Off-the-Grid Line Spectrum Denoising and Estimation with Multiple Measurement Vectors

Yuanxin Li and Yuejie Chi

August 12, 2014

Abstract

Despite the fact the required number of samples for reconstructing a signal can be greatly reduced if it is sparse in a known basis, many real world signals are however sparse in an unknown and continuous dictionary. One example is the spectrally-sparse signal, which is composed of a small number of spectral atoms with arbitrary frequencies on the unit interval. In this paper we study the problem of denoising and estimating an ensemble of spectrally-sparse signals from their partial and noisy observations, and simultaneously recovering the set of signals if necessary.

Two approaches are developed based on atomic norm minimization and structured matrix completion, both of which can be solved efficiently via semidefinite programming. The first approach aims to estimate and denoise the set of signals from their partial and noisy observations via atomic norm minimization, and recover the frequencies via examining the dual polynomial of the convex program. We characterize the optimality condition of the proposed algorithm and derive the expected convergence rate for denoising, demonstrating the benefit of including multiple measurement vectors. The second approach aims to recover the population covariance matrix from the partially observed sample covariance matrix by motivating its low-rank Toeplitz structure without recovering the signal ensemble. The frequencies can be recovered via conventional spectrum estimation methods such as MUSIC from the estimated covariance matrix. Performance guarantee is derived with a finite number of measurement vectors. Finally, numerical examples are provided to validate the performance of the proposed algorithms, with comparisons against several existing approaches.

1 Introduction

Many signal processing applications encounter a signal ensemble where each signal in the ensemble can be represented as a sparse superposition of \( r \) complex sinusoids sharing the same frequencies, for example in remote sensing and super-resolution imaging, and the goal is to recover the set of signals and corresponding frequencies from a small number of linear measurements of the signal ensemble. While there has been a long line of traditional approaches [3], convex optimization [4, 5] has been recently proposed as an efficient way to reduce the number of measurements with provable performance guarantees by promoting the sparsity prior in the reconstruction in a tractable manner. In particular, it is shown that if the frequencies of a signal all lie on the DFT grid, the signal of length \( n \) can then be recovered exactly from an order of \( r \log n \)
randomly selected samples with high probability \([6]\), where \(r \ll n\). Other approaches also arise by exploiting the sparsity prior, altogether known as the field of Compressed Sensing (CS), finding many important applications in analog-to-digital conversion \([7,8]\) and spectrum estimation \([9]\).

However, most existing CS theories act as a model selection principle, where the signal is assumed sparse in an a priori basis, and the goal is to identify which atoms in the basis are activated. There is a modeling gap, however, from physical signals that are actually composed of a small number of parameterized atoms with continuous and unknown parameters determined by the mother nature. An example in this category that garnered much attention is the spectrally-sparse signal, where the signal is composed of a small number of spectral atoms with arbitrary frequencies on the unit interval. Performance degeneration of CS algorithms is observed and studied systematically in \([10-12]\) when there is an unavoidable basis mismatch between the actual frequencies and the assumed DFT basis. Subsequent work has been proposed to mitigate the effect of basis mismatch to a great extent \([13-16]\), but a general consensus is still lacking.

Therefore, it is necessary to develop a parameter estimation principle, which does not need an a priori basis for reconstruction while maintaining the capabilities of subsampling. One recent approach is via the atomic norm minimization \([17]\), which is a general recipe for designing convex solutions to parsimonious model selection. It has been successfully applied to recover a spectrally-sparse signal from a small number of consecutive samples \([18]\) or randomly selected samples \([19]\) from the time domain. In particular, Tang et. al. showed that a spectrally-sparse signal can be recovered from an order of \(r \log n \log r\) random samples with high probability when the frequencies satisfy a mild separation condition \([19]\) for line spectra with random amplitudes. This approach is extended to higher dimensional frequencies in \([20,21]\). Another approach is proposed in \([22,23]\) based on structured matrix completion, where the problem is reformulated into a structured multi-fold Hankel matrix completion inspired by matrix pencil \([24]\). For this approach, it is shown that an order of \(r \text{polylog} n\) randomly selected samples are sufficient to guarantee perfect recovery with high probability under some mild incoherence conditions and the approach is also amenable to higher-dimensional frequencies. Both approaches allow recovering off-the-grid frequencies at an arbitrary precision from a number of samples much smaller than \(n\). We refer interested readers for respective papers for details.

1.1 Our Contributions and Comparisons to Related Work

It has been shown in the traditional CS framework that the availability of Multiple Measurement Vectors (MMV) can further reduce the required number of samples and improve performance \([25-30]\) by harnessing the joint sparsity pattern of different signals, also known as group sparsity. In this paper, we study the problem of denoising and estimating multiple spectrally-sparse signals that share the same set of continuous-valued frequencies from their possibly partial and noisy observations. We leverage the power of MMV without assuming the frequencies to lie exactly on a grid. Two approaches are developed based on atomic norm minimization and structured matrix completion that can be solved efficiently using semidefinite programming.

The first approach, based on atomic norm minimization, aims to recover the signal ensemble and can be regarded as a continuous counterpart of the MMV model in CS. We first define the atomic norm of multiple spectrally-sparse signals and characterize its semidefinite program formulation. The atomic norm for a single spectrally-sparse signal becomes a special case \([19]\). We then consider signal recovery from their partial observations and denoising in Additive White Gaussian Noise (AWGN) based on atomic norm minimization under the respective observation models. We characterize the dual problem of the proposed algorithm and outline frequency recovery by examining the dual polynomial. We further derive the expected convergence rate for denoising as a function of the number of measurement vectors, demonstrating the benefit of including MMV. We also provide an efficient implementation based on ADMM \([31]\) for the proposed denoising algorithm. Through numerical comparisons between conventional CS algorithms and the proposed algorithm, we further show that the number of samples per signal can be reduced by harnessing
the joint sparsity pattern of multiple signals in atomic norm minimization.

A disadvantage of the above approach is that the computational complexity becomes expensive when the number of measurement vectors is high if we wish to recover them all. Recognizing that in many scenarios one only wishes to recover the set of frequencies, we switch our focus on reconstructing the covariance matrix rather than the signal ensemble. Covariance structures can be explored when multiple observations of a stochastic signal are available [32]. With a mild second-order statistical assumption on the sparse coefficients, a correlation-aware approach is proposed in [33, 34] to improve the size of recoverable support by exploring the sparse representation of the covariance matrix in the Khatri-Rao product of the signal sparsity basis. However, due to the earlier-mentioned basis mismatch issue, the correlation-aware approach can not estimate frequencies off the grid.

Under the statistical assumption that the frequencies are uncorrelated which holds in a variety of applications in array signal processing [3], the full covariance matrix is a Hermitian Toeplitz matrix whose rank is the number of distinct frequencies. In the second approach, we first calculate the partial sample covariance matrix from partial observations of the measurement vectors. A convex optimization algorithm is formulated to estimate the full Hermitian Toeplitz covariance matrix whose submatrix on the set of observed entries is close to the partial sample covariance matrix, with an additional trace regularization that promotes the Toeplitz low-rank structure. Trace regularization for positive semidefinite matrices is a widely adopted convex relaxation of the non-convex rank constraint, which is equivalent to atomic norm minimization in conjunction with the Toeplitz constraint in this case. We derive performance guarantees of the proposed structured low-rank matrix estimation algorithm with a finite number of measurement vectors assuming full observation or a complete sparse ruler. Finally, the set of frequencies can be obtained from the estimated full covariance matrix using conventional methods such as MUSIC [35]. Compared with directly applying MUSIC to the partial sample covariance matrix, the proposed algorithm has the potential to recover a higher number of frequencies than the number of samples per measurement vector by taking advantages of the array geometry, for example the co-prime array [36] or the minimum sparse ruler [37]. We provide numerical examples to validate the proposed algorithm with comparisons to other existing approaches, and demonstrate the performance gain when the number of measurement vectors increases. As our algorithm only requires the partially observed sample covariance matrix rather than the observed signals, the computational complexity does not grow with the number of measurement vectors, in contrast to CS and the first approach based on atomic norm minimization that aim to recover the signal ensemble.

Several recent papers [38, 39] have also proposed discretization-free approaches for direction-of-arrival estimation by exploiting low-rank properties of the covariance matrix under different setups. However, only statistical consistency is established for the algorithm in [38] without a finite sample analysis. The paper [39] assumes completed observation of the covariance matrix and applies low-rank matrix denoising under specific array geometries without performance guarantees.

1.2 Organization of This Paper

The rest of the paper is organized as below. Section 2 describes the signal model with MMV and defines its atomic norm. Section 3 considers line spectrum estimation and denoising based on atomic norm minimization that aims to recover the set of signals, and Section 4 presents the second algorithm based on structured matrix completion that aims to recover the covariance matrix. Numerical experiments are provided in Section 5 to validate the proposed algorithms. Finally, conclusions and future work are discussed in Section 6. Throughout the paper, matrices are denoted by bold capitals and vectors by bold lowercases. The transpose is denoted by \((\cdot)^T\), the complex conjugate or Hermitian is denoted by \((\cdot)^*\).
2 Signal Model with MMV and its Atomic Norm

In this section we first describe the spectrally-signal model with multiple vectors, then define and characterize the atomic norm associated with the MMV model for spectrally-sparse signals.

2.1 Signal Model with MMV

Let \( x = [x_1, \ldots, x_n]^T \in \mathbb{C}^n \) be a spectrally-sparse signal with \( r \) distinct frequency components, written as

\[
x = \sum_{k=1}^{r} c_k a(f_k) \triangleq V c,
\]

where each atom \( a(f) \) is defined as

\[
a(f) = \frac{1}{\sqrt{n}} \left[ 1, e^{j2\pi f}, \ldots, e^{j2\pi (n-1)f} \right]^T, \quad f \in [0,1),
\]

the matrix \( V \) is given as \( V = [a(f_1), \ldots, a(f_r)] \in \mathbb{C}^{n \times r} \), and \( c = [c_1, \ldots, c_r]^T \in \mathbb{C}^r \). The set of frequencies \( \mathcal{F} = \{f_k\}_{k=1}^{r} \) can lie anywhere on the unit interval, such that \( f_k \) is continuously valued in \([0,1)\).

In an MMV model, we consider \( L \) signals, stacked in a matrix, \( X = [x_1, \ldots, x_L] \), where each \( x_l \in \mathbb{C}^n \), \( l = 1, \ldots, L \), is composed of

\[
x_l = \sum_{k=1}^{r} c_{k,l} a(f_k) = V c_l,
\]

with \( c_l = [c_{1,l}, \ldots, c_{r,l}]^T \). Hence \( X \) can be expressed as

\[
X = VC,
\]

where \( C = [c_1 \cdots c_L] \in \mathbb{C}^{r \times L} \).

2.2 Atomic Norm of the MMV Model

We follow the general recipe proposed in \[17\] to define the atomic norm of a spectrally-sparse signal ensemble \( X \). We first define an atom for representing (4) as

\[
A(f, b) = a(f) b^*,
\]

where \( f \in [0,1), b \in \mathbb{C}^L \) with \( \|b\|_2 = 1 \). The set of atoms is given as \( \mathcal{A} = \{A(f, b) | f \in [0,1), \|b\|_2 = 1 \} \).

Define

\[
\|X\|_{A,0} = \inf_r \left\{ X = \sum_{k=1}^{r} c_k A(f_k, b_k), c_k \geq 0 \right\},
\]

as the smallest number of atoms to describe \( X \). A natural objective to describe \( X \) is to minimize \( \|X\|_{A,0} \), i.e., to seek the atomic decomposition of \( X \) with the minimal number of atoms. It is easy to show that \( \|X\|_{A,0} \) can be represented equivalently as \[19\]

\[
\|X\|_{A,0} = \inf_{u,W} \left\{ \text{rank}(\mathcal{T}(u)) \mid \mathcal{T}(u) X^* W \succeq 0 \right\},
\]
where $T(u)$ is the Hermitian Toeplitz matrix with vector $u$ as its first column. Since minimizing (6) is NP-hard, we consider the convex relaxation of $\|X\|_{\mathcal{A},0}$, called the atomic norm of $X$, as

$$\|X\|_{\mathcal{A}} = \inf \{ t > 0 : X \in t \operatorname{conv}(\mathcal{A}) \}$$

$$= \inf \left\{ \sum_k c_k |X = \sum_k c_k A(f_k, b_k), c_k \geq 0 \right\},$$

(7)

where $\operatorname{conv}(\mathcal{A})$ is the convex hull of $\mathcal{A}$. The atomic norm of a single vector $x_l$ defined in [19] becomes a special case of (7) for $L = 1$.

Encouragingly, the atomic norm $\|X\|_{\mathcal{A}}$ admits the following equivalent SDP characterization, which may be solved efficiently. The proof can be found in Appendix A.

Theorem 1. The atomic norm $\|X\|_{\mathcal{A}}$ can be written equivalently as

$$\|X\|_{\mathcal{A}} = \inf_{u \in \mathbb{C}^n, W \in \mathbb{C}^{L \times L}} \left\{ \frac{1}{2} \operatorname{Tr}(T(u)) + \frac{1}{2} \operatorname{Tr}(W) \left[ \begin{array}{cc} T(u) & X \\ X^* & W \end{array} \right] \succeq 0 \right\},$$

where $\operatorname{Tr}(X)$ is the trace of $X$.

3 Line Spectrum Estimation and Denoising with MMV

In this section, we consider two problems based on atomic norm minimization: (a) signal recovery from their partial observations; and (b) signal denoising from their full observations in AWGN.

3.1 Signal Recovery from Partial Observations for MMV Model

We assume that a random or deterministic (sub)set of entries of each vector in $X$ defined in (4) is observed, and the observation pattern is denoted by $\Omega \subset \{0, \ldots, n-1\} \times \{1, \ldots, L\}$. In the absence of noise, the partially observed signal matrix is given as

$$Z_{\tilde{\Omega}} = X_{\tilde{\Omega}} = P_{\tilde{\Omega}}(X),$$

(8)

where $P_{\tilde{\Omega}}$ is a projection matrix on the set indexed by $\tilde{\Omega}$. Note that we allow the observation pattern of each vector of $X$ to be different, possibly randomly selected.

We propose the following atomic norm minimization algorithm to recover the complete signal $X$:

$$\hat{X} = \arg\min_X \|X\|_{\mathcal{A}} \quad \text{s.t.} \quad X_{\tilde{\Omega}} = Z_{\tilde{\Omega}}.$$  

(9)

When the measurements are corrupted by noise, give as

$$Z_{\tilde{\Omega}} = X_{\tilde{\Omega}} + N_{\tilde{\Omega}},$$

where $N_{\tilde{\Omega}}$ is the noise term, we consider the atomic norm regularized algorithm:

$$\hat{X} = \arg\min_X \frac{1}{2} \|X_{\tilde{\Omega}} - Z_{\tilde{\Omega}}\|_F^2 + \tau \|X\|_{\mathcal{A}},$$

(10)

where $\tau$ is a carefully-selected regularization parameter. For now we only discuss the noiseless algorithm (9) and later discuss (10) with full observations in Section 3.2.
From Theorem 1, we can now equivalently write (9) as the following semidefinite program:

\[
\hat{X} = \arg\min_{X, W} \inf_u \frac{1}{2} \text{Tr}(T(u)) + \frac{1}{2} \text{Tr}(W) \\
\text{s.t. } \begin{bmatrix} T(u) & X \\ X^* & W \end{bmatrix} \succeq 0, X_{\bar{\Omega}} = Z_{\bar{\Omega}}. \tag{11}
\]

Similarly, (10) can be recast as a semidefinite program as well.

Interestingly, one can recover the set of frequencies from the solution of the dual problem of (9). Define \( \langle Y, X \rangle = \text{Tr}(X^*Y) \), and \( \langle Y, X \rangle_R = \text{Re}(\langle Y, X \rangle) \). The dual norm of \( \|X\|_A \) can be defined as

\[
\|X\|_A^* = \sup_{\|X\|_A \leq 1} \langle Y, X \rangle_R = \sup_{f \in [0,1], \|b\| = 1} \langle b, Y^*a(f) \rangle = \sup_{f \in [0,1]} \|Y^*a(f)\|_2 = \sup_{f \in [0,1]} \|Q(f)\|_2.
\]

Denote the optimal solution of (9) as \( \hat{X} \). The dual problem of (9) can be written as

\[
\max_Y \langle Y_{\bar{\Omega}}, X^*_{\bar{\Omega}} \rangle_R \text{ s.t. } \|Y\|_A^* \leq 1, Y_{\bar{\Omega}^c} = 0. \tag{12}
\]

Let \((X, Y)\) be primal-dual feasible to (9) and (12), we have \( \langle Y, X \rangle_R = \langle Y, X^* \rangle_R \). Strong duality holds since Slater’s condition holds, and it implies that the solutions of (9) and (12) equal if and only if \( Y \) is dual optimal and \( X \) is primal optimal [40]. Using strong duality, we can obtain a dual certification to the optimality of the solution of (9).

**Proposition 1.** The solution of (9) \( \hat{X} = X^* \) is the unique optimizer if there exists a \( Q(f) = Y^*a(f) \) such that

\[
Q(f_k) = b_k, \quad \forall f_k \in \mathcal{F}, \\
\|Q(f)\|_2 < 1, \quad \forall f \notin \mathcal{F}, \\
Y_{\bar{\Omega}^c} = 0. \tag{13}
\]

The proof can be found in Appendix B. Proposition 1 indicates that one can recover the set of frequencies using the dual polynomial \( \|Q(f)\|_2 = \|Y^*a(f)\|_2 \) constructed from the dual solution \( Y \) by identifying frequencies that satisfy \( \{ f \in [0,1) : \|Q(f)\|_2 = 1 \} \).

**Proposition 1** offers a way to certify the optimality of (9) as long as we can find a dual polynomial \( Q(f) \) that satisfies (13). Borrowing the dual polynomials constructed for the single measurement vector case in [19], we can easily show that the atomic norm minimization for MMV models succeeds with high probability under the same frequency separation condition when the size of \( \bar{\Omega} \) exceeds certain threshold. For completeness, we include the following theorem.

**Theorem 2.** Let \( \bar{\Omega} \) be a set of indices selected uniformly at random from \( \{0, \ldots, n - 1\} \times \{1, \ldots, L\} \). Additionally, assume the signs \( c_{k,l}/|c_{k,l}| \) are drawn i.i.d. from the uniform distribution on the complex unit circle and that

\[
\Delta := \min_{k \neq l} |f_k - f_l| \geq \frac{1}{[(n-1)/4]} \tag{14}
\]
which is the minimum separation between frequency pairs wrapped around on the unit circle. Then there exists a numerical constant $C$ such that

$$|\bar{\Omega}| \geq CL \max \left\{ \log^2 \frac{n}{\delta}, r \log \frac{r}{\delta} \log \frac{n}{\delta} \right\}$$

is sufficient to guarantee that we can recover $X$ via (9) with probability at least $1 - L\delta$ respect to the random samples and signs.

From Theorem 2 we can see that the atomic norm minimization succeeds with high probability as soon as the number of samples is slightly above the information-theoretical lower bound $\Theta(rL)$ by logarithmic factors, given a mild separation condition is satisfied. Theorem 2 is a straightforward extension of the single vector case $L = 1$ studied in [19], by constructing each row of $Q(f)$ in the same manner as [19], hence the proof is omitted. It will be interesting to see whether one can relax either (14) or (15) given more measurement vectors. We demonstrate in the numerical examples that indeed the inclusion of multiple vectors can improve the reconstruction performance in Section 5.

Remark 1. (Connection to EMaC Algorithm) Chen and Chi proposed an algorithm called Enhanced Matrix Completion (EMaC) to recover a partially observed spectrally-sparse signal [22, 23]. Consider the matrix pencil constructed from $x$ in (1) as

$$H(x, p) = \begin{bmatrix} x_1 & x_2 & \cdots & x_{n-p+1} \\ x_2 & x_3 & \cdots & x_{n-p+2} \\ \vdots & \vdots & \ddots & \vdots \\ x_p & x_{p+1} & \cdots & x_n \end{bmatrix},$$

where $p$ is a pencil parameter. In [22, 23], by choosing $p = \lceil n/2 \rceil = \Theta(n)$, the Hankel matrix $H(x, p)$ is a low-rank matrix satisfying $\text{rank}(H(x, p)) \leq r$. Therefore it is proposed to recover $x$ by minimizing the nuclear norm of $H(x, p)$ as

$$\hat{x} = \arg\min_{x} \|H(x, p)\|_* \quad \text{s.t.} \quad x_\Omega = z_\Omega.$$  

Equivalently, (17) can be reformulated as

$$\min_{W_1, W_2, x} \text{Tr}(W_1) + \text{Tr}(W_2)$$

$$\text{s.t.} \begin{bmatrix} W_1 & H(x, p) \\ H(x, p)^* & W_2 \end{bmatrix} \succeq 0, \ x_\Omega = z_\Omega.$$  

Interestingly, it is possible to regard each column of $H(x, p)$ as a spectrally-sparse signal that shares the frequencies. We may propose to minimize the atomic norm of $H(x, p)$ as

$$\hat{x}_A = \arg\min_{x} \|H(x, p)\|_A \quad \text{s.t.} \quad x_\Omega = z_\Omega,$$

which can be reformulated as

$$\min_{u, W_2, x} \text{Tr}(T(u)) + \text{Tr}(W_2)$$

$$\text{s.t.} \begin{bmatrix} T(u) & H(x, p) \\ H(x, p)^* & W_2 \end{bmatrix} \succeq 0, \ x_\Omega = z_\Omega.$$  

Comparing (18) and (20), the EMaC algorithm can be regarded as a relaxation of (20) by dropping the Toeplitz constraint (which allows handling of damping modes) of the first diagonal block in EMaC. When $p = 1$, (20) is equivalent to the atomic norm minimization algorithm for $L = 1$.

\[\text{Remark 1. (Connection to EMaC Algorithm) Chen and Chi proposed an algorithm called Enhanced Matrix Completion (EMaC) to recover a partially observed spectrally-sparse signal [22, 23]. Consider the matrix pencil constructed from } x \text{ in (1) as}
\]

$$H(x, p) = \begin{bmatrix} x_1 & x_2 & \cdots & x_{n-p+1} \\ x_2 & x_3 & \cdots & x_{n-p+2} \\ \vdots & \vdots & \ddots & \vdots \\ x_p & x_{p+1} & \cdots & x_n \end{bmatrix},$$

\[\text{where } p \text{ is a pencil parameter. In [22, 23], by choosing } p = \lceil n/2 \rceil = \Theta(n), \text{ the Hankel matrix } H(x, p) \text{ is a low-rank matrix satisfying } \text{rank}(H(x, p)) \leq r. \text{ Therefore it is proposed to recover } x \text{ by minimizing the nuclear norm of } H(x, p) \text{ as}
\]

$$\hat{x} = \arg\min_{x} \|H(x, p)\|_* \quad \text{s.t.} \quad x_\Omega = z_\Omega.$$  

\[\text{Equivalently, (17) can be reformulated as}
\]

$$\min_{W_1, W_2, x} \text{Tr}(W_1) + \text{Tr}(W_2)$$

$$\text{s.t.} \begin{bmatrix} W_1 & H(x, p) \\ H(x, p)^* & W_2 \end{bmatrix} \succeq 0, \ x_\Omega = z_\Omega.$$  

\[\text{Interestingly, it is possible to regard each column of } H(x, p) \text{ as a spectrally-sparse signal that shares the frequencies. We may propose to minimize the atomic norm of } H(x, p) \text{ as}
\]

$$\hat{x}_A = \arg\min_{x} \|H(x, p)\|_A \quad \text{s.t.} \quad x_\Omega = z_\Omega,$$

\[\text{which can be reformulated as}
\]

$$\min_{u, W_2, x} \text{Tr}(T(u)) + \text{Tr}(W_2)$$

$$\text{s.t.} \begin{bmatrix} T(u) & H(x, p) \\ H(x, p)^* & W_2 \end{bmatrix} \succeq 0, \ x_\Omega = z_\Omega.$$  

\[\text{Comparing (18) and (20), the EMaC algorithm can be regarded as a relaxation of (20) by dropping the Toeplitz constraint (which allows handling of damping modes) of the first diagonal block in EMaC. When } p = 1, \text{ (20) is equivalent to the atomic norm minimization algorithm for } L = 1.
\]
Remark 2. We briefly review the conventional CS approaches for completeness. In CS, it is assumed that $x$ is sparse in a DFT basis or a fine-grained DFT frame, and they share the same sparsity pattern. Hence, the signal $X$ is modeled as
\[ X = F\Theta, \]  
(21)
where $\Theta = [\theta_1, \ldots, \theta_n]^* \in \mathbb{C}^{n \times L}$, and the number of nonzero rows of $\Theta$ is small. Define the group sparsity $\ell_1$ norm of $\Theta$ as
\[ \| \Theta \|_{2,1} = \sum_{i=1}^{n} \| \theta_i \|_2. \] 
A convex optimization algorithm [26] that motivates group sparsity can be posed to solve the MMV model as
\[ \hat{\Theta} = \arg\min_{\Theta} \| \Theta \|_{2,1} \quad \text{s.t.} \quad P_{\bar{\Omega}} (F\Theta - Z) = 0. \]  
(22)
The signal then is recovered as $\hat{X} = F\hat{\Theta}$. However, when the frequencies $F$ are off-the-grid, the model (21) becomes highly inaccurate due to spectral leakage along the Dirichlet kernel, losing the performance guarantee of (22).

3.2 Signal Denoising for MMV Model

In this section, we consider the problem of line spectrum denoising in AWGN when full measurements are available. The algorithm (10) can be rewritten as
\[ \hat{X} = \arg\min_X \frac{1}{2} \| X - Z \|_F^2 + \tau \| X \|_A, \] 
(23)
where the subscript $\bar{\Omega}$ is dropped with $Z = X^* + N$ and $N$ is the additive noise. This algorithm can be efficiently implemented via ADMM, of which we provide the procedure in Appendix E. We have the following theorem for the expected convergence rate of (23) when the noise is AWGN. The proof is in Appendix C.

**Theorem 3.** Assume the entries of $N$ are composed of i.i.d. Gaussian entries $\mathcal{CN}(0, \sigma^2)$. Set $\tau = \sigma \left( 1 + \frac{1}{\log n} \right)^{\frac{1}{2}} \left( L \log (\alpha L) + \sqrt{2L \log (\alpha L)} + \sqrt{\frac{\pi L}{2}} + 1 \right)^{\frac{1}{2}}$, where $\alpha = 8\pi n \log n$, then the expected convergence rate is bounded as
\[ \frac{1}{L} \mathbb{E} \| \hat{X} - X^* \|_F^2 \leq \frac{2\tau}{L} \| X^* \|_A. \]  
(24)

From Theorem 3, $\tau$ is set on the order of $\sqrt{L}$. If $\| X^* \|_A = o(\sqrt{L})$, then the per-measurement vector Mean-Squared Error (MSE) vanishes as $L$ increases. This is satisfied, for example by a correlated signal ensemble where the norm of each row of coefficient matrix $C$ is $o(\sqrt{L})$. On the other hand, if all entries in $C$ are selected with unit amplitude, then $\| X^* \|_A = O(\sqrt{L})$ and the per-measurement vector MSE may not vanish with the increase of $L$. Nonetheless, our numerical examples in Section 5.2 demonstrate that the per-measurement vector MSE decreases gracefully with the increase of $L$.

4 Structured Covariance Estimation for MMV Model

While the availability of MMV improves the performance as demonstrated in Section 3, the computational cost also increases dramatically when $L$ is large. In many applications, one is only interested in the set of frequencies, and the covariance matrix of the signal carries sufficient information to recover the frequencies. In this section, we develop a covariance estimation algorithm based on structured matrix completion that takes advantages of statistical properties of the frequency coefficients and the low-dimensional structures of the covariance matrix.
In particular, we assume that the coefficients \( c_{k,l} \)'s satisfy
\[
\mathbb{E}[c_{k,l}] = 0
\]
and the following second-order statistical property:
\[
\mathbb{E}[c_{k,l}c_{k',l'}] = \begin{cases} \sigma_k^2, & \text{if } k = k', \ l = l', \\ 0, & \text{otherwise}. \end{cases}
\] (25)

To put it differently, the coefficients from different signals are uncorrelated, and the coefficients for different frequencies in the same signal are also uncorrelated. As an example, (25) is satisfied if \( c_{k,l} \)'s are generated i.i.d. from \( \mathcal{CN}(0, \sigma_k^2) \).

Assume each vector in \( \mathbf{X} \) is observed at the same location \( \Omega \) of size \( m \). Without ambiguity, we use \( \Omega \) to denote both the observation pattern of the signal ensemble \( \mathbf{X}_\Omega \) and each signal \( \mathbf{x}_{\Omega,l} \). Instead of focusing on reconstructing the complete signal matrix \( \mathbf{X} \), we explore the low-dimensional structure of its covariance matrix. Given (25), it is straightforward that the covariance matrix of the signal \( \mathbf{x}_l \) in (3) can be written as
\[
\Sigma^* = \mathbb{E} [\mathbf{x}_l \mathbf{x}_l^*] = \sum_{k=1}^{r} \sigma_k^2 \mathbf{a}(f_k) \mathbf{a}(f_k)^*,
\] (26)
which is a positive semidefinite (PSD) Hermitian Toeplitz matrix. This matrix is low-rank with rank \( (\Sigma^*) = r \ll n \). In other words, the spectral sparsity translates into the small rank of the covariance matrix. Let the first column of \( \Sigma^* \) be \( \mathbf{u}^* = \frac{1}{\sqrt{m}} \sum_{k=1}^{r} \sigma_k^2 \mathbf{a}(f_k) \in \mathbb{C}^n \), then \( \Sigma^* \) can be rewritten as \( \Sigma^* = \mathcal{T}(\mathbf{u}^*) \). From \( \mathbf{u}^* \) or \( \Sigma^* \), the set of frequencies can be estimated accurately by well-studied spectrum estimation algorithms such as MUSIC \([35]\), ESPRIT \([42]\) or other linear prediction methods. Therefore, we can focus ourselves on reconstruction of the covariance matrix \( \Sigma^* \).

### 4.1 Structured Covariance Matrix Estimation

The covariance matrix of the partially observed samples \( \mathbf{x}_{\Omega,l} \) can be given as
\[
\Sigma^*_{\Omega} = \mathbb{E}[\mathbf{x}_{\Omega,l} \mathbf{x}_{\Omega,l}^*] = \mathcal{P}_\Omega(\Sigma^*) \in \mathbb{C}^{m \times m},
\] (27)
where \( \mathcal{P}_\Omega \) is a mask operator that only preserves the submatrix in the rows and columns indexed by \( \Omega \).

If \( \Sigma^*_{\Omega} \) can be perfectly estimated, e.g. using an infinite number of measurement vectors, one might directly seek a low-rank Hermitian Toeplitz matrix \( \mathcal{T}(\mathbf{u}) \) which agrees with \( \Sigma^*_{\Omega} \) restricted on the submatrix indexed by \( \Omega \). Unfortunately, the ideal covariance matrix in (27) cannot be perfectly obtained; rather, we will first construct the sample covariance matrix of the partially observed samples as
\[
\Sigma_{\Omega,L} = \frac{1}{L} \sum_{l=1}^{L} \mathbf{x}_{\Omega,l} \mathbf{x}_{\Omega,l}^* = \frac{1}{L} \mathbf{X}_\Omega \mathbf{X}_\Omega^* \in \mathbb{C}^{m \times m}.
\] (28)

Further denote the sample covariance matrix as \( \Sigma_L = \frac{1}{L} \sum_{l=1}^{L} \mathbf{x}_l \mathbf{x}_l^* \). We then seek a low-rank PSD Hermitian Toeplitz matrix whose restriction on the submatrix indexed by \( \Omega \) is close to the sample covariance matrix \( \Sigma_{\Omega,L} \) in (28). A natural algorithm would be
\[
\hat{\mathbf{u}} = \arg\min_{\mathbf{u} \in \mathbb{C}^n} \frac{1}{2} \left\| \mathcal{P}_\Omega(\mathcal{T}(\mathbf{u}) - \hat{\Sigma}_{\Omega,L}\right\|_F^2 + \lambda \text{rank}(\mathcal{T}(\mathbf{u}))
\]
s.t. \( \mathcal{T}(\mathbf{u}) \succeq 0, \) (29)

where \( \lambda \) is a regularization parameter balancing the data fitting term and the rank regularization term. However, as directly minimizing the rank is NP-hard, we consider a convex relaxation for rank minimization over the PSD cone, which replaces the rank minimization by trace minimization, resulting in
\[
\hat{\mathbf{u}} = \arg\min_{\mathbf{u} \in \mathbb{C}^n} \frac{1}{2} \left\| \mathcal{P}_\Omega(\mathcal{T}(\mathbf{u}) - \hat{\Sigma}_{\Omega,L}\right\|_F^2 + \lambda \text{Tr}(\mathcal{T}(\mathbf{u}))
\]
s.t. \( \mathcal{T}(\mathbf{u}) \succeq 0. \) (30)

9
The algorithm (30) can be solved efficiently using off-the-shelf semidefinite program solvers. Interestingly, the trace minimization of \( T(u) \) is equivalent to minimizing the atomic norm of \( u \) under a nonnegative constraint since \( \|u\|_A = \text{Tr}(T(u)) \) if \( T(u) \succeq 0 \). Therefore we can regard (30) as an atomic norm regularized algorithm:

\[
\hat{u} = \arg\min_{u \in \mathbb{C}^n} \frac{1}{2} \| P_\Omega (T(u)) - \Sigma_{\Omega,L} \|_F^2 + \lambda \|u\|_A \\
\text{s.t. } T(u) \succeq 0.
\]

The proposed algorithm works with the sample covariance matrix \( \Sigma_{\Omega,L} \) rather than \( X_\Omega \) directly. Therefore our algorithm does not require storing \( X_\Omega \) of size \( mL \), but only \( \Sigma_{\Omega,L} \) of size \( m^2 \), which greatly reduces the necessary memory space when \( m \ll L \) and may be updated online if the measurement vectors arrive sequentially.

It is also worthwhile to compare the proposed algorithm (30) with the correlation-aware method in [33, 34]. The method in [33, 34], when specialized to a unitary linear array, can be regarded as a discretized version of our algorithm (30), where the atoms \( a(f_k) \)'s in the covariance matrix (27) are discretized over a discrete grid. Further numerical comparisons are carried out in Section 5.

### 4.2 Performance Guarantees with Finite Samples

We analyze the performance of (30) under an additional Gaussian assumption, where each \( c_l \) is i.i.d. generated as \( c_l \sim \mathcal{CN}(0, \Lambda) \), and therefore \( x_l \sim \mathcal{CN}(0, \Sigma^*) \). Define the effective rank of a matrix \( \Sigma \) as \( r_{\text{eff}}(\Sigma) = \text{Tr}(\Sigma)/\|\Sigma\| \) which is strictly smaller than \( r \) and allows the signal to be approximately sparse. We have the following theorem.

**Theorem 4.** Suppose that \( c_l \sim \mathcal{CN}(0, \Lambda) \). Let \( \mathbf{u}^* \) be the ground truth. Set

\[
\lambda \geq C \max \left\{ \sqrt{\frac{r_{\text{eff}}(\Sigma_{\Omega}) \log(Ln)}{L}}, \frac{r_{\text{eff}}(\Sigma^*_\Omega) \log(Ln)}{L} \right\} \|\Sigma_{\Omega}^*\|
\]

for some constant \( C \), then with probability at least \( 1 - L^{-1} \), the solution to (30) satisfies

\[
\| T(\hat{u} - \mathbf{u}^*) \|_F \leq 16\sqrt{r}
\]

if \( \Omega \) corresponds to full observation; and

\[
\frac{1}{\sqrt{n}} \|\hat{u} - \mathbf{u}^*\|_F \leq 16\sqrt{r}
\]

if \( \Omega \) is a complete sparse ruler such that the unobserved entries can be deduced from differences of observed ones.

The proof is in Appendix D. Note that the observation set \( \Omega \) is assumed deterministic in Theorem 4.

When full observations are available, our algorithm yields reliable estimate of the covariance matrix as soon as the number of measurement vectors \( L \) is on the order of \( r_{\text{eff}}(\Sigma^*)r \log n \leq r^2 \log n \), which is much smaller than the ambient dimension \( n \). When \( \Omega \) forms a complete sparse ruler, the average per-entry MSE vanishes as soon as \( L \) is on the order of \( r_{\text{eff}}(\Sigma^*_\Omega)r \log n \leq r^2 \log n \).

### 5 Numerical Experiments

In this section, we evaluate the performance of the proposed algorithms (9), (23) and (30). In particular, we examine the influence of the number of measurement vectors and the number of samples per signal on the performance of frequency estimation, and compare the proposed algorithms against several existing approaches.
5.1 Atomic Norm Minimization (9) for MMV Model

Let $n = 64$ and $m = 32$. In each Monte Carlo experiment, we generate a spectrally-sparse signal with $r$ frequencies randomly located in $[0, 1)$ that satisfy a separation condition $\Delta = \min_{k \neq l} |f_k - f_l| \geq 1/n$. This separation condition is slightly weaker than the condition asserted in Theorem 2 to guarantee the success of (9) with high probability. For each frequency component, we randomly generate its amplitudes for each signal. We run (9) and calculate the reconstruction Normalized Mean Squared Error (NMSE) as $\|X - X^*\|_F / \|X^*\|_F$, and claim the experiment is successful if NMSE $\leq 10^{-5}$. For each pair of $r$ and $L$, we run a total of 50 Monte Carlo experiments and output the average success rate. Fig. 1 shows the success rate of reconstruction versus the sparsity level $r$ for $L = 1, 2, \text{ and } 3$ respectively. As we increase $L$, the success rate becomes higher for the same sparsity level.

![Figure 1: Success rate of reconstruction versus the sparsity level $r$ for $L = 1, 2, 3$ when $n = 64, m = 32$ and the frequencies are generated satisfying a separation condition $\Delta \geq 1/n$ for the same observation across signals.](image)

Fig. 2 shows the reconstructed dual polynomial for a randomly generated spectrally-sparse signal with $r = 10, n = 64$ and $m = 32$. The amplitudes are generated randomly with $CN(0, 1)$ entries when no noise is present. It can be seen that although the algorithm achieves perfect recovery with $L = 1$ and $L = 3$, the reconstructed dual polynomial has a much better localization property when $L = 3$.

![Figure 2: The reconstructed dual polynomial for a randomly generated spectrally-sparse signal with $n = 64, r = 10, \text{ and } m = 32$: (a) $L = 1$, (b) $L = 3$.](image)
5.2 Atomic Norm based Denoising \((23)\) for MMV Model

Let \(n = 64\) and the sparsity level \(r = 8\). The frequencies are selected to satisfy the separation condition \(\Delta \geq 1/n\). We generate the coefficient matrix \(C\) with \(c_{k,l} \sim \mathcal{CN}(0, 1)\). The noise matrix \(N\) is randomly generated with \(\mathcal{CN}(0, \sigma^2)\), where \(\sigma = 0.05\). We solve \((23)\) via ADMM and calculate per-measurement vector MSE as \(\| \hat{X} - X^* \|_F^2 / L\). Fig. 3 shows the per-measurement vector MSE of reconstruction from noisy measurements with respect to the number of measurement vectors. The per-measurement vector MSE decreases with increasing of \(L\), which demonstrates the benefit of more accurate denoising results brought by MMV.

![Figure 3: The per-measurement vector MSE of reconstruction versus the number of measurement vectors \(L\) when \(n = 64\), \(r = 8\) and \(\sigma = 0.05\).](image)

We further examine the influence of \(L\) on the accuracy of frequency estimation. Let \(n = 14\) and \(r = 2\). The coefficients \(C\) is generated with each entry \(c_{k,l} \sim \mathcal{CN}(0, 1)\). And each entry in additive noise \(N\) is randomly generated with \(\mathcal{CN}(0, \sigma^2)\), where \(\sigma = 0.3\). For each \(L\), we obtain the frequency estimates from the dual solution of \((23)\). The average MSE of each frequency estimate as \((\hat{f}_k - f_k)^2\), where \(\hat{f}_k\) is the estimate of real frequency \(f_k\), is averaged over 500 Monte Carlo runs. We compare this against the Cramér Rao Bound (CRB), which can be derived from the following Fisher information matrix \(J(f)\):

\[
J(f) = \frac{4\pi^2}{\sigma^2} \sum_{l=1}^{L} \left[ \sum_{i=1}^{n-1} \frac{|c_{1,l}|^2}{\sum_{i=0}^{n-1} |c_{1,l}|^2} e^{j2\pi(f_1-f_2)i} e^{j2\pi(f_2-f_1)i} \right].
\]

Fig. 4 shows the average MSE and the corresponding CRB with respect to the number of measurement vectors \(L\). With the increase of \(L\), the average MSE of frequency estimates approaches to CRB while CRB approaches to 0.

5.3 Structured Covariance Estimation \((30)\) for MMV Model

We conduct several numerical experiments to validate \((30)\). In particular, we examine the influence of the number of measurement vectors \(L\) on the performance of covariance estimation and frequency estimation, and compare \((30)\) against several existing approaches. Unfortunately we can not directly use Theorem 4 to set \(\lambda\) since \(\Sigma^*\) is not known. In all the experiments, we set \(\lambda = 5 \times 10^{-3} / \left( (\log L)^2 \log m \right)\) which gives good performance empirically.

We first examine the influence of \(L\) on estimating the structured covariance matrix. We fix \(n = 64\), and select \(m = 15\) entries uniformly at random from each measurement vector. The frequencies are selected...
uniformly from \([0,1]\), and the coefficients for each frequency are randomly drawn from \(CN(0,1)\). For various number of measurement vectors \(L\) and sparsity level \(r\), we conduct the algorithm (30) and record the NMSE defined as \(\|\hat{u} - u^*\|_2 / \|u^*\|_2\), where \(\hat{u}\) is the estimate obtained from (30) while \(u^*\) is the first column of the true covariance matrix. Each experiment is repeated 50 times, and the average NMSE is calculated. Fig. 5 shows the average NMSE with respect to the sparsity level \(r\) for \(L = 20, 100, 500, 1000\) and 5000. It can be seen that as \(L\) increases, the average NMSE decreases for a fixed sparsity level.

We next examine the influence of \(L\) on frequency estimation using the obtained Toeplitz covariance matrix. This is done in MATLAB via applying the "rootmusic" function with the true model order (i.e. the sparsity level \(r\)). We fix \(n = 64\), and pick \(m = 8\) entries uniformly at random from each measurement vector. Fig. 6(a) shows the ground truth of the set of frequencies, where the amplitude of each frequency is given as the variance in (25). Fig. 6(b)–(d) demonstrate the set of estimated frequencies when \(L = 50, 200, 400\) respectively. As we increase \(L\), the estimates of the frequencies get more accurate, especially at separating close-located frequencies. It is also worth noticing that the amplitudes of the frequencies are not as well estimated, due to the small value of \(m\).
5.4 Comparisons with Existing Approaches

We compare the performance of covariance matrix estimation for different algorithms, including CS using group sparsity [26], the correlation-aware approach [34], atomic norm minimization [9], and structured covariance estimation (30). We will calculate the NMSE as $\|\hat{\Sigma} - \Sigma^*\|_F / \|\Sigma^*\|_F$, where $\hat{\Sigma}$ is the reconstructed covariance matrix, and $\Sigma^*$ is the truth covariance matrix. Since the first three methods aim at recovering the signal matrix $X$, we will construct the estimated covariance matrix as the sample covariance matrix of the reconstructed signal matrix $\hat{X}$ as $\hat{\Sigma} = \hat{X}\hat{X}^*/L$.

Let $n = 64$, $m = 15$, $r = 8$, and the signals are generated in the same manner as before. For CS and correlation-aware method, we assume a DFT frame with an oversampling factor 4. For the correlation-aware method, we empirically set its regularization parameter as $h = 2 \times 10^{-4}/(\log L \cdot \log m)^2$ which gives good performance [34]. Fig. 7 shows the NMSE with respect to the number of measurement vectors $L$ for different algorithms. Both our method and the correlation-aware method outperform the CS and the atomic norm minimization approach, and ours is slightly better than the correlation-aware approach.

Figure 6: Frequency estimation using (30) for different $L$’s when $n = 64$, $m = 8$ and $r = 6$. (a) Ground truth; (b) $L = 50$; (c) $L = 200$; and (d) $L = 400$.

Figure 7: The NMSE with respect to the number of measurement vectors $L$ for different algorithms when $n = 64$, $m = 15$ and $r = 8$. 

14
Figure 8: Frequency estimation with noise-free measurements using different algorithms when $n = 64$ and $L = 400$. First row: $m = 8$, $r = 6$; Second row: $m = 5$, $r = 6$. (a) Ground truth; (b) CS with the DFT frame; (c) Correlation-aware with the DFT frame; (d) Atomic norm minimization; (e) Structured covariance estimation.

We next compare the performance of frequency estimation using CS with a DFT frame, atomic norm minimization, correlation-aware with a DFT frame, and structured covariance estimation. Let $n = 64$, $L = 400$ and $r = 6$. We generate a spectrally sparse ground truth scene in Fig. 8(a) in the same way as Fig. 6(a). Fig. 8(b)–(e) respectively show the estimated frequencies on a unit circle for different methods, with $m = 8$ and $m = 5$ at $\Omega = \{0, 32, 39, 47, 57\}$ respectively in the first row and the second row. The proposed algorithm works well to locate all the frequencies accurately in both cases. Due to the off-the-grid mismatch, CS and correlation-aware techniques predict frequencies on the lattice of the DFT frame, and result in a larger number of estimated frequencies. On the other hand, atomic norm minimization fails to distinguish the two close frequencies and misses one frequency due to insufficient number of measurements per vector.

Then, we repeat the experiment of Fig. 8(a) expect that the signals are corrupted by noise $\mathcal{CN}(0, \sigma^2)$, where $\sigma = 0.2$. Fig. 9 shows the performance of each method in a unit circle. Notice that the proposed algorithm can still work well to locate all the frequencies accurately, despite there is certain inaccuracy on the corresponding amplitude estimation.

6 Conclusion

We proposed an algorithm to estimate off-the-grid frequencies from partially observed samples of multiple spectrally-sparse signals via atomic norm minimization. Two approaches are developed and solved efficiently via semidefinite programming. The first algorithm aims to recover the set of signal ensembles, which has a higher computational cost when the number of measurement vectors is high. The second algorithm aims to recover the structured covariance matrix from the partially observed sample covariance matrix. The set of frequencies can be recovered either via characterization of the dual polynomial, or using directly traditional methods such as MUSIC. The effectiveness of the proposed methods are demonstrated through numerical examples with comparison against existing CS approaches.
A Proof of Theorem 1

Proof. Denote the value of the right hand side as $\|X\|_\mathcal{T}$. Suppose that $X = \sum_{k=1}^r c_k a(f_k) b_k^*$, there exists a vector $u$ such that $T(u) = \sum_{k=1}^r c_k a(f_k) a(f_k)^*$, by the Vandermonde decomposition lemma. It is obvious that

$$
\begin{bmatrix}
T(u) \\
X^* \\
\sum_{k=1}^r c_k b_k b_k^*
\end{bmatrix} =
\begin{bmatrix}
\sum_{k=1}^r c_k a(f_k) a(f_k)^* \\
\sum_{k=1}^r c_k b_k a(f_k)^* \\
\sum_{k=1}^r c_k b_k b_k^*
\end{bmatrix}
= \sum_{k=1}^r c_k \begin{bmatrix}
a(f_k) \\
b_k
\end{bmatrix} \begin{bmatrix}
a(f_k)^* \\
b_k^*
\end{bmatrix} \succeq 0,
$$

and $\frac{1}{2} \text{Tr}(T(u)) + \frac{1}{2} \text{Tr}(W) = \sum_{k=1}^r c_k = \|X\|_A$, therefore $\|X\|_\mathcal{T} \leq \|X\|_A$. On the other hand, suppose that for any $u$ and $W$ that satisfy

$$
\begin{bmatrix}
T(u) \\
X^* \\
W
\end{bmatrix} \succeq 0,
$$

with $T(u) = VDV^*$, $D = \text{diag}(d_i)$, $d_i \geq 0$. It follows that $X$ is in the range of $V$, hence $X = VB$ with the columns of $B^T$ given by $b_i$. Since $W$ is also PSD, $W$ can be written as $W = B^* EB$ where $E$ is also PSD. We now have

$$
\begin{bmatrix}
T(u) \\
X^* \\
W
\end{bmatrix} =\begin{bmatrix}
V \\
B^*
\end{bmatrix} \begin{bmatrix}
D & I \\
I & E
\end{bmatrix} \begin{bmatrix}
V^* \\
B
\end{bmatrix} \succeq 0,
$$

which yields

$$
\begin{bmatrix}
D & I \\
I & E
\end{bmatrix} \succeq 0
$$

and $E \succeq D^{-1}$ by the Schur complement lemma. Now observe

$$
\text{Tr}(W) = \text{Tr}(B^* EB) \geq \text{Tr}(B^* D^{-1} B)
= \text{Tr}(D^{-1} BB^*) = \sum_i d_i^{-1} \|b_i\|^2.
$$
Therefore,

\[
\frac{1}{2} \text{Tr}(T(u)) + \frac{1}{2} \text{Tr}(W) = \frac{1}{2} \text{Tr}(D) + \frac{1}{2} \text{Tr}(W) \\
\geq \sqrt{\text{Tr}(D) \cdot \text{Tr}(W)} \\
\geq \sqrt{\left( \sum_i d_i \right) \left( \sum_i d_i^{-1} \|b_i\|^2 \right)} \\
\geq \sum \|b_i\| \geq \|X\|_A,
\]

which gives \(\|X\|_T \geq \|X\|_A\). Therefore, \(\|X\|_T = \|X\|_A\).

\[\Box\]

\section*{B Proof of Proposition 1}

\textit{Proof.} First, any \(Y\) satisfying (13) is dual feasible. We have

\[
\|X^*\|_A \geq \|X^*\|_A \|Y\|_A^* \\
\geq \langle Y, X^* \rangle_{\mathbb{R}} = \langle Y, \sum_{k=1}^r c_k a(f_k)b_k^* \rangle_{\mathbb{R}} \\
= \sum_{k=1}^r \text{Re} \langle Y, a(f_k)b_k^* \rangle \\
= \sum_{k=1}^r \text{Re} \langle c_k(b_k, Q(f_k)) \rangle \\
= \sum_{k=1}^r \text{Re} \langle c_k(b_k, b_k) \rangle = \sum_{k=1}^r c_k \geq \|X^*\|_A.
\]

Hence \(\langle Y, X^* \rangle_{\mathbb{R}} = \|X^*\|_A\). By strong duality we have \(X^*\) is primal optimal and \(Y\) is dual optimal.

For uniqueness, suppose \(X\) is another optimal solution which has support outside \(F\). It is trivial to justify if \(\hat{X}\) and \(X^*\) have the same support, they must coincide since the set of atoms with frequencies in \(F\) is independent. Let \(X = \sum_k \hat{c}_k a(\hat{f}_k)b_k^*\). We then have the dual certificate

\[
\langle Y, X \rangle_{\mathbb{R}} = \sum_{\hat{f}_k \in F} \text{Re} \langle \hat{c}_k(b_k, Q(\hat{f}_k)) \rangle + \sum_{\hat{f}_l \notin F} \text{Re} \langle \hat{c}_l(b_l, Q(\hat{f}_l)) \rangle \\
< \sum_{\hat{f}_k \in F} \hat{c}_k + \sum_{\hat{f}_l \notin F} \hat{c}_l = \|\hat{X}\|_A,
\]

which contradicts strong duality. Therefore the optimal solution of (9) is unique.

\[\Box\]

\section*{C Proof of Theorem 3}

\textit{Proof.} We first record [44 Proposition 1 and Theorem 1] that applies to our atomic norm denoising formulation.

\textbf{Lemma 1.} \(\text{If } \mathbb{E}\|N\|^*_A \leq \tau, \text{ the solution to (23) satisfies that} \)

\[
\mathbb{E} \left\| \hat{X} - X^* \right\|_F^2 \leq 2\tau \|X^*\|_A.
\]

(31)
Lemma 1 immediately implies that we can characterize the expected convergence rate of the atomic norm denoising algorithm (23) if the behavior of $\mathbb{E} \|N\|_A^*$ can be understood. According to the definition of dual norm, we can write ($\|N\|_A^*$) as

$$\|N\|_A^* = \sup_{f \in [0,1)} \|N^* \alpha(f)\|_2$$

$$= \sup_{f \in [0,1)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{i,t} e^{j2\pi(i-1)f} \right|^2$$

$$= \sup_{f \in [0,1)} \sum_{l=1}^{L} |w_l(f)|^2,$$

(32)

where $\phi_{i,t}$ is the $(i,t)$th entry of $N$, and $w_l(f) \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{i,t} e^{j2\pi(i-1)f}$. For $f_1, f_2 \in [0,1)$, by Bernstein’s theorem [43] and also a partial result in [44, Appendix C], we can obtain that

$$|w_l(f_1)| - |w_l(f_2)| \leq 4\pi n |f_1 - f_2| \sup_{f \in [0,1)} |w_l(f)|.$$  

(33)

Therefore, we can write

$$\sum_{l=1}^{L} |w_l(f_1)|^2 - \sum_{l=1}^{L} |w_l(f_2)|^2$$

$$\leq \sum_{l=1}^{L} (|w_l(f_1)| + |w_l(f_2)|) \left(4\pi n |f_1 - f_2| \sup_{f \in [0,1)} |w_l(f)|\right)$$

(34)

$$\leq 8\pi n |f_1 - f_2| L \sup_{f \in [0,1)} \sum_{l=1}^{L} |w_l(f)|^2$$

$$= 8\pi n |f_1 - f_2| L (\|N\|_A^*)^2,$$

(35)

where (34) follows by plugging in (33), (35) follows from $\sum_{l=1}^{L} \sup_{f \in [0,1)} |w_l(f)|^2 \leq L \sup_{f \in [0,1)} \sum_{l=1}^{L} |w_l(f)|^2$, and the last equality follows from (32).

Let $D$ be a grid size that we will specify later, then by allowing $f_2$ to take any of the $D$ values on the grid points $\{0, \frac{1}{D}, \ldots, \frac{D-1}{D}\}$, we have

$$\|N\|_A^* \leq \max_{d=0,\ldots,D-1} \sum_{l=1}^{L} w_l \left(\frac{d}{D}\right)^2 + \frac{4\pi nL}{D} (\|N\|_A^*)^2.$$  

Thus, $\|N\|_A^*$ can be bounded as

$$\|N\|_A^* \leq \left(1 - \frac{4\pi nL}{D}\right)^{-\frac{1}{2}} \left(\max_{d=0,\ldots,D-1} \sum_{l=1}^{L} w_l \left(\frac{d}{D}\right)^2\right)^{\frac{1}{2}}.$$  

(36)

Denote $Q_d \triangleq \frac{2}{\pi^2} \sum_{l=1}^{L} |w_l \left(\frac{d}{D}\right)|^2$ which is a chi-squared random variable with $2L$ degrees of freedom. We first analyze $\mathbb{E}(\|N\|_A^*)$. From (36), we have that

$$\mathbb{E}(\|N\|_A^*) \leq \left(1 - \frac{4\pi nL}{D}\right)^{-\frac{1}{2}} \left(\frac{\sigma^2}{2}\right)^{\frac{1}{2}} \left(\mathbb{E}\left[\max_{d=0,\ldots,D-1} Q_d\right]\right)^{\frac{1}{2}}$$

$$\leq \left(\frac{\sigma^2}{2}\right)^{\frac{1}{2}} \left(1 + \frac{8\pi nL}{D}\right)^{\frac{1}{2}} \left(\mathbb{E}\left[\max_{d=0,\ldots,D-1} Q_d\right]\right)^{\frac{1}{2}}.$$  

(37)
Note that
\[
E \left[ \max_{d=0, \ldots, D-1} Q_d \right] = \int_0^\infty P \left[ \max_{d=0, \ldots, D-1} Q_d \geq t \right] dt \\
\leq \delta + \int_\delta^\infty P \left[ \max_{d=0, \ldots, D-1} Q_d \geq t \right] dt \\
\leq \delta + D \int_\delta^\infty P \{ Q_d \geq t \} dt,
\]
where the last line follows by the union bound. Recall the following lemma which bounds the tail behavior of a chi-squared random variable.

**Lemma 2.** Let \( U \) be a standard chi-squared distribution of degrees of freedom \( 2L \), for any \( \gamma > 0 \), we have
\[
P \left[ U \geq (1 + \gamma + \frac{1}{2} \gamma^2)2L \right] \leq \exp \left( -\frac{L}{2} \gamma^2 \right).
\]

Plugging (39) into (38), we obtain
\[
E \left[ \max_{d=0, \ldots, D-1} Q_d \right] \leq \delta + 2LD \int_\delta^\infty P \left[ Q_d \geq (1 + \gamma + \frac{1}{2} \gamma^2)2L \right] (1 + \gamma) d\gamma \\
\leq \delta + 2LD \int_{1+\frac{L}{\sqrt{D}}(\delta-L)}^\infty \exp \left( -\frac{L}{2} \gamma^2 \right) (1 + \gamma) d\gamma \\
= \delta + 2L \sqrt{D} \pi Q \left( -\sqrt{L} + \sqrt{\delta - L} \right) + 2DE^{-\frac{1}{2} \delta} \left( 1 + \frac{1}{2} \sqrt{L(\delta - L)} \right)^2 \\
\leq \delta + 2L \sqrt{D} \pi Q \left( -\sqrt{L} + \sqrt{\delta - L} \right) + 2DE^{-\frac{1}{2} \delta} \left( 1 + \frac{1}{2} \sqrt{L(\delta - L)} \right)^2 \\
= \delta + \left( \sqrt{2\pi \sqrt{L}} + 2 \right) DE^{-\frac{1}{2} \delta} \left( 1 + \frac{1}{2} \sqrt{L(\delta - L)} \right)^2,
\]
where (40) follows by letting \( t = (1 + \gamma + \frac{1}{2} \gamma^2)2L \), (41) follows from (39), (42) follows by straight calculations using the definition of the \( Q \) function, and (43) follows by the Chernoff bound \( Q(x) \leq \frac{1}{2} e^{-\frac{x^2}{2}} \).

Let \( -\frac{1}{2} \left( -\sqrt{L} + \sqrt{\delta - L} \right)^2 = -\log D \), i.e. \( \delta = 2L + 2 \log D + 2\sqrt{2L \log D} \), then we can obtain that
\[
E \left[ \max_{d=0, \ldots, D-1} Q_d \right] \leq 2L + 2 \log D + 2\sqrt{2L \log D} + \sqrt{2\pi L} + 2.
\]
Therefore, by plugging this into (37) and letting \( D = 8\pi nL \log n \), we obtain
\[
E \left[ \| N \|_\ast^2 \right] \leq \left( \frac{a^2}{2} \right)^{\frac{1}{2}} \left( 1 + \frac{1}{\log n} \right)^{\frac{1}{2}} \left( 2L + 2 \log 8\pi nL \log n + 2\sqrt{2L \log 8\pi nL \log n} + \sqrt{2\pi \sqrt{L}} + 2 \right)^{\frac{1}{2}}.
\]
The proof is completed by setting the right hand side as \( \tau \).

\[\square\]

**D Proof of Theorem 4**

**Proof.** As the trace norm is equivalent to the nuclear norm \( \| \cdot \|_\ast \) for PSD matrices, we consider the equivalent algorithm
\[
\hat{u} = \arg \min_u \frac{1}{2} \| \mathcal{P}_\Omega (u) - \Sigma_{\Omega, L} \|_F^2 + \lambda \| T(u) \|_\ast.
\]

19
Denote the tangent space of $\mathcal{T}(u)$ spanned its column and row space as $T$, and its orthogonal tangent space as $T^\perp$. Decompose the error term $\mathcal{T}(\hat{u} - u^*) = H_1 + H_2$ into two terms satisfying $\text{rank}(H_1) \leq 2r$ and $H_2 \in T^\perp$ [46]. Rephrasing straightforwardly [46] Lemma 1, we have that as long as $\lambda \geq \|\Sigma_{\Omega}^* - \Sigma_{\Omega,T}\|$, where $\Sigma_{\Omega}^* = \mathcal{P}_\Omega(\mathcal{T}(u^*))$,

$$
\|H_2\|_* \leq 3\|H_1\|_* .
$$

(45)

To obtain a reasonable regularization parameter $\lambda$, we utilize the following bound in [47].

**Lemma 3 ( [47] ).** Suppose that $x_i$ is a Gaussian random vector with mean zero and covariance $\Sigma$. Define the sample covariance matrix $\Sigma_L = \frac{1}{L} \sum_{i=1}^L x_ix_i^T$. Then with probability at least $1 - L^{-1}$,

$$
\|\Sigma_L - \Sigma\| \leq C \max \left\{ \sqrt{\frac{r_{\text{eff}}(\Sigma) \log(Ln)}{L}}, \frac{r_{\text{eff}}(\Sigma) \log(Ln)}{L} \right\} \|\Sigma\|,
$$

(46)

for some constant $C$.

Instantiating (46) we have with probability at least $1 - L^{-1}$,

$$
\|\Sigma_{\Omega,L}^* - \Sigma_{\Omega,L}\| \leq C \max \left\{ \sqrt{\frac{r_{\text{eff}}(\Sigma_{\Omega,L}^*) \log(Ln)}{L}}, \frac{r_{\text{eff}}(\Sigma_{\Omega,L}^*) \log(Ln)}{L} \right\} \|\Sigma_{\Omega,L}^*\| =: \lambda.
$$

(47)

Now by the triangle inequality,

$$
\|\mathcal{T}(u^*)\|_* = \|\mathcal{T}(u^*) - \mathcal{T}(\hat{u}) + \mathcal{T}(\hat{u})\|_* \leq \|H\|_* + \|\mathcal{T}(\hat{u})\|_* ,
$$

and by the optimality of $\hat{u}$:

$$
\frac{1}{2}\|\mathcal{P}_\Omega(\mathcal{T}(\hat{u})) - \Sigma_{\Omega,L}\|_F^2 + \lambda\|\mathcal{T}(\hat{u})\|_* \leq \frac{1}{2}\|\mathcal{P}_\Omega(\mathcal{T}(u^*)) - \Sigma_{\Omega,L}\|_F^2 + \lambda\|\mathcal{T}(u^*)\|_* ,
$$

which gives

$$
\lambda\|\mathcal{T}(u^*) - \mathcal{T}(\hat{u})\|_* \geq \frac{1}{2}\|\mathcal{P}_\Omega(\mathcal{T}(\hat{u})) - \Sigma_{\Omega,L}\|_F^2 - \frac{1}{2}\|\mathcal{P}_\Omega(\mathcal{T}(u^*)) - \Sigma_{\Omega,L}\|_F^2 .
$$

(48)

Further since

$$
\|\Sigma_{\Omega,L} - \mathcal{P}_\Omega(\mathcal{T}(\hat{u}))\|_F^2 = \|\Sigma_{\Omega,L} - \mathcal{P}_\Omega(\mathcal{T}(\hat{u}))\|_F^2 - \|\Sigma_{\Omega,L} - \mathcal{P}_\Omega(\mathcal{T}(u^*))\|_F^2 - 2\langle \Sigma_{\Omega,L} - \Sigma_{\Omega,L}^* , \mathcal{T}(u^*) - \mathcal{T}(\hat{u}) \rangle ,
$$

which gives

$$
\|\mathcal{P}_\Omega(\mathcal{T}(\hat{u} - u^*))\|_F^2 \leq \|\Sigma_{\Omega,L} - \mathcal{P}_\Omega(\mathcal{T}(\hat{u}))\|_F^2 - \|\Sigma_{\Omega,L} - \mathcal{P}_\Omega(\mathcal{T}(u^*))\|_F^2 - 2\|\Sigma_{\Omega,L} - \Sigma_{\Omega,L}^* , \mathcal{T}(u^*) - \mathcal{T}(\hat{u}) \|_*
\leq 2\lambda\|\mathcal{T}(u^*) - \mathcal{T}(\hat{u})\|_* + 2\|\Sigma_{\Omega,L} - \Sigma_{\Omega,L}^* \| \cdot \|\mathcal{T}(u^*) - \mathcal{T}(\hat{u})\|_*
\leq 4\lambda\|\mathcal{T}(u^*) - \mathcal{T}(\hat{u})\|_*
\leq 16\lambda\|H_1\|_* \leq 16\sqrt{r}\|H_1\|_F \leq 16\sqrt{r}\|\mathcal{T}(\hat{u} - u^*)\|_F ,
$$

where the first inequality follows from (48) and the Cauchy-Schwartz inequality, the second inequality follows from (47), and the third inequality follows from (45). We consider two cases:

- With full observation $\mathcal{P}_\Omega(\mathcal{T}(\hat{u} - u^*)) = \mathcal{T}(\hat{u} - u^*)$, we have $\|\mathcal{T}(\hat{u} - u^*)\|_F \leq 16\sqrt{r}$. 

20
• When $\Omega$ is a complete sparse ruler, we have

$$\|P_\Omega(T(\hat{u} - u^*))\|_F^2 \geq \|\hat{u} - u^*\|_F^2,$$

which gives

$$\|\hat{u} - u^*\|_F^2 \leq 16\lambda\sqrt{r}\|T(\hat{u} - u^*)\|_F \leq 16\lambda\sqrt{rn}\|\hat{u} - u^*\|_F.$$

Therefore we have $\frac{1}{\sqrt{n}}\|\hat{u} - u^*\|_F \leq 16\lambda\sqrt{r}$.

\[ \square \]

E ADMM Implementation of (23)

In order to apply ADMM \[31\], we reformulate (23) as

$$\min X \quad \frac{1}{2} \|X - Z\|_F^2 + \frac{\tau}{2} (\text{Tr}(T(u)) + \text{Tr}(W))$$

subject to

$$Y = \begin{bmatrix} T(u) & X \\ X^* & W \end{bmatrix}, Y \succeq 0,$$

whose augmented Lagrangian can then be cast as

$$\Psi(X, u, W, \Lambda, Y) = \frac{1}{2} \|X - Z\|_F^2 + \frac{\tau}{2} (\text{Tr}(T(u)) + \text{Tr}(W))$$

$$+ \begin{langle} \Lambda, Y - \begin{bmatrix} T(u) & X \\ X^* & W \end{bmatrix} \end{rangle}$$

$$+ \frac{\rho}{2} \|Y - \begin{bmatrix} T(u) & X \\ X^* & W \end{bmatrix}\|_F^2,$$

where $Y$, $W$ and $\Lambda$ are all Hermitian matrices. For notation simplicity, let $\Lambda = \begin{bmatrix} \Lambda_{n \times n} & \Lambda_{n \times L} \\ \Lambda_{L \times n} & \Lambda_{L \times L} \end{bmatrix}$. Then the update steps of ADMM are as follows

$$(X^{t+1}, u^{t+1}, W^{t+1}) = \arg\min_{X, u, W} \Psi(X, u, W, \Lambda^t, Y^t);$$

$$Y^{t+1} = \arg\min_{Y \succeq 0} \Psi(X^{t+1}, u^{t+1}, W^{t+1}, \Lambda^t, Y);$$

$$\Lambda^{t+1} = \Lambda^t + \rho \left( Y^{t+1} - \begin{bmatrix} T(u^{t+1}) & X^{t+1} \\ X^{t+1*} & W^{t+1} \end{bmatrix}\right),$$

where the superscript $t$ denotes the $t$th iteration. Fortunately, closed-form solutions to the above updates exist and can be given as

$$W^{t+1} = \frac{1}{2} Y_{L \times L}^t + \frac{1}{2}(Y_{L \times L}^t)^* + \frac{1}{\rho} \left( \Lambda_{L \times L}^t - \frac{\tau}{2} I \right);$$

$$X^{t+1} = \frac{1}{2\rho + 1} \left( Z + 2(\Lambda_{L \times n}^t)^* + \rho Y_{n \times L}^t + \rho(Y_{L \times n}^t)^* \right);$$

$$u^{t+1} = \frac{1}{\rho} \cdot \mathbf{y} \cdot \text{conj} \left( \mathcal{G}(\Lambda_{n \times n}^t) + \frac{\rho}{2} (\mathcal{G}(Y_{n \times n}^t)^*) + \mathcal{G}(Y_{n \times n}^t) - \frac{\tau}{2} n e_1 \right),$$

21
where \( \text{conj}(\cdot) \) means the conjugate operation on each entry of a vector or a matrix, \( e_1 \) is the first vector in the standard basis, \( a = \mathcal{G}(A) \) is a mapping from a matrix to a vector where the \( i \)th entry in \( a \) is the sum of all the entries \( A_{p,q} \)'s of \( A \) satisfying \( q - p + 1 = i \), and \( \Upsilon \) is a diagonal matrix with diagonal entries \( \Upsilon_{i,i} = \frac{1}{n-i+1}, \quad i = 1, \ldots, n. \)

Let \( \Xi^t = \begin{bmatrix} T(u^t+1) & X^t+1 \\ X^{*t+1} & W^t+1 \end{bmatrix} - \frac{1}{\rho} \Lambda^t = U^t \text{diag}(\{\sigma^t_i\})(U^t)^* \) be its eigenvalue decomposition, then the update of \( Y \) can be given as

\[
Y^{t+1} = U^t \text{diag}(\{\sigma^t_i\}_+)(U^t)^*.
\]

We run the above iterations until both primal and dual residuals satisfy the pre-set tolerance level.

**References**

[1] Y. Chi, “Joint sparsity recovery for spectral compressed sensing,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[2] Y. Li and Y. Chi, “Compressive parameter estimation with multiple measurement vectors via structured low-rank covariance estimation,” in *Proceedings of Statistical Signal Processing Workshop (SSP)*, Gold Coast, Australia, Jul. 2014.

[3] L. L. Scharf, *Statistical signal processing*. Addison-Wesley Reading, 1991, vol. 98.

[4] E. Candes, 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, February 2006.

[5] D. Donoho, “Compressed sensing,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289 –1306, April 2006.

[6] E. Candes and J. Romberg, “Sparsity and incoherence in compressive sampling,” *Inverse problems*, vol. 23, no. 3, p. 969, 2007.

[7] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, “Beyond nyquist: Efficient sampling of sparse bandlimited signals,” *Information Theory, IEEE Transactions on*, vol. 56, no. 1, pp. 520–544, 2010.

[8] M. Mishali and Y. C. Eldar, “From theory to practice: Sub-nyquist sampling of sparse wideband analog signals,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 2, pp. 375–391, 2010.

[9] Z. Tian and G. B. Giannakis, “Compressed sensing for wideband cognitive radios,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4. IEEE, 2007, pp. IV–1357.

[10] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, “Sensitivity to basis mismatch in compressed sensing,” *Signal Processing, IEEE Transactions on*, vol. 59, no. 5, pp. 2182–2195, 2011.

[11] L. L. Scharf, E. K. Chong, A. Pezeshki, and J. R. Luo, “Sensitivity considerations in compressed sensing,” in *Proceedings of the Asilomar Conference on Signals, Systems and Computers*. IEEE, 2011, pp. 744–748.
[12] P. Pakrooh, L. L. Scharf, A. Pezeshki, and Y. Chi, “Analysis of fisher information and the cram-
erao bound for nonlinear parameter estimation after compressed sensing,” in Proceedings of the IEEE
International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

[13] E. J. Candes, Y. C. Eldar, D. Needell, and P. Randall, “Compressed sensing with coherent and redu-
dant dictionaries,” Applied and Computational Harmonic Analysis, vol. 31, no. 1, pp. 59–73, 2011.

[14] A. Fannjiang and W. Liao, “Coherence pattern-guided compressive sensing with unresolved grids,”
SIAM Journal on Imaging Sciences, vol. 5, no. 1, pp. 179–202, 2012.

[15] H. Zhu, G. Leus, and G. B. Giannakis, “Sparsity-cognizant total least-squares for perturbed compres-
sive sampling,” Signal Processing, IEEE Transactions on, vol. 59, no. 5, pp. 2002–2016, 2011.

[16] K. Fyhnn, M. F. Duarte, and S. H. Jensen, “Compressive parameter estimation for sparse translation-
\textit{invariant signals using polar interpolation},” arXiv preprint arXiv:1305.3483, 2013.

[17] 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.

[18] E. Candes and C. Fernandez-Granda, “Towards a mathematical theory of super-resolution,” March
2012. [Online]. Available: http://arxiv.org/abs/1203.5871

[19] G. Tang, B. Bhaskar, P. Shah, and B. Recht, “Compressed sensing off the grid,” Information Theory,
IEEE Transactions on, vol. 59, no. 11, pp. 7465–7490, Nov 2013.

[20] Y. Chi and Y. Chen, “Compressive recovery of 2-D off-grid frequencies,” in Proceedings of Asilomar
Conference on Signals, Systems, and Computers, Pacific Grove, CA, Nov. 2013.

[21] W. Xu, J.-F. Cai, K. V. Mishra, M. Cho, and A. Kruger, “Precise semidefinite programming formulation
of atomic norm minimization for recovering d-dimensional ($d \geq 2$) off-the-grid frequencies,” arXiv
preprint arXiv:1312.0485, 2013.

[22] Y. Chen and Y. Chi, “Spectral compressed sensing via structured matrix completion,” International
Conference on Machine Learning (ICML), 2013.

[23] ———, “Robust spectral compressed sensing via structured matrix completion,” IEEE Trans. on Infor-
mation Theory, 2014, accepted.

[24] Y. Hua, “Estimating two-dimensional frequencies by matrix enhancement and matrix pencil,” IEEE
Transactions on Signal Processing, vol. 40, no. 9, pp. 2267 –2280, Sep 1992.

[25] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, “Algorithms for simultaneous sparse approximation. part
\text{i: Greedy pursuit},” Signal Processing, vol. 86, no. 3, pp. 572–588, 2006.

[26] J. A. Tropp, “Algorithms for simultaneous sparse approximation. part ii: Convex relaxation,” Signal
Processing, vol. 86, no. 3, pp. 589–602, 2006.

[27] K. Lee, Y. Bresler, and M. Junge, “Subspace methods for joint sparse recovery,” Information Theory,
IEEE Transactions on, vol. 58, no. 6, pp. 3613–3641, 2012.

[28] J. M. Kim, O. K. Lee, and J. C. Ye, “Compressive music: revisiting the link between compressive
sensing and array signal processing,” Information Theory, IEEE Transactions on, vol. 58, no. 1, pp.
278–301, 2012.
[29] M. E. Davies and Y. C. Eldar, “Rank awareness in joint sparse recovery,” Information Theory, IEEE Transactions on, vol. 58, no. 2, pp. 1135–1146, 2012.

[30] M. Mishali and Y. C. Eldar, “Reduce and boost: Recovering arbitrary sets of jointly sparse vectors,” Signal Processing, IEEE Transactions on, vol. 56, no. 10, pp. 4692–4702, 2008.

[31] S. Boyd, N. Parikh, B. Chu, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Foundations and Trends in Machine Learning, vol. 3, no. 1, pp. 1–122, December 2011.

[32] P. J. Tourtier and L. L. Scharf, “Maximum likelihood identification of correlation matrices for estimation of power spectra at arbitrary resolutions,” in Acoustics, Speech, and Signal Processing, 1987. Proceedings.(ICASSP’87). IEEE International Conference on, vol. 12. IEEE, 1987, pp. 2066–2069.

[33] P. Pal and P. Vaidyanathan, “Correlation-aware techniques for sparse support recovery,” in Statistical Signal Processing Workshop (SSP), 2012 IEEE. IEEE, 2012, pp. 53–56.

[34] ——, “On application of lasso for sparse support recovery with imperfect correlation awareness,” in Signals, Systems and Computers (ASILOMAR), 2012 Conference Record of the Forty Sixth Asilomar Conference on. IEEE, 2012, pp. 958–962.

[35] R. Schmidt, “Multiple emitter location and signal parameter estimation,” Antennas and Propagation, IEEE Transactions on, vol. 34, no. 3, pp. 276–280, 1986.

[36] P. Pal and P. Vaidyanathan, “Coprime sampling and the music algorithm,” in Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), 2011 IEEE. IEEE, 2011, pp. 289–294.

[37] S. Shakeri, D. Ariananda, and G. Leus, “Direction of arrival estimation using sparse ruler array design,” in Signal Processing Advances in Wireless Communications (SPAWC), 2012 IEEE 13th International Workshop on, June 2012, pp. 525–529.

[38] Z. Yang, L. Xie, and C. Zhang, “A discretization-free sparse and parametric approach for linear array signal processing,” arXiv preprint arXiv:1312.7695, 2013.

[39] P. Pal and P. Vaidyanathan, “A grid-less approach to underdetermined direction of arrival estimation via low rank matrix denoising,” Signal Processing Letters, IEEE, vol. 21, no. 6, pp. 737 – 741, June 2014.

[40] S. P. Boyd and L. Vandenberghe, Convex optimization. Cambridge university press, 2004.

[41] Z. Yang and L. Xie, “Exact joint sparse frequency recovery via optimization methods,” arXiv preprint arXiv:1405.6585, 2014.

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

[43] C. Carathéodory and L. Fejér, “Über den zusammengang der extemen von harmonischen funktionen mit ihren koeffizienten und über den picard-landauschen satz,” Rendiconti del Circolo Matematico di Palermo, vol. 32, pp. 218–239, 1911.

[44] B. Bhaskar, G. Tang, and B. Recht, “Atomic norm denoising with applications to line spectral estimation,” Signal Processing, IEEE Transactions on, vol. 61, no. 23, pp. 5987–5999, Dec 2013.
[45] A. Schaeffer, “Inequalities of a. markoff and s. bernstein for polynomials and related functions,” *Bull. Amer. Math. Soc*, vol. 47, pp. 565–579, 1941.

[46] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, “A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers,” *Statistical Science*, vol. 27, no. 4, pp. 538–557, 2012.

[47] F. Bunea and L. Xiao, “On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca,” *arXiv preprint arXiv:1212.5321*, 2012.