsymbol{S}=\boldsymbol{Y}}} \frac{\sqrt{N}}{2} \sum_j p_j + \frac{1}{2\sqrt{N}} \text{tr}\left(\boldsymbol{S}^H \boldsymbol{P}^{-1} \boldsymbol{S}\right)$$

$$= \min_{\substack{\boldsymbol{f}, \boldsymbol{p} \succeq 0, \boldsymbol{S} \\ \boldsymbol{A}(\boldsymbol{f})\boldsymbol{S}=\boldsymbol{Y}}} \frac{\sqrt{N}}{2} \sum_j p_j + \frac{1}{2\sqrt{N}} \sum_j \|\boldsymbol{S}_j\|_2^2 p_j^{-1} \quad (22)$$

$$= \min_{\boldsymbol{f}, \boldsymbol{S}} \sum_j \|\boldsymbol{S}_j\|_2, \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{A}\left(\boldsymbol{f}\right)\boldsymbol{S}$$

$$= \|\boldsymbol{Y}\|_{\mathcal{A}},$$

where the optimal solution of p_j equals $\frac{1}{\sqrt{N}}\|\boldsymbol{S}_j\|_2$ and the last equality follows from (17).

C. Proof of Theorem 1

We use contradiction. Suppose that there exists $\widetilde{\boldsymbol{Y}} \neq \boldsymbol{Y}^o$ satisfying that $\widetilde{\boldsymbol{Y}}_{\boldsymbol{\Omega}} = \boldsymbol{Y}_{\boldsymbol{\Omega}}^o$ and $\widetilde{K} := \left\|\widetilde{\boldsymbol{Y}}\right\|_{\mathcal{A},0} \leq \|\boldsymbol{Y}^o\|_{\mathcal{A},0} = K$. Let $\widetilde{\boldsymbol{Y}} = \sum_{k=1}^{\widetilde{K}} \boldsymbol{a}\left(\widetilde{f}_j\right)\widetilde{\boldsymbol{s}}_j$ be an atomic decomposition. Also let $\boldsymbol{A}_1 = [\boldsymbol{a}\left(f\right)]_{f \in \mathcal{T} \setminus \{\widetilde{f}_j\}}$ (the matrix consisting of those $\boldsymbol{a}\left(f\right)$, $f \in \mathcal{T} \setminus \left\{\widetilde{f}_j\right\}$), $\boldsymbol{A}_{12} = [\boldsymbol{a}\left(f\right)]_{f \in \mathcal{T} \cap \{\widetilde{f}_j\}}$ and $\boldsymbol{A}_2 = [\boldsymbol{a}\left(f\right)]_{f \in \{\widetilde{f}_j\} \setminus \mathcal{T}}$. In addition, let $K_{12} = \left|\mathcal{T} \cap \left\{\widetilde{f}_j\right\}\right|$ and $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_{12} & \boldsymbol{A}_2 \end{bmatrix}$. Then we have $\boldsymbol{Y}^o = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_{12} \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_{12} \end{bmatrix}$ and $\widetilde{\boldsymbol{Y}} = \begin{bmatrix} \boldsymbol{A}_{12} & \boldsymbol{A}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_{21} \\ \boldsymbol{S}_2 \end{bmatrix}$, where $\boldsymbol{S}_1, \boldsymbol{S}_{12}, \boldsymbol{S}_{21}$ and $\boldsymbol{S}_2$ are properly defined. It follows that $\boldsymbol{Y}^o - \widetilde{\boldsymbol{Y}} = \boldsymbol{A}\boldsymbol{\Upsilon} \neq \boldsymbol{0}$, where $\boldsymbol{\Upsilon} = \begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_{12} - \boldsymbol{S}_{21} \\ -\boldsymbol{S}_2 \end{bmatrix} \neq \boldsymbol{0}$. On the other hand, it follows from $\widetilde{\boldsymbol{Y}}_{\boldsymbol{\Omega}} = \boldsymbol{Y}_{\boldsymbol{\Omega}}^o$ that $\boldsymbol{A}_{\boldsymbol{\Omega}}\boldsymbol{\Upsilon} = \boldsymbol{0}$. Note that $\boldsymbol{A}_{\boldsymbol{\Omega}}$ is composed of atoms in $\mathcal{A}_{\boldsymbol{\Omega}}^1$ and has a nontrivial null space since we have shown that $\boldsymbol{\Upsilon} \neq \boldsymbol{0}$. Then,

$$\text{rank}\left(\boldsymbol{A}_{\boldsymbol{\Omega}}\right) \geq \text{spark}\left(\mathcal{A}_{\boldsymbol{\Omega}}^1\right) - 1. \quad (23)$$

Moreover, for the nullity (dimension of the null space) of $\boldsymbol{A}_{\boldsymbol{\Omega}}$ it holds that

$$\text{nullity}\left(\boldsymbol{A}_{\boldsymbol{\Omega}}\right) \geq \text{rank}\left(\boldsymbol{\Upsilon}\right)$$

$$\geq \text{rank}\left(\begin{bmatrix} \boldsymbol{S}_1 \\ \boldsymbol{S}_{12} \end{bmatrix}\right) - \text{rank}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{S}_{21} \end{bmatrix}\right) \quad (24)$$

$$\geq \text{rank}\left(\boldsymbol{Y}_{\boldsymbol{\Omega}}^o\right) - K_{12}.$$

Consequently, the equality

$$\#\text{columns of } \boldsymbol{A}_{\boldsymbol{\Omega}} = \text{rank}\left(\boldsymbol{A}_{\boldsymbol{\Omega}}\right) + \text{nullity}\left(\boldsymbol{A}_{\boldsymbol{\Omega}}\right)$$

together with (23) and (24) yields that $K + \widetilde{K} - K_{12} \geq \text{spark}\left(\mathcal{A}_{\boldsymbol{\Omega}}^1\right) - 1 + \text{rank}\left(\boldsymbol{Y}_{\boldsymbol{\Omega}}^o\right) - K_{12}$. Therefore,

$$2K \geq K + \widetilde{K} \geq \text{spark}\left(\mathcal{A}_{\boldsymbol{\Omega}}^1\right) - 1 + \text{rank}\left(\boldsymbol{Y}_{\boldsymbol{\Omega}}^o\right),$$

which contradicts the condition in (9).

To show the uniqueness part, note that the condition in (9) implies that $K < \text{spark}\left(\mathcal{A}_{\boldsymbol{\Omega}}^1\right) - 1$ since $\text{rank}\left(\boldsymbol{Y}_{\boldsymbol{\Omega}}^o\right) \leq K$. According to the definition of spark, any K atoms in $\mathcal{A}_{\boldsymbol{\Omega}}^1$ are linearly independent. Therefore, the atomic decomposition is unique given the set of frequencies $\mathcal{T} = \{f_j\}_{j=1}^K$. Now suppose there exists another decomposition $\boldsymbol{Y}^{o'} = \sum_{j=1}^{\widetilde{K}} \boldsymbol{a}\left(\widetilde{f}_j\right)\widetilde{\boldsymbol{s}}_j$, where $\widetilde{K} \leq K$ and $\widetilde{f}_{j_0} \notin \mathcal{T}$ for some $j_0 \in \left[\widetilde{K}\right]$. Note that we have used the same notations for simplicity and we similarly define the other notations. Once again we have that $\boldsymbol{\Upsilon} \neq \boldsymbol{0}$ since $\boldsymbol{A}_2$ is nonempty and $\boldsymbol{S}_2 \neq \boldsymbol{0}$. The rest of the proof follows from the same arguments as above.

D. Proof of Theorem 4

The proof of Theorem 4 generalizes that in [20] (and reorganized in [23]) from the SMV to the MMV case. The main challenge is how to construct and deal with a vector-valued dual polynomial induced by the MMV problem, instead of the scalar-valued one in [20]. Since our proof follows similar procedures as in [20] and because of the page limit, we only highlight the key steps. Readers are referred to Section 5 of the technical report [32] for the detailed proof.

Following from [23], we can consider an equivalent case of symmetric data index set $\boldsymbol{J} = \{-2n, \ldots, 2n\}$, where $n = \left\lfloor \frac{N-1}{4} \right\rfloor$, instead of the set specified in (1). As in [20], we link Theorem 4 to a dual polynomial. In particular, Theorem 4 holds if there exists a vector-valued dual polynomial $Q : \mathbb{T} \to \mathbb{C}^{1 \times L}$,

$$Q(f) = \boldsymbol{a}(f)^H \boldsymbol{V} \quad (25)$$

satisfying that

$$Q\left(f_k\right) = \boldsymbol{\phi}_k, \quad f_k \in \mathcal{T}, \quad (26)$$

$$\|Q\left(f\right)\|_2 < 1, \quad f \in \mathbb{T} \setminus \mathcal{T}, \quad (27)$$

where the coefficient matrix $\boldsymbol{V} \in \mathbb{C}^{|\boldsymbol{J}| \times L}$. The following proof is devoted to construction of $Q(f)$ under the assumptions of Theorem 4.

Inspired by [20], we let

$$Q\left(f\right) = \sum_{f_j \in \mathcal{T}} \boldsymbol{\alpha}_j \mathcal{K}\left(f - f_j\right) + \sum_{f_j \in \mathcal{T}} \boldsymbol{\beta}_j \mathcal{K}'\left(f - f_j\right), \quad (28)$$

where $\mathcal{K}\left(f\right)$ is the squared Fejér kernel

$$\mathcal{K}\left(f\right) = \left[\frac{\sin(\pi(n+1)f)}{(n+1)\sin(\pi f)}\right]^4 = \sum_{j=-2n}^{2n} g_j e^{-i2\pi j f} \quad (29)$$

in which g_j are constant, $\mathcal{K}'$ denotes the first-order derivative of $\mathcal{K}$, and the coefficients $\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j \in \mathbb{C}^{1 \times L}$ are specified by imposing (26) and

$$Q'\left(f_k\right) = \boldsymbol{0}, \quad f_k \in \mathcal{T}. \quad (30)$$

The equations in (26) and (30) can be combined into the linear system of equations:

$$\begin{bmatrix} \boldsymbol{D}_0 & c_0^{-1}\boldsymbol{D}_1 \\ -c_0^{-1}\boldsymbol{D}_1 & -c_0^{-2}\boldsymbol{D}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ c_0\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{0} \end{bmatrix}, \qquad (31)$$

where the coefficient matrix $\boldsymbol{D} := \begin{bmatrix} \boldsymbol{D}_0 & c_0^{-1}\boldsymbol{D}_1 \\ -c_0^{-1}\boldsymbol{D}_1 & -c_0^{-2}\boldsymbol{D}_2 \end{bmatrix}$ only depends on the frequency set $\mathcal{T}$, $c_0 = \sqrt{\mathcal{K}''(0)}$ is a constant, $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_K^T \end{bmatrix}^T \in \mathbb{C}^{K \times L}$, $\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_K^T \end{bmatrix}^T \in \mathbb{C}^{K \times L}$ and $\boldsymbol{\beta} \in \mathbb{C}^{K \times L}$ is similarly defined. Using the fact that the coefficient matrix in (31) is close to identity [20], we next prove that $\begin{bmatrix} \boldsymbol{\alpha} \\ c_0\boldsymbol{\beta} \end{bmatrix}$ is close to $\begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{0} \end{bmatrix}$. Different from the SMV case in which $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are scalars, the difficulty in our proof is how to quantify this closeness. To do this, we define the $\ell_{2,\infty}$ matrix norm and its induced operator norm as follows.

Definition 1. *We define the $\ell_{2,\infty}$ norm of $\boldsymbol{X} \in \mathbb{C}^{d_1 \times d_2}$ as*

$$\|\boldsymbol{X}\|_{2,\infty} = \max_j \|\boldsymbol{X}_j\|_2$$

and its induced norm of a linear operator $\boldsymbol{\mathcal{P}} : \mathbb{C}^{d_1 \times d_2} \to \mathbb{C}^{d_3 \times d_2}$ as

$$\|\boldsymbol{\mathcal{P}}\|_{2,\infty} = \sup_{\boldsymbol{X} \neq \boldsymbol{0}} \frac{\|\boldsymbol{\mathcal{P}}\boldsymbol{X}\|_{2,\infty}}{\|\boldsymbol{X}\|_{2,\infty}} = \sup_{\|\boldsymbol{X}\|_{2,\infty} \leq 1} \|\boldsymbol{\mathcal{P}}\boldsymbol{X}\|_{2,\infty} ,$$

where $\boldsymbol{X}_j$ denotes the jth row of $\boldsymbol{X}$, and d_1, d_2 and d_3 are positive integers.

By Definition 1, we have that $\|\boldsymbol{\Phi}\|_{2,\infty} = 1$ and expect to bound $\|\boldsymbol{\alpha}\|_{2,\infty}$ and $\|\boldsymbol{\beta}\|_{2,\infty}$ using the induced norm of the operators $\boldsymbol{D}_j$, $j = 0, 1, 2$. To do so, we calculate the induced norm first. Interestingly, the induced $\ell_{2,\infty}$ norm is identical to the ℓ_∞ norm, which is stated in the following result.

Lemma 2 ([32]). $\|\boldsymbol{\mathcal{P}}\|_{2,\infty} = \|\boldsymbol{\mathcal{P}}\|_\infty$ *for any linear operator $\boldsymbol{\mathcal{P}}$ defined by a matrix $\boldsymbol{P}$ such that $\boldsymbol{\mathcal{P}}\boldsymbol{X} = \boldsymbol{P}\boldsymbol{X}$ for any $\boldsymbol{X}$ of proper dimension.*

By Lemma 2 the $\ell_{2,\infty}$ operator norm of $\boldsymbol{D}_j$, $j = 0, 1, 2$ equals their ℓ_∞ norm that has been derived in [20]. Then, under the assumptions of Theorem 4 and using the results in [20], we can show that

$$\|\boldsymbol{\alpha} - \boldsymbol{\Phi}\|_{2,\infty} \leq 8.824 \times 10^{-3}, \qquad (32)$$

$$\|\boldsymbol{\beta}\|_{2,\infty} \leq \frac{1.647}{n} \times 10^{-2}. \qquad (33)$$

Finally, we complete the proof by showing that the constructed polynomial $Q(f)$ satisfies (27) using (32), (33) and the bounds on $\mathcal{K}(f - f_k)$ and its derivatives given in [20]. As in [20], we divide $\mathbb{T}$ into several intervals that are either neighborhood of or far from some $f_k \in \mathcal{T}$. If f is far from every $f_k \in \mathcal{T}$, then we can show that $\|Q(f)\|_2 \leq 0.99992$. Otherwise, we can show that on the neighborhood of $f_k \in \mathcal{T}$, the second derivative of $\|Q(f)\|_2^2$ is negative. This means that $\|Q(f)\|_2^2$ is a strictly concave function and achieves its maximum 1 at the only stationary point f_k by (30). So we can conclude (27) and complete the proof.

E. Proof of Theorem 5

The proof of Theorem 4 in the last subsection forms the basis of the proof of Theorem 5 that will be given following similar steps as in [23]. As in the full data case, we only highlight the key steps of our proof and interested readers are referred to [32, Section 6] for the details. Similarly, we can also consider the symmetric case of $\boldsymbol{J} = \{-2n, \ldots, 2n\}$ and start with the dual certificate. In particular, $\boldsymbol{Y}^o = \sum_{k=1}^K c_k \boldsymbol{a}(f_k, \boldsymbol{\phi}_k)$ is the unique optimizer to (13) and provides the unique atomic decomposition satisfying that $\|\boldsymbol{Y}^o\|_{\mathcal{A}} = \sum_{k=1}^K c_k$ if 1) $\{\boldsymbol{a}_{\boldsymbol{\Omega}}(f_k)\}_{f_k \in \mathcal{T}} \subset \mathcal{A}_{\boldsymbol{\Omega}}^1$ are linearly independent and 2) there exists a vector-valued dual polynomial $\overline{Q}(f) = \boldsymbol{a}^H(f)\boldsymbol{V} \in \mathbb{C}^{1 \times L}$ as in (25) satisfying (26), (27) and the additional constraint that

$$\boldsymbol{V}_j = \boldsymbol{0}, \quad j \notin \boldsymbol{\Omega}. \qquad (34)$$

Note that the condition of linear independence above is necessary to prove the uniqueness part but is neglected in [23]. We will show later that this condition is satisfied for free when we construct the dual polynomial $\overline{Q}(f)$ under the assumptions of Theorem 5. As in [23], we consider an equivalent Bernoulli observation model in which the samples indexed by $\boldsymbol{J}$ are observed independently with probability $p = \frac{M}{4n}$. In mathematics, let $\{\delta_j\}_{j \in \boldsymbol{J}}$ be i.i.d. Bernoulli random variables such that

$$\mathbb{P}(\delta_j = 1) = p, \qquad (35)$$

where $\delta_j = 1$ or 0 indicates whether the jth entry in $\boldsymbol{J}$ is observed or not. It follows that the sampling index set $\boldsymbol{\Omega} = \{j : \delta_j = 1\}$.

Inspired by [23], we let

$$\overline{Q}(f) = \sum_{f_j \in \mathcal{T}} \boldsymbol{\alpha}_j \overline{\mathcal{K}}(f - f_j) + \sum_{f_j \in \mathcal{T}} \boldsymbol{\beta}_j \overline{\mathcal{K}}'(f - f_j), \quad (36)$$

where $\overline{\mathcal{K}}(f)$ is a random analog of $\mathcal{K}(f)$ as defined in (29):

$$\overline{\mathcal{K}}(f) = \sum_{j=-2n}^{2n} \delta_j g_n(j) e^{-i2\pi j f}. \qquad (37)$$

It is clear that $\mathbb{E}\overline{\mathcal{K}}(f) = p\mathcal{K}(f)$ and similar result holds for its derivatives. Again, we impose for the coefficients $\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j \in \mathbb{C}^{1 \times L}$ that

$$\overline{\boldsymbol{D}} \begin{bmatrix} \boldsymbol{\alpha} \\ c_0\boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi} \\ \boldsymbol{0} \end{bmatrix}, \qquad (38)$$

where $\overline{\boldsymbol{D}}$ is a random analog of $\boldsymbol{D}$ in (31) with $\mathbb{E}\overline{\boldsymbol{D}} = p\boldsymbol{D}$. It is clear that $\overline{Q}(f)$ above already satisfies (26) and (34). The remaining task is showing that it also satisfies (27) under the assumptions of Theorem 5.

Let $Q(f)$ be the dual polynomial in (25) that is the full data case counterpart of $\overline{Q}(f)$. As in [23], we need to show that $\overline{Q}(f)$ (and its derivatives) is tightly concentrated around $Q(f)$ (and its derivatives) when the sample size M satisfies (16). To do this, define two events

$$\mathcal{E}_1 = \left\{ \|p^{-1}\overline{\boldsymbol{D}} - \boldsymbol{D}\|_2 \leq \frac{1}{4} \right\}, \qquad (39)$$

$$\mathcal{E}_2 = \left\{ \sup_{f \in \mathbb{T}_{\text{grid}}} c_0^{-l} \|\overline{Q}^{(l)} - Q^{(l)}\|_2 \leq \frac{\epsilon}{3}, l = 0, 1, 2, 3 \right\} \quad (40)$$

where $\mathbb{T}_{\text{grid}} \subset \mathbb{T}$ and $\epsilon > 0$ are a set of discrete points and a small number, respectively, to specify. It has been shown in [23] that $\overline{D}$ is invertible on $\mathcal{E}_1$ which happens with probability at least $1 - \delta$ if

$$M \geq C_1 K \log \frac{K}{\delta} \tag{41}$$

and if the frequency separation condition is satisfied, where C_1 is constant. Note that the aforementioned linear independence of $\{a_\Omega(f_k)\}_{f_k \in \mathcal{T}} \subset \mathcal{A}_\Omega^1$ can be shown based on this result (see [32, Lemma 6.4]). We next focus on the case when $\mathcal{E}_1$ happens. It follows that

$$\begin{bmatrix} \alpha \\ c_0 \beta \end{bmatrix} = \overline{D}^{-1} \begin{bmatrix} \Phi \\ 0 \end{bmatrix} = \overline{L} \Phi, \tag{42}$$

where $\overline{L} \in \mathbb{C}^{2K \times K}$ denotes the left part of $\overline{D}^{-1}$. Therefore, as in [23], we have that

$$c_0^{-l} \left[\overline{Q}^{(l)}(f) - Q^{(l)}(f) \right] = H_1(f)\Phi + H_2(f)\Phi, \tag{43}$$

where $H_1(f), H_2(f) \in \mathbb{C}^{1 \times K}$ are as defined and bounded in [23]. The main difference from [23] lies in the fact that Φ is a $K \times L$ matrix instead of a $K \times 1$ vector. To show that both $\|H_1(f)\Phi\|_2$ and $\|H_2(f)\Phi\|_2$ are concentrated around 0 with high probability, we need the following vector-form Hoeffding's inequality that can be proven based on [37, Theorem 1.3].

Lemma 3 ([32]). *Let the rows of $\Phi \in \mathbb{C}^{K \times L}$ be sampled independently on the complex hyper-sphere $\mathbb{S}^{2L-1}$ with zero mean. Then, for all $w \in \mathbb{C}^K$, $w \neq 0$, and $t \geq 0$,*

$$\mathbb{P}\left(\left\| w^H \Phi \right\|_2 \geq t \right) \leq (L+1) e^{-\frac{t^2}{8\|w\|_2^2}}.$$

Using Lemma 3 we can show that $\mathcal{E}_2$ happens with probability at least $1 - \delta$ if

$$M \geq C_2 \frac{1}{\epsilon^2} \max \left\{ \log \frac{|\mathbb{T}_{\text{grid}}|}{\delta} \log \frac{L |\mathbb{T}_{\text{grid}}|}{\delta}, \right. \\ \left. K \log \frac{K}{\delta} \log \frac{L |\mathbb{T}_{\text{grid}}|}{\delta} \right\} \tag{44}$$

among other assumptions in Theorem 5, where C_2 is constant. This result is then extended, as in [23], from $\mathbb{T}_{\text{grid}}$ to the whole unit circle $\mathbb{T}$ by choosing some $\mathbb{T}_{\text{grid}}$ satisfying that

$$|\mathbb{T}_{\text{grid}}| < \frac{3C_3 \sqrt{L} n^3}{\epsilon}, \tag{45}$$

where C_3 is also constant. This means that $\overline{Q}(f)$ (and its derivatives) is concentrated around $Q(f)$ (and its derivatives) with high probability. Now we are ready to complete the proof by showing that $\|\overline{Q}(f)\|_2$ satisfies (27) using the properties of $Q(f)$ shown in the last subsection and by properly choosing ϵ. In particular, letting $\epsilon = 10^{-5}$, $\|\overline{Q}(f)\|_2$ can still be well bounded by 1 from above when f is far from every $f_k \in \mathcal{T}$. When f is in the neighborhood of some $f_k \in \mathcal{T}$, the second derivative of $\|\overline{Q}(f)\|_2^2$ is concentrated around the second derivative of $\|Q(f)\|_2^2$ and thus it is negative. It follows that $\|\overline{Q}(f)\|_2^2$ is strictly concave and achieves the maximum 1 at the only stationary point f_k. Finally, to close the proof, note that inserting (45) into (44) resulting in the bound in (16).
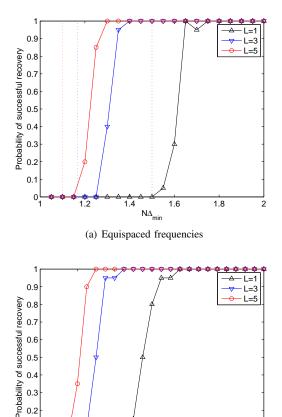

(a) Equispaced frequencies

(b) Random frequencies

Fig. 1. Frequency recovery results with respect to the number of measurement vectors L in the case of full data and uncorrelated sources.

V. NUMERICAL SIMULATIONS

A. Full Data

We consider the full data case and test the frequency recovery performance of the proposed atomic norm method with respect to the frequency separation condition. In particular, we consider two types of frequencies, equispaced and random, and two types of source signals, uncorrelated and coherent. We fix $N = 128$ and vary $\Delta_{\min}$ (a lower bound of the minimum separation of frequencies) from $1.05N^{-1}$ (or $0.9N^{-1}$ for random frequencies) to $2N^{-1}$ at a step of $0.05N^{-1}$. In the case of equispaced frequencies, for each $\Delta_{\min}$ we generate a set of frequencies $\mathcal{T}$ of the maximal cardinality $\lfloor \Delta_{\min}^{-1} \rfloor$ with frequency separation $\Delta_{\mathcal{T}} = \frac{1}{\lfloor \Delta_{\min}^{-1} \rfloor} \geq \Delta_{\min}$. In the case of random frequencies, we generate the frequency set $\mathcal{T}$, $\Delta_{\mathcal{T}} \geq \Delta_{\min}$, by repetitively adding new frequencies (generated uniformly at random) till no more can be added. Therefore, any two adjacent frequencies in $\mathcal{T}$ are separated by a value in the interval $[\Delta_{\min}, 2\Delta_{\min})$. It follows that $|\mathcal{T}| \in \left(\frac{1}{2}\Delta_{\min}^{-1}, \Delta_{\min}^{-1}\right]$. We empirically find that $\mathbb{E}|\mathcal{T}| \approx \frac{3}{4}\Delta_{\min}^{-1}$ which is the midpoint of the interval above.

We first consider uncorrelated sources, where the source signals $S = [s_{kt}] \in \mathbb{C}^{K \times L}$ in (1) are drawn i.i.d. from a standard complex Gaussian distribution. Moreover, we consider
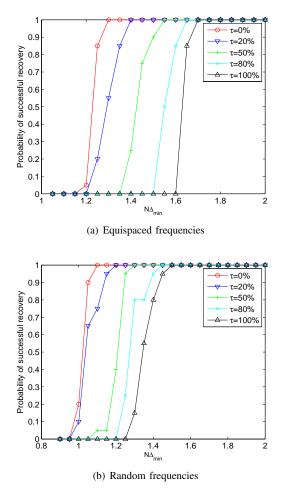
(a) Equispaced frequencies

(b) Random frequencies

Fig. 2. Frequency recovery results with respect to the percentage of coherent sources τ in the case of full data and coherent sources, with $L = 5$.

the number of measurement vectors $L = 1, 3,$ and 5. For each value of $\Delta_{\min}$ and each type of frequencies, we carry out 20 Monte Carlo runs and calculate the success rate of frequency recovery. In each run, we generate $\mathcal{T}$ and $\boldsymbol{S} \in \mathbb{C}^{K \times 5}$ and obtain the full data $\boldsymbol{Y}^o$. For each value of L, we attempt to recover the frequencies using the proposed atomic norm method, implemented by SDPT3 [38] in Matlab, based on the first L columns of $\boldsymbol{Y}^o$. The frequencies are considered to be successfully recovered if the root mean squared error (RMSE) is less than 10^{-8}.

The simulation results are presented in Fig. 1, which verify the conclusion of Theorem 4 that the frequencies can be exactly recovered using the proposed atomic norm method under a frequency separation condition. When more measurement vectors are available, the recovery performance improves and it seems that a weaker frequency separation condition is sufficient to guarantee exact frequency recovery. By comparing Fig. 1(a) and Fig. 1(b), it also can be seen that a stronger frequency separation condition is required in the case of equispaced frequencies where more frequencies are present and they are located more closely.

We next consider coherent sources. In this simulation, we fix $L = 5$ and consider different percentages, denoted by τ, of the K source signals which are coherent (identical up to a

scaling factor). It follows that $\tau = 0\%$ refers to the case of uncorrelated sources considered previously. $\tau = 100\%$ means that all the sources signals are coherent and the problem is equivalent to the SMV case. For each type of frequencies, we consider five values of τ ranging from 0% to 100% and calculate each success rate over 20 Monte Carlo runs.

Our simulation results are presented in Fig. 2. It can be seen that, as τ increases, the success rate decreases and a stronger frequency separation condition is required for exact frequency recovery. As τ equals the extreme value 100%, the curves of success rate approximately match those for $L = 1$ in Fig. 1, verifying that taking MMVs does not necessarily improve the performance of frequency recovery.[2]

Finally, we report the computational speed of the proposed atomic norm method. It takes about 11s to solve one SDP on average on a PC and the CPU times differ slightly for the three values of L. About 22 hours are used in total to produce the data generating Fig. 1 and Fig. 2.

B. Compressive Data

In the compressive data case, we study the so-called phase transition phenomenon in the (M, K) plane. In particular, we fix $N = 128$, $L = 5$ and $\Delta_{\min} = 1.2N^{-1}$, and study the performance of the proposed ANM method in signal and frequency recovery with different settings of the source signal. The frequency set $\mathcal{T}$ is randomly generated with $\Delta_{\mathcal{T}} \geq \Delta_{\min}$ and $|\mathcal{T}| = K$ (differently from that in the last subsection, the process of adding frequencies is terminated as $|\mathcal{T}| = K$). In our simulation, we vary $M = 8, 12, \ldots, 128$ and at each M, $K = 2, 4, \ldots, \min(M, 84)$ since it is difficult to generate a set of frequencies with $K > 84$ under the aforementioned frequency separation condition. In this simulation, we consider temporarily correlated sources. In particular, suppose that each row of $\boldsymbol{S}$ has a Toeplitz covariance matrix $\boldsymbol{R}(r) = \begin{bmatrix} 1 & r & \ldots & r^4 \\ r & 1 & \ldots & r^3 \\ \vdots & \vdots & \ddots & \vdots \\ r^4 & r^3 & \ldots & 1 \end{bmatrix} \in \mathbb{R}^{5 \times 5}$ (up to a positive scaling factor). Therefore, $r = 0$ means that the source signals at different snapshots are uncorrelated while $r = \pm 1$ means completely correlated and corresponds to the trivial case. We first generate $\boldsymbol{S}_0$ from an i.i.d. standard complex Gaussian distribution and then let $\boldsymbol{S}(r) = \boldsymbol{S}_0 \boldsymbol{R}(r)^{\frac{1}{2}}$, where we consider $r = 0, 0.5, 0.9, 1$. For each combination (M, K), we carry out 20 Monte Carlo runs and calculate the rate of successful recovery with respect to r. The recovery is considered successful if the relative RMSE of data recovery, measured by $\|\boldsymbol{Y}^* - \boldsymbol{Y}^o\|_F / \|\boldsymbol{Y}^o\|_F$, is less than 10^{-8} and the RMSE of frequency recovery is less than 10^{-6}, where $\boldsymbol{Y}^*$ denotes the solution of $\boldsymbol{Y}$.

The simulation results are presented in Fig. 3, where the phase transition phenomenon from perfect recovery to complete failure can be observed in each subfigure. It can be seen that more frequencies can be recovered when more samples are

[2]The slight differences between the curves in Fig. 1 and Fig. 2 are partially caused by the fact that, to simulate coherence sources, the two datasets are generated slightly differently.
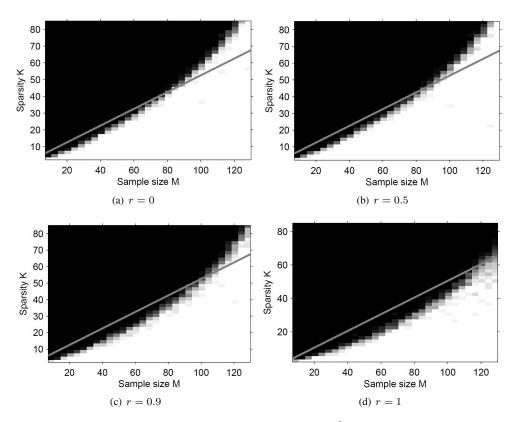
(a) $r = 0$

(b) $r = 0.5$

(c) $r = 0.9$

(d) $r = 1$

Fig. 3. Phase transition results in the compressive data case with $N = 128$ and $\Delta_{\min} = 1.2N^{-1}$. White means complete success and black means complete failure. The straight lines are $K = \frac{1}{2}(M + L)$ in (a)-(c) and $K = \frac{1}{2}(M + 1)$ in (d).

observed. When the correlation level of the MMVs, indicated by r, increases, the phase of successful recovery decreases. On the other hand, note that Fig. 3(d) actually corresponds to the SMV case. By comparing Fig. 3(d) and the other three subfigures, it can be seen that the frequency recovery performance can be greatly improved by taking MMVs, even in the presence of strong temporal correlations.

We also plot the line $K = \frac{1}{2}(M + L)$ in Figs. 3(a)-3(c) and the line $K = \frac{1}{2}(M + 1)$ in Fig. 3(d) (straight gray lines) which are upper bounds of the sufficient condition in Theorem 1 for the atomic ℓ_0 norm minimization (note that spark $\left(\mathcal{A}_{\boldsymbol{\Omega}}^1\right) \leq M + 1$). It can be seen that successful recoveries can be obtained even above these lines, indicating good performance of the proposed ANM method. It requires about 13s on average to solve one problem and almost 200 hours in total to generate the whole data set used in Fig. 3.

C. The Noisy Case

While this paper has been focused on the noiseless case, we provide a simple simulation to illustrate the performance of the proposed method in the practical noisy case. We consider $N = 50$, $M = 20$ with $\boldsymbol{\Omega}$ randomly generated, $K = 3$ sources with frequencies of 0.1, 0.12 and 0.3 and powers of 2, 3 and 1 respectively, and $L = 5$. The source signals of each source are generated with constant amplitude and random phases. Complex white Gaussian noise is added to the measurements with noise variance $\sigma^2 = 0.1$. We propose to denoise the observed noisy signal $\boldsymbol{Y}_{\boldsymbol{\Omega}}^o$ and recover the

frequency components by solving the following optimization problem:

$$\min_{\boldsymbol{Y}} \|\boldsymbol{Y}\|_{\mathcal{A}}, \text{ subject to } \|\boldsymbol{Y}_{\boldsymbol{\Omega}} - \boldsymbol{Y}_{\boldsymbol{\Omega}}^o\|_{\mathrm{F}}^2 \leq \eta^2, \qquad (46)$$

where η^2, set to $\left(ML + 2\sqrt{ML}\right)\sigma^2$ (mean + twice standard deviation), bounds the noise energy from above with large probability. The spectral MUSIC method is also considered for comparison. Note that MUSIC estimates the frequencies from the sample covariance, while the proposed ANM method carries out covariance fitting by exploiting its structures. While the proposed method requires the noise level, MUSIC needs the source number K.

Typical simulation results of one Monte Carlo run are presented in Fig. 4. The SMV case is studied in Fig. 4(a) where only the first measurement vector is used for frequency recovery. It is shown that the three frequency components are correctly identified using the ANM method while MUSIC fails. The MMV case is studied in Fig. 4(b) with uncorrelated sources, where both ANM and MUSIC succeed to identify the three frequency components. The case of coherent sources is presented in Fig. 4(c), where source 3 in Fig. 4(b) is modified such that it is coherent with source 1. MUSIC fails to detect the two coherent sources as expected while the proposed method still performs well. It is shown in all the three subfigures that spurious frequency components can be present using the ANM method. But their powers are low. To be specific, the spurious components have about 0.4% of the total powers in Fig. 4(a), and this number is on the order of 10^{-6} in the latter
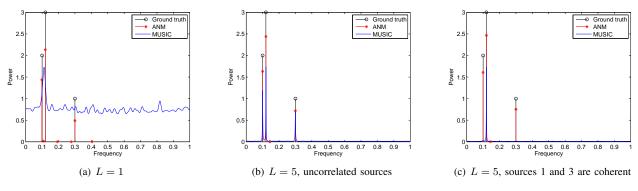
Fig. 4. Frequency recovery/estimation results of ANM and MUSIC in the presence of noise, with (a) $L = 1$, (b) $L = 5$ and uncorrelated sources, and (c) $L = 5$ and coherent sources.

two subfigures. While these numerical results imply that the proposed method is robust to noise, a theoretical analysis will be investigated in future studies. The proposed method needs about 1.5s in each scenario.

VI. CONCLUSION

In this paper we studied the JSFR problem by exploiting the joint sparsity in the MMVs. We proposed an atomic ℓ_0 norm approach and showed the advantage of MMVs. We also proposed an atomic norm approach that can be efficiently solved by semidefinite programming and studied its theoretical guarantees for frequency recovery. These results extend the existing ones either from the SMV to the MMV case or from the discrete to the continuous frequency setting. We also discussed the connections between the proposed approaches and conventional subspace methods as well as the recent grid-based and gridless sparse techniques. Though the worst case analysis we provided for the atomic norm approach does not indicate performance gains in the presence of MMVs, simulation results indeed imply that when the source signals are located at general positions the number of required measurements can be reduced and/or the frequency separation condition can be relaxed. This average case analysis should be investigated in future studies under stronger assumptions.

REFERENCES

[1] Z. Yang and L. Xie, "Continuous compressed sensing with a single or multiple measurement vectors," in *IEEE Workshop on Statistical Signal Processing (SSP)*, 2014, pp. 308–311.

[2] H. Krim and M. Viberg, "Two decades of array signal processing research: The parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, 1996.

[3] P. Stoica and R. L. Moses, *Spectral analysis of signals*. Pearson/Prentice Hall Upper Saddle River, NJ, 2005.

[4] C. Carathéodory and L. Fejér, "Über den Zusammenhang der Extremen von harmonischen Funktionen mit ihren Koeffizienten und über den Picard-Landau'schen Satz," *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, vol. 32, no. 1, pp. 218–239, 1911.

[5] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophysical Journal International*, vol. 33, no. 3, pp. 347–366, 1973.

[6] R. Schmidt, "A signal subspace approach to multiple emitter location spectral estimation," Ph.D. dissertation, Stanford University, 1981.

[7] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

[8] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[9] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[10] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2197–2202, 2003.

[11] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[12] D. Malioutov, M. Cetin, and A. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.

[13] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4634–4643, 2006.

[14] M. Hyder and K. Mahata, "Direction-of-arrival estimation using a mixed $\ell_{2,0}$ norm approximation," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4646–4655, 2010.

[15] Y. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 505–519, 2010.

[16] L. Hu, Z. Shi, J. Zhou, and Q. Fu, "Compressed sensing of complex sinusoids: An approach based on dictionary refinement," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3809–3822, 2012.

[17] Z. Yang, C. Zhang, and L. Xie, "Robustly stable signal recovery in compressed sensing with structured matrix perturbation," *IEEE Transactions on Signal Processing*, vol. 60, no. 9, pp. 4658–4671, 2012.

[18] Z. Yang, L. Xie, and C. Zhang, "Off-grid direction of arrival estimation using sparse Bayesian inference," *IEEE Transactions on Signal Processing*, vol. 61, no. 1, pp. 38–43, 2013.

[19] C. Austin, J. Ash, and R. Moses, "Dynamic dictionary algorithms for model order and parameter estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 5117–5130, 2013.

[20] E. J. Candès 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.

[21] W. Rudin, *Real and complex analysis*. New York, USA: Tata McGraw-Hill Education, 1987.

[22] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational Mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[23] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7465–7490, 2013.

[24] B. N. Bhaskar, G. Tang, and B. Recht, "Atomic norm denoising with applications to line spectral estimation," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5987–5999, 2013.

[25] E. J. Candès and C. Fernandez-Granda, "Super-resolution from noisy data," *Journal of Fourier Analysis and Applications*, vol. 19, no. 6, pp. 1229–1254, 2013.

[26] G. Tang, B. N. Bhaskar, and B. Recht, "Near minimax line spectral estimation," in *47th Annual Conference on Information Sciences and Systems (CISS)*, 2013, pp. 1–6.

[27] J. Fang, J. Li, Y. Shen, H. Li, and S. Li, "Super-resolution compressed sensing: An iterative reweighted algorithm for joint parameter learning and sparse signal recovery," *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 761–765, 2014.

[28] J.-M. Azais, Y. De Castro, and F. Gamboa, "Spike detection from inaccurate samplings," *Applied and Computational Harmonic Analysis*, vol. 38, no. 2, pp. 177–195, 2015.

[29] Z. Yang and L. Xie, "On gridless sparse methods for line spectral estimation from complete and incomplete data," *IEEE Transactions on Signal Processing*, vol. 63, no. 12, pp. 3139–3153, 2015.

[30] Z. Yang, L. Xie, and C. Zhang, "A discretization-free sparse and parametric approach for linear array signal processing," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 4959–4973, 2014.

[31] Z. Tan, Y. C. Eldar, and A. Nehorai, "Direction of arrival estimation using co-prime arrays: A super resolution viewpoint," *IEEE Transactions on Signal Processing*, vol. 62, no. 21, pp. 5565–5576, 2014.

[32] Z. Yang and L. Xie, "Exact joint sparse frequency recovery via optimization methods," Tech. Rep., May 2014. [Online]. Available: http://arxiv.org/abs/1405.6585v1

[33] Y. Chi, "Joint sparsity recovery for spectral compressed sensing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3938–3942.

[34] Y. Li and Y. Chi, "Off-the-grid line spectrum denoising and estimation with multiple measurement vectors," August 2014. [Online]. Available: http://arxiv.org/abs/1408.2242

[35] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear algebra and its applications*, vol. 18, no. 2, pp. 95–138, 1977.

[36] P. Stoica, P. Babu, and J. Li, "SPICE: A sparse covariance-based estimation method for array processing," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 629–638, 2011.

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

[38] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, "SDPT3–a MATLAB software package for semidefinite programming, version 1.3," *Optimization Methods and Software*, vol. 11, no. 1-4, pp. 545–581, 1999.