Super-Resolution Compressed Sensing for Line Spectral Estimation: An Iterative Reweighted Approach

Jun Fang, *Member, IEEE*, Feiyu Wang, Yanning Shen, Hongbin Li, *Senior Member, IEEE*, and Rick S. Blum, *Fellow, IEEE*

***Abstract*—Conventional compressed sensing theory assumes signals have sparse representations in a known dictionary. Nevertheless, in many practical applications such as line spectral estimation, the sparsifying dictionary is usually characterized by a set of unknown parameters in a continuous domain. To apply the conventional compressed sensing technique to such applications, the continuous parameter space has to be discretized to a finite set of grid points, based on which a "nominal dictionary" is constructed for sparse signal recovery. Discretization, however, inevitably incurs errors since the true parameters do not necessarily lie on the discretized grid. This error, also referred to as grid mismatch, leads to deteriorated recovery performance. In this paper, we consider the line spectral estimation problem and propose an iterative reweighted method which jointly estimates the sparse signals and the unknown parameters associated with the true dictionary. The proposed algorithm is developed by iteratively decreasing a surrogate function majorizing a given log-sum objective function, leading to a gradual and interweaved iterative process to refine the unknown parameters and the sparse signal. A simple yet effective scheme is developed for adaptively updating the regularization parameter that controls the tradeoff between the sparsity of the solution and the data fitting error. Theoretical analysis is conducted to justify the proposed method. Simulation results show that the proposed algorithm achieves super resolution and outperforms other state-of-the-art methods in many cases of practical interest.**

***Index Terms*—Super-resolution compressed sensing, line spectra estimation, grid mismatch, iterative reweighted methods.**

Manuscript received December 08, 2015; revised April 07, 2016 and May 10, 2016; accepted May 10, 2016. Date of publication May 24, 2016; date of current version July 25, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Fauzia Ahmad. This work was supported in part by the National Science Foundation of China under Grants 61428103, 61522104, the National Science Foundation under Grants ECCS-1408182 and ECCS-1405579, and the Air Force Office of Scientific Research under award number FA9550-16-1-0243.

J. Fang and F. Wang are with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: JunFang@uestc.edu.cn; 15196637434@163.com).

Y. Shen is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: shenx513@umn.edu).

H. Li is with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030 USA (e-mail: Hongbin.Li@stevens.edu).

R. S. Blum is with the Department of Electrical and Computer Engineering, Lehigh University, Bethlehem, PA 18015 USA (e-mail: rblum@lehigh.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2016.2572041

I. Introduction

COMPRESSED sensing finds a variety of applications in practice as many natural signals admit a sparse or an approximate sparse representation in a certain basis. Nevertheless, accurate reconstruction of a sparse signal relies on the knowledge of the sparsifying dictionary, while in many applications, it is often impractical to pre-specify a dictionary that can sparsely represent the signal. For example, for the line spectral estimation problem, using a preset discrete Fourier transform (DFT) matrix suffers from considerable performance degradation because the true frequency components may not lie on the pre-specified frequency grid. The same is true for direction-of-arrival (DOA) estimation, where the true directions of sources may not align with the presumed grid. In these and similar applications, the sparsifying dictionary is characterized by a set of unknown parameters in a continuous domain. In order to apply compressed sensing to such applications, the continuous parameter space has to be discretized to a finite set of grid points, based on which a nominal dictionary is constructed for sparse signal recovery. Discretization, however, inevitably incurs errors since the true parameters do not necessarily lie on the discretized grid. This error, also referred to as the grid mismatch, leads to deteriorated performance or even failure in recovering the sparse signal. Finer grids can certainly be used to reduce grid mismatch and improve the reconstruction accuracy. Nevertheless, recovery algorithms may become numerically instable and computationally prohibitive when very fine discretized grids are employed.

The grid mismatch problem has attracted a lot of attention over the past few years. Specifically, in [1], the problem was addressed in a general framework of "basis mismatch" where the mismatch is modeled as a perturbation (caused by grid discretization, calibration errors or other factors) between the presumed and the actual dictionaries, and the impact of the basis mismatch on the reconstruction error was analyzed. In [2], [3], to deal with grid mismatch, the true dictionary is approximated as a summation of a presumed dictionary and a structured parameterized matrix via a Taylor expansion. The recovery performance of this method, however, depends on the accuracy of the Taylor expansion in approximating the true dictionary. The grid mismatch problem was also examined in [4], [5], where a highly coherent dictionary (very fine grids) is used to mitigate the discretization error, and a class of greedy algorithms which use the technique of band exclusion (coherence-inhibiting) were proposed for sparse signal recovery. Besides these efforts,

another line of work [6]–[9] studied the problem of grid mismatch in a more fundamental way: they circumvent the discretization issue by working directly on the continuous parameter space, leading to the so-called super-resolution technique. In [6]–[9], an atomic norm-minimization approach was proposed to handle the infinite dictionary with continuous atoms. It was shown that given that the frequency components are sufficiently separated, the frequency components of a mixture of complex sinusoids can be super-resolved with infinite precision from only coarse-scale measurements. Other related works include [10] for the extension of the atomic-norm techniques and [11] based on a low rank Hankel matrix completion. In [12], [13], by treating the sparse signal as hidden variables, Bayesian approaches were proposed to iteratively refine the dictionary, and are shown able to achieve super-resolution. Parametric dictionary learning for sparse signal recovery was also considered in [14], where the objective is to optimize the dictionary parameters in order to minimize the distance of the dictionary to an equiangular tight frame (ETF).

In this paper, we address the line spectral estimation problem in a super-resolution compressed sensing framework[1]. Specifically, line spectral estimation is formulated as a sparse signal recovery problem with an unknown parametric dictionary. We propose an iterative reweighted method for joint dictionary parameter learning and sparse signal recovery. The proposed method is developed by iteratively decreasing a surrogate function that majorizes the original log-sum objective function. Note that the use of the iterative reweighted scheme for sparse signal recovery is not new and has achieved great success over past few years (e.g., [15]–[19]). Nevertheless, previous works concern only recovery of the sparse signal. The current work, instead, generalizes the iterative reweighted scheme for joint dictionary parameter learning and sparse signal recovery. Moreover, previous iterative reweighted algorithms usually involve iterative minimization of a surrogate function majorizing a given objective function, while we propose to iteratively decrease a surrogate function. This generalization extends the applicability of the iterative reweighted scheme since finding a simple and convex surrogate function which admits an analytical solution could be difficult for many complex problems.

The current work is an extension of our previous work [20] to more general scenarios involving noisy and/or multiple measurement vectors. As will be shown in this paper, this extension is technically non-trivial for the following two reasons. First, the extension to the noisy case inevitably involves a choice of a regularization parameter controlling the tradeoff between the data fitting error and the sparsity of the solution, which is tricky but meanwhile critical to the recovery performance. In the paper, to address this issue, a simple yet effective scheme was developed for adaptively updating the regularization parameter. Second, a pruning operation is introduced to remove those small coefficients along with their associated frequency components during the iterative process, which brings in improved stability and substantial reduction in computational complexity. Note that such a pruning operation is not applicable to [20] because performing the pruning will result in an ill-posed inverse problem (see Section III.B for additional details). In this paper, we also provide a theoretical analysis for the log-sum minimization considered in this paper. Our theoretical analysis shows that the global minimum of the log-sum minimization approaches the true solution as the parameter of the log-sum functional reduces to zero, i.e., $\epsilon \to 0$, which offers a theoretical justification for the proposed algorithm.

Note that estimating the frequencies of a mixture of complex sinusoids is an extensively studied problem in radar, sonar and seismic applications. Classical techniques include subspace-based methods such as the MUSIC [21], ESPRIT [22], matrix-pencil [23]. Some recent works on frequency parameter estimation include, e.g., [24], [25]. Sparse signal representation (compressed sensing) provides a new perspective and methodology to address the line spectral estimation problem. Compressed sensing-based methods present several unique advantages over conventional subspace-based methods, e.g., they are able to achieve super-resolution accuracy with a small number of time samples, and can deal with coherent sources. A more detailed discussion and comparison between compressed-sensing methods and classical subspace-based methods, however, is beyond the scope of this paper as our major objective is to devise a new approach to overcome the grid mismatch limitation inherent in conventional compressed sensing techniques. Also, although our paper focuses on line spectral estimation problem, the proposed method is quite general and may be applied to other parametric dictionary models.

The rest of the paper is organized as follows. In Section II, the line spectral estimation problem is formulated as a joint sparse representation and dictionary parameter estimation problem. An iterative reweighted algorithm is developed in Section III. The choice of the regularization parameter controlling the tradeoff between sparsity and data fitting is discussed in Section IV, where a simple and effective update rule for the regularization parameter is proposed. In Section V, we provide a theoretical analysis of the considered optimization problem for the noiseless case. Extension of the proposed algorithm to the multiple measurement vector scenario is studied in Section VI. Simulation results are provided in Section VII, followed by concluding remarks in Section VIII.

II. Problem Formulation

Consider the line spectral estimation problem where the observed signal is a summation of a number of complex sinusoids:

$$y_m = \sum_{k=1}^{K} \alpha_k e^{-j\omega_k m} + e_m \quad m = 1, \ldots, M \tag{1}$$

where $\omega_k \in [0, 2\pi)$ and α_k denote the frequency and the complex amplitude of the k-th component, respectively, and e_m represents the observation noise. Define $\boldsymbol{a}(\omega) \triangleq [e^{-j\omega}\ e^{-j2\omega}\ \ldots\ e^{-jM\omega}]^T$, the model (1) can be rewritten in a vector-matrix form as

$$\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{\omega})\boldsymbol{\alpha} + \boldsymbol{e} \tag{2}$$

[1]The notion of "super-resolution" was introduced in [6] to refer to the ability of resolving the true parameters with infinite precision. In some other works, "super-resolution compressed sensing" also refers to "off-grid compressed sensing".

where $\boldsymbol{y} \triangleq [y_1 \ \ldots \ y_M]^T$, $\boldsymbol{\alpha} \triangleq [\alpha_1 \ \ldots \ \alpha_K]^T$, and $\boldsymbol{A}(\boldsymbol{\omega}) \triangleq [\boldsymbol{a}(\omega_1) \ \ldots \ \boldsymbol{a}(\omega_K)]$. Note that in some applications, to facilitate data acquisition, we may wish to estimate $\{\omega_k\}$ and $\{\alpha_k\}$ from a subset of measurements randomly extracted from $\{y_m\}_{m=1}^M$. This random sampling operation amounts to constructing a new dictionary by retaining the associated rows and removing other rows of $\boldsymbol{A}(\boldsymbol{\omega})$. This modification, however, makes no difference to our algorithm development.

We first discuss how conventional compressed sensing techniques are applied to address this line spectral estimation problem. For conventional compressed sensing techniques, the continuous frequency parameter space has to be discretized into a finite set of grid points. The unknown frequency components $\{\omega_k\}$ are assumed to lie on some of the discretized grid points. Estimating $\{\omega_k\}$ and $\{\alpha_k\}$ can then be formulated as a sparse signal recovery problem $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{e}$, where $\boldsymbol{A} \in \mathbb{C}^{M \times N} (M \ll N)$ is an overcomplete dictionary constructed based on the discretized grid points. In practical applications, the true parameters do not necessarily lie on the discretized grid, in which case the recovery performance deteriorated.

To circumvent this issue, in this paper, the parameters $\{\theta_n\}$ associate with the overcomplete dictionary $\boldsymbol{A}(\boldsymbol{\theta}) \triangleq [\boldsymbol{a}(\theta_1) \ \ldots \ \boldsymbol{a}(\theta_N)]$, where each atom $\boldsymbol{a}(\theta_n)$ is determined by a frequency parameter θ_n, are treated unknown. Estimating $\{\omega_k\}$ and $\{\alpha_k\}$ can therefore be formulated as a sparse signal recovery problem with an unknown parametric dictionary. In this framework, the objective is not only to estimate the sparse signal, but also to optimize/refine the frequency parameters $\boldsymbol{\theta} \triangleq \{\theta_n\}_{n=1}^N$ such that the parametric dictionary will approach the true sparsifying dictionary. More specifically, the goal is to search for a set of unknown parameters $\{\theta_n\}_{n=1}^N$ with which the observed signal $\boldsymbol{y}$ can be represented by as few atoms as possible with a specified error tolerance. Such a problem can be readily formulated as

$$\begin{aligned} \min_{\boldsymbol{z},\boldsymbol{\theta}} \quad & \| \boldsymbol{z} \|_0 \\ \text{s.t.} \quad & \| \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{z} \|_2 \leq \xi \end{aligned} \tag{3}$$

where $\| \boldsymbol{z} \|_0$ stands for the number of the nonzero components of $\boldsymbol{z}$, and ξ is an error tolerance parameter related to noise statistics. The optimization (3), however, is an NP-hard problem. Alternative sparsity-promoting functions such as ℓ_1-norm can be used to replace ℓ_0-norm to find a sparse solution of $\boldsymbol{z}$ more computationally efficient. In this paper, we consider the use of the log-sum sparsity-encouraging functional. Log-sum penalty function has been used for sparse signal recovery [17], [19] and was shown to be more sparsity-encouraging than the ℓ_1-norm [19], [26]. Replacing the ℓ_0-norm in (3) with the log-sum functional leads to

$$\begin{aligned} \min_{\boldsymbol{z},\boldsymbol{\theta}} \quad & L(\boldsymbol{z}) = \sum_{n=1}^N \log\left(|z_n|^2 + \epsilon\right) \\ \text{s.t.} \quad & \| \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{z} \|_2 \leq \xi \end{aligned} \tag{4}$$

where z_n denotes the nth component of the vector $\boldsymbol{z}$, and $\epsilon > 0$ is a positive parameter to ensure that the function is well-defined. The choice of ϵ will be discussed later in our paper. The optimization (4) can be formulated as an unconstrained optimization problem by removing the constraint and adding a data fitting term, $\lambda \| \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{z} \|_2^2$, to the objective functional, which yields the following optimization

$$\begin{aligned} \min_{\boldsymbol{z},\boldsymbol{\theta}} \quad G(\boldsymbol{z},\boldsymbol{\theta}) &\triangleq \sum_{n=1}^N \log\left(|z_n|^2 + \epsilon\right) + \lambda \| \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{z} \|_2^2 \\ &= L(\boldsymbol{z}) + \lambda \| \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{z} \|_2^2 \end{aligned} \tag{5}$$

where $\lambda > 0$ is a regularization parameter controlling the trade-off between data fitting and the sparsity of the solution, and its choice will be more thoroughly discussed later in this paper. The above optimization (5) can be solved via a two-stage iterative algorithm [27] that alternates between a sparse coding stage and a dictionary update stage. Nevertheless, this two-stage algorithm does not guarantee a monotonically decreasing objective function value. The algorithm is also very likely to be trapped in undesirable local minima. In the following, we develop an iterative reweighted algorithm which shall overcome these limitations.

III. Proposed Iterative Reweighted Algorithm

A. Algorithm Development

We now develop an iterative reweighted algorithm for joint dictionary parameter learning and sparse signal recovery. We resort to a bounded optimization approach, also known as the majorization-minimization (MM) approach [17], [28], to solve the optimization (5). The idea of the MM approach is to iteratively minimize a simple surrogate function majorizing the given objective function. Nevertheless, in this paper we will show that through iteratively decreasing (not necessarily minimizing) the surrogate function, the iterative process also yields a non-increasing objective function value and eventually converges to a stationary point of $G(\boldsymbol{z},\boldsymbol{\theta})$. To obtain an appropriate surrogate function for (5), we first find a suitable surrogate function for the log-sum functional $L(\boldsymbol{z})$. It can be verified that a differentiable and convex surrogate function majorizing $L(\boldsymbol{z})$ is given by

$$Q\left(\boldsymbol{z}|\widehat{\boldsymbol{z}}^{(t)}\right) \triangleq \sum_{n=1}^N \left(\frac{|z_n|^2 + \epsilon}{\left|\hat{z}_n^{(t)}\right|^2 + \epsilon} + \log\left(\left|\hat{z}_n^{(t)}\right|^2 + \epsilon\right) - 1 \right) \tag{6}$$

where $\widehat{\boldsymbol{z}}^{(t)} \triangleq [\hat{z}_1^{(t)} \ \ldots \ \hat{z}_N^{(t)}]^T$ denotes an estimate of $\boldsymbol{z}$ at iteration t. We can easily verify that $Q(\boldsymbol{z}|\widehat{\boldsymbol{z}}^{(t)}) - L(\boldsymbol{z}) \geq 0$, with the equality attained when $\boldsymbol{z} = \widehat{\boldsymbol{z}}^{(t)}$. Consequently the surrogate function for the objective function $G(\boldsymbol{z},\boldsymbol{\theta})$ is

$$S\left(\boldsymbol{z},\boldsymbol{\theta}|\widehat{\boldsymbol{z}}^{(t)}\right) \triangleq Q\left(\boldsymbol{z}|\widehat{\boldsymbol{z}}^{(t)}\right) + \lambda \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{z}\|_2^2 \tag{7}$$

Solving (5) now reduces to minimizing the surrogate function iteratively. Ignoring terms independent of $\{\boldsymbol{z},\boldsymbol{\theta}\}$, optimizing the

surrogate function (7) is simplified as

$$\min_{\boldsymbol{z},\boldsymbol{\theta}} \quad \boldsymbol{z}^H \boldsymbol{D}^{(t)} \boldsymbol{z} + \lambda \| \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta}) \boldsymbol{z} \|_2^2 \tag{8}$$

where $[\cdot]^H$ denotes the conjugate transpose, and $\boldsymbol{D}^{(t)}$ is a diagonal matrix given as

$$\boldsymbol{D}^{(t)} \triangleq \text{diag}\left\{ \frac{1}{\left|\hat{z}_1^{(t)}\right|^2 + \epsilon}, \ldots, \frac{1}{\left|\hat{z}_N^{(t)}\right|^2 + \epsilon} \right\}$$

Conditioned on $\boldsymbol{\theta}$, the optimal $\boldsymbol{z}$ of (8) can be readily obtained as

$$\boldsymbol{z}^*(\boldsymbol{\theta}) = \left(\boldsymbol{A}^H(\boldsymbol{\theta}) \boldsymbol{A}(\boldsymbol{\theta}) + \lambda^{-1} \boldsymbol{D}^{(t)} \right)^{-1} \boldsymbol{A}^H(\boldsymbol{\theta}) \boldsymbol{y} \tag{9}$$

Substituting (9) back into (8), the optimization simply becomes searching for the unknown parameter $\boldsymbol{\theta}$:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \ f(\boldsymbol{\theta}) \triangleq & -\boldsymbol{y}^H \boldsymbol{A}(\boldsymbol{\theta}) \left(\boldsymbol{A}^H(\boldsymbol{\theta}) \boldsymbol{A}(\boldsymbol{\theta}) + \lambda^{-1} \boldsymbol{D}^{(t)} \right)^{-1} \\ & \times \boldsymbol{A}^H(\boldsymbol{\theta}) \boldsymbol{y} \end{aligned} \tag{10}$$

An analytical solution of the above optimization (10) is difficult to obtain. Nevertheless, in our algorithm, we only need to search for a new estimate $\widehat{\boldsymbol{\theta}}^{(t+1)}$ such that the following inequality holds

$$f\left(\widehat{\boldsymbol{\theta}}^{(t+1)}\right) \leq f\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) \tag{11}$$

Since $f(\boldsymbol{\theta})$ is differentiable for our case, such an estimate can be easily obtained by using a gradient descent method. Given $\widehat{\boldsymbol{\theta}}^{(t+1)}$, $\widehat{\boldsymbol{z}}^{(t+1)}$ can be obtained via (9), with $\boldsymbol{\theta}$ replaced by $\widehat{\boldsymbol{\theta}}^{(t+1)}$, i.e.,

$$\widehat{\boldsymbol{z}}^{(t+1)} = \boldsymbol{z}^*\left(\widehat{\boldsymbol{\theta}}^{(t+1)}\right) \tag{12}$$

In the following, we show that the new estimate $\{\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)}\}$ results in a non-increasing objective function value, that is,

$$G\left(\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)}\right) \leq G\left(\widehat{\boldsymbol{z}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}\right) \tag{13}$$

To this goal, we first show the following inequality

$$\begin{aligned} S\left(\widehat{\boldsymbol{z}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) & \overset{(a)}{\geq} S\left(\boldsymbol{z}^*\left(\widehat{\boldsymbol{\theta}}^{(t)}\right), \widehat{\boldsymbol{\theta}}^{(t)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \\ & = f\left(\widehat{\boldsymbol{\theta}}^{(t)}\right) + \text{constant} \\ & \overset{(b)}{\geq} f\left(\widehat{\boldsymbol{\theta}}^{(t+1)}\right) + \text{constant} \\ & = S\left(\boldsymbol{z}^*\left(\widehat{\boldsymbol{\theta}}^{(t+1)}\right), \widehat{\boldsymbol{\theta}}^{(t+1)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \\ & \overset{(c)}{=} S\left(\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \end{aligned} \tag{14}$$

where (a) comes from the fact that $\boldsymbol{z}^*(\boldsymbol{\theta})$ is the optimal solution to the optimization (8); (b) and (c) follow from (11) and (12), respectively. Moreover, we have

$$\begin{aligned} & G\left(\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)}\right) - S\left(\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \\ & = L\left(\widehat{\boldsymbol{z}}^{(t+1)}\right) - Q\left(\widehat{\boldsymbol{z}}^{(t+1)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \\ & \overset{(a)}{\leq} L\left(\widehat{\boldsymbol{z}}^{(t)}\right) - Q\left(\widehat{\boldsymbol{z}}^{(t)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \\ & = G\left(\widehat{\boldsymbol{z}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}\right) - S\left(\widehat{\boldsymbol{z}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \end{aligned} \tag{15}$$

where (a) follows from the fact that $Q(\boldsymbol{z}|\widehat{\boldsymbol{z}}^{(t)}) - L(\boldsymbol{z})$ attains its minimum when $\boldsymbol{z} = \widehat{\boldsymbol{z}}^{(t)}$. Combining (14)–(15), we eventually arrive at

$$\begin{aligned} G\left(\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)}\right) & = G\left(\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)}\right) - S\left(\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \\ & \quad + S\left(\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \\ & \leq G\left(\widehat{\boldsymbol{z}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}\right) - S\left(\widehat{\boldsymbol{z}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \\ & \quad + S\left(\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)} \middle| \widehat{\boldsymbol{z}}^{(t)}\right) \\ & \leq G\left(\widehat{\boldsymbol{z}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}\right) \end{aligned} \tag{16}$$

We see that through iteratively decreasing (not necessarily minimizing) the surrogate function, the objective function $G(\boldsymbol{z}, \boldsymbol{\theta})$ is guaranteed to be non-increasing at each iteration and the iterative process eventually converges to a stationary point of $G(\boldsymbol{z}, \boldsymbol{\theta})$. To better evaluate the proposed algorithm, it is also meaningful to examine the convergence rate of our proposed method. Several works [29], [30] reported results on the convergence rate of iterative reweighted methods. These theoretical results, however, are inapplicable for our proposed algorithm because our proposed method is different from conventional iterative reweighted methods that consider only the recovery of sparse signals. Hence it still remains an open issue to examine the convergence property of the proposed method.

For clarification, we summarize our algorithm as follows.

Iterative Reweighted Algorithm I

1. Given an initialization $\widehat{\boldsymbol{z}}^{(0)}, \widehat{\boldsymbol{\theta}}^{(0)}$, and a pre-selected regularization parameter λ.
2. At iteration $t = 0, 1, \ldots$: Based on the estimate $\widehat{\boldsymbol{z}}^{(t)}$, construct the surrogate function as depicted in (7). Search for a new estimate of the unknown parameter vector, denoted as $\widehat{\boldsymbol{\theta}}^{(t+1)}$, by using the gradient descent method such that the inequality (11) is satisfied. Compute a new estimate of the sparse signal, denoted as $\widehat{\boldsymbol{z}}^{(t+1)}$, via (12).
3. Go to Step 2 if $\| \widehat{\boldsymbol{z}}^{(t+1)} - \widehat{\boldsymbol{z}}^{(t)} \|_2 > \varepsilon$, where ε is a prescribed tolerance value; otherwise stop.

B. Discussions

We see that in our algorithm, the unknown parameters and the signal are refined in a gradual and interweaved manner. This gradual and interweaved refinement allows different atoms in the dictionary, meanwhile gradually refined, fully competing against each other to represent the observed data. Thus the true frequency components are more likely to stand out, and as a result, the algorithm is less likely to get stuck in undesirable local minima. In addition, similar to [18], the parameter ϵ used throughout our optimization can be gradually decreased instead of remaining fixed. For example, at the beginning, ϵ can be set to a relatively large value, say 1. We then gradually reduce the value of ϵ in the subsequent iterations until ϵ attains a sufficiently small value, e.g., 10^{-8}. Numerical results demonstrate that this gradual refinement of the parameter ϵ can further improve the probability of finding the correct solution.

The second step of the proposed algorithm involves searching for a new estimate of the unknown parameter vector to meet the condition (11). As mentioned earlier, this can be accomplished via a gradient-based search algorithm. Details of computing the gradient of $f(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ are provided in Appendix A. Also, to achieve a better reconstruction accuracy, the estimates of $\{\theta_i\}$ can be refined in a sequential manner. Our experiments suggest that a new estimate which satisfies (11) can be easily obtained within only a few iterations.

The main computational task of our proposed algorithm at each iteration is to calculate $\boldsymbol{z}^*(\boldsymbol{\theta})$ (as per (9)) and the first derivative of $f(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, both of which involve computing the inverse of the following $N \times N$ matrix: $\boldsymbol{A}^H(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta}) + \lambda^{-1}\boldsymbol{D}^{(t)}$. The computational complexity of the proposed method can be reduced by introducing a pruning operation, that is, at each iteration, we remove those small coefficients along with their associated frequency components. Thus the dimension of the signal $\boldsymbol{z}$ and the parameter $\boldsymbol{\theta}$ keeps shrinking as the iterative process evolves. As a consequence, the dimension of the matrix $\boldsymbol{A}^H(\boldsymbol{\theta})\boldsymbol{A}(\boldsymbol{\theta}) + \lambda^{-1}\boldsymbol{D}^{(t)}$ to be inverted decreases accordingly, and hence the computational complexity is reduced. A hard thresholding rule can be used to remove those irrelevant frequency components. Specifically, if the coefficient $\hat{z}_n^{(t)}$ is less than a pre-specified small value τ, i.e., $\hat{z}_n^{(t)} \leq \tau$, then the associated frequency component $\hat{\theta}_n^{(t)}$ can be removed from further consideration since its contribution to the signal synthesis is negligible.

Note that the above pruning procedure cannot be applied to our previous algorithm [20] developed for the scenario of noiseless measurements. To see this, the previous algorithm requires computing the inverse of the following $M \times M$ matrix $\boldsymbol{A}(\boldsymbol{\theta})(\boldsymbol{D}^{(t)})^{-1}\boldsymbol{A}^H(\boldsymbol{\theta})$ at each iteration. Performing pruning operations will result in an ill-posed inverse problem since the above matrix will eventually become singular as the dimension of $\boldsymbol{A}(\boldsymbol{\theta})$ shrinks. The proposed method in the current work is therefore computationally more attractive than our previous algorithm, particularly when the number of observed data samples, M, is large.

Our proposed method requires a finite set of grid points as initial estimates of the dictionary parameters. Nevertheless, unlike conventional compressed sensing methods, the dictionary parameters are jointly optimized/refined along with the sparse signal during the iterative process. For our method, a general guideline for choosing N is to let $N \gg K$ in order to obtain a sufficiently fine grid, because a finer initial grid makes the search for the dictionary parameters easier, faster, and less likely to be trapped in undesirable local minima. With a fine initial grid, some of the grid points are close to the true frequency components. The estimated coefficients associated with these grid points are expected to be prominent during the first few iterations, thus serving as a reliable estimate for subsequent refinement/optimization. On the contrary, a coarse grid, for example, considering the extreme case $N = K$, would make the search for the dictionary parameters more difficult and vulnerable to convergence to local minima since the initial grid points are far away from the true frequency components. Also note that N does not necessarily need to be greater than the number of measurements M. The proposed algorithm works well as long as $N \gg K$.

IV. Adaptive Update of λ

As mentioned earlier, λ is a regularization parameter controlling the tradeoff between the sparsity of the solution and the data fitting error. Clearly, a small λ leads to a sparse solution, whereas a larger λ renders a less sparse but better-fitting solution. As a consequence, in scenarios where frequency components are closely-spaced, choosing a small λ may result in an underestimation of the frequency components while an excessively large λ may lead to an overestimation. Thus the choice of λ is critical to the recovery performance.

When the knowledge of the noise level is known *a priori*, the regularization parameter λ can be chosen such that the norm of the residual matches the noise level of the data. This selection rule is also known as the discrepancy principle [31]. For the case of unknown noise variance, the L-curve method has been shown to provide a reasonably good and robust parameter choice [31] in some experiments. Nevertheless, the L-curve method is computationally expensive for our case since, in order to plot the L-curve, it requires us to solve the optimization problem (5) for a number of different values of λ. To the best of our knowledge, a general rule for regularization parameter selection remains an open issue. In this section, we propose a simple yet effective scheme for adaptively updating the parameter λ during the iterative process. The developed scheme does not require the knowledge of the noise variance.

Note that iterative reweighted methods have a close connection with sparse Bayesian learning algorithms [32]–[34]. In fact, a dual-form analysis [19] reveals that sparse Bayesian learning can be considered as a non-separable iterative reweighted strategy solving a non-separable penalty function. Inspired by this insight, it is expected the mechanism inherent in the sparse Bayesian learning method to achieve automatic balance between the sparsity and the fitting error should also work for the iterative reweighted methods. Let us first briefly examine how the sparse Bayesian learning algorithm works. In the sparse Bayesian learning framework, the observation noise is

assumed to be white Gaussian noise with zero mean and variance $\delta \triangleq \sigma^2$, and the sparse signal $\boldsymbol{z}$ is assigned a Gaussian prior distribution [32]

$$p(\boldsymbol{z}|\boldsymbol{\alpha}) = \prod_{n=1}^{N} p(z_n|\alpha_n)$$

where $p(z_n|\alpha_n) = \mathcal{N}(z_n|0, \alpha_n^{-1})$ and $\boldsymbol{\alpha} \triangleq \{\alpha_n\}$. Here each α_n is the inverse variance (precision) of the Gaussian distribution, and a non-negative sparsity-controlling hyperparameter. For each iteration, given a set of estimated hyperparameters $\{\alpha_n^{(t)}\}$, the maximum a posterior (MAP) estimator of $\boldsymbol{z}$ can be obtained via

$$\widehat{\boldsymbol{z}}^{(t)} = \arg\min_{\boldsymbol{z}} \quad \boldsymbol{z}^H \boldsymbol{D}^{(t)} \boldsymbol{z} + \delta^{-1} \|\boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{z}\|_2^2 \tag{17}$$

where $\boldsymbol{D}^{(t)} \triangleq \text{diag}(\alpha_1^{(t)}, \ldots, \alpha_N^{(t)})$. Meanwhile, given the estimated sparse signal $\widehat{\boldsymbol{z}}^{(t)}$ and its posterior covariance matrix, the hyperparameters $\{\alpha_i\}$ are re-estimated. In this Bayesian framework, the tradeoff between the sparsity and the data fitting is automatically achieved by employing a probabilistic model for the sparse signal $\boldsymbol{z}$, and the tradeoff tuning parameter is equal to the inverse of the noise variance δ (cf. (17)).

Comparing (8) and (17), we see that the sparse Bayesian learning method and our proposed iterative reweighted algorithm have similar formulations in updating the spares signal $\boldsymbol{z}$, except that the diagonal matrix $\boldsymbol{D}^{(t)}$ is calculated in different ways. The formulation of (17) sheds light on the choice of λ in (8). Following (17), we can set λ inversely proportional to the noise variance, i.e., $\lambda = d\delta^{-1}$, where d is a constant scaling factor. Note that when the noise variance is unknown *a priori*, the noise variance δ can be iteratively estimated, based on which the tuning parameter λ can be iteratively updated. A reasonable estimate of the noise variance is given by

$$\hat{\delta}^{(t)} = \frac{\left\|\boldsymbol{y} - \boldsymbol{A}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right)\widehat{\boldsymbol{z}}^{(t)}\right\|_2^2}{M} \tag{18}$$

and accordingly $\lambda^{(t)}$ can be updated as

$$\lambda^{(t)} = \frac{d}{\hat{\delta}^{(t)}} = \frac{dM}{\left\|\boldsymbol{y} - \boldsymbol{A}\left(\widehat{\boldsymbol{\theta}}^{(t)}\right)\widehat{\boldsymbol{z}}^{(t)}\right\|_2^2} \tag{19}$$

As mentioned earlier, the choice of λ is always a tricky issue. In order to achieve a reasonable balance between data fitting and sparsity, we may need to tune λ for different numbers of samples, different signal-to-noise ratios, and different frequency spacings. The proposed adaptive update scheme helps circumvent this issue. Although another parameter d was introduced, its choice is much easier: the value of d can be fixed irrespective of other factors such as signal-to-noise ratios and frequency spacings. In other words, once d is chosen, no further tuning of d is needed. The iterative update of λ can be seamlessly integrated into our algorithm, which is summarized as follows.

The above algorithm, in fact, can be interpreted as an alternating procedure for solving the following optimization problem

Iterative Reweighted Algorithm II

1. Given an initialization $\widehat{\boldsymbol{z}}^{(0)}, \widehat{\boldsymbol{\theta}}^{(0)}$, and $\lambda^{(0)}$.
2. At iteration $t = 0, 1, \ldots$: Based on $\widehat{\boldsymbol{z}}^{(t)}$ and $\lambda^{(t)}$, construct the surrogate function as depicted in (7). Search for a new estimate of the unknown parameter vector, denoted as $\widehat{\boldsymbol{\theta}}^{(t+1)}$, by using the gradient descent method such that the inequality (11) is satisfied. Compute a new estimate of the sparse signal, denoted as $\widehat{\boldsymbol{z}}^{(t+1)}$, via (12). Compute a new regularization parameter $\lambda^{(t+1)}$ according to (19).
3. Go to Step 2 if $\| \widehat{\boldsymbol{z}}^{(t+1)} - \widehat{\boldsymbol{z}}^{(t)} \|_2 > \varepsilon$, where ε is a prescribed tolerance value; otherwise stop.

which involves optimization of $\boldsymbol{z}$, $\boldsymbol{\theta}$, and λ:

$$\begin{aligned}\min_{\boldsymbol{z},\boldsymbol{\theta},\lambda} \tilde{G}(\boldsymbol{z},\boldsymbol{\theta},\lambda) &\triangleq \sum_{n=1}^{N} \log\left(|z_n|^2 + \epsilon\right) + \lambda \parallel \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{z} \parallel_2^2 \\ &\quad - dM \log\lambda \\ &= L(\boldsymbol{z}) + \lambda \parallel \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{z} \parallel_2^2 - dM\log\lambda\end{aligned} \tag{20}$$

To see this, note that given an estimate of $\{\widehat{\boldsymbol{z}}^{(t)}, \widehat{\boldsymbol{\theta}}^{(t)}\}$, the optimal λ of (20) is given by (19). On the other hand, for a fixed $\lambda^{(t)}$, the above optimization reduces to (5), in which case a new estimate $\{\widehat{\boldsymbol{z}}^{(t+1)}, \widehat{\boldsymbol{\theta}}^{(t+1)}\}$ can be obtained according to (11) and (12). Therefore the proposed algorithm ensures that the objective function of (20) keeps non-increasing at each iteration, and the proposed algorithm eventually converges to a stationary point of (20). The last term $dM \log\lambda$ in (20) is a regularization term included to pull λ away from zero. Without this term, the optimization (20) becomes meaningless since the optimal λ in this case equals to zero.

V. Theoretical Analysis

In this section, we consider the noiseless case and provide a theoretical analysis to show that the global minimum of the optimization (4) approaches the true solution as $\epsilon \to 0$. For the noiseless case, the optimization (4) simply becomes

$$\begin{aligned}\min_{\boldsymbol{z},\boldsymbol{\theta}} \quad & L(\boldsymbol{z}) = \sum_{i=1}^{N} \log\left(|z_i|^2 + \epsilon\right) \\ \text{s.t.} \quad & \boldsymbol{y} = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{z}\end{aligned} \tag{21}$$

Let $\overline{\boldsymbol{\theta}}_0 \triangleq [\omega_1 \ \ldots \ \omega_K]$ and $\overline{\boldsymbol{z}}_0 \triangleq [\alpha_1 \ \ldots \ \alpha_K]$ denote the true frequencies and complex amplitudes, respectively; $\boldsymbol{\theta}_0$ and $\boldsymbol{z}_0$ denote the associated augmented N-dimensional vectors. Also, define $S_\omega \triangleq \{\omega_i\}_{i=1}^K$. Let $\{\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{z}}\}$ denote a *feasible solution* of (21). Clearly, we have

$$\boldsymbol{A}(\overline{\boldsymbol{\theta}}_0)\overline{\boldsymbol{z}}_0 = \boldsymbol{A}\left(\widehat{\boldsymbol{\theta}}\right)\widehat{\boldsymbol{z}} \tag{22}$$

where $\widehat{\boldsymbol{z}}$ is an N-dimensional vector. We assume $\widehat{\boldsymbol{z}}$ has $J \leq N$ nonzero coefficients, which are denoted as $\{\hat{z}_i\}_{i=1}^J$. The

associated estimated frequencies are denoted by $\{\hat{\theta}_i\}_{i=1}^{J}$. Without loss of generality, we partition $\{\hat{\theta}_i\}_{i=1}^{J}$ into two subsets: $S_1 = \{\hat{\theta}_i\}_{i=1}^{p}$ and $S_2 = \{\hat{\theta}_i\}_{i=p+1}^{J}$, where $S_1 \cap S_\omega = \varnothing$ and $S_2 \subseteq S_\omega$.

If $S_2 = S_\omega$ and $\{\hat{z}_i\}_{i=p+1}^{J} = \{\alpha_i\}_{i=1}^{K}$, we clearly have $L(\widehat{\boldsymbol{z}}) \geq L(\boldsymbol{z}_0)$, and the inequality becomes equality when $\{\hat{z}_i\}_{i=1}^{p}$ are equal to zero, which implies that the solution $\{\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{z}}\}$ is equivalent to the groundtruth. Hence we only need to consider the general case where $S_2 \subset S_\omega$, or $S_2 = S_\omega$ but $\{\hat{z}_i\}_{i=p+1}^{J} \neq \{\alpha_i\}_{i=1}^{K}$. For this general case, it is easy to know that

$$p \geq M - K + 1 \tag{23}$$

otherwise there is no way to satisfy (22). Without loss of generality, we assume

$$|\hat{z}_1| \geq |\hat{z}_2| \geq \cdots \geq |\hat{z}_p| \tag{24}$$

Also, since $S_1 \cap S_\omega = \varnothing$, we can assume

$$|\hat{\theta}_i - \omega_j| > \nu \quad \forall i \in \{1, \ldots, p\}, j \in \{1, \ldots, K\} \tag{25}$$

where ν is a very small value, say 10^{-6}, to ensure that $\hat{\theta}_i$ and ω_j are differentiable, otherwise $\hat{\theta}_i$ can be considered essentially equivalent to ω_j. Similarly, we assume

$$|\hat{\theta}_i - \hat{\theta}_j| > \nu \quad \forall i, j \in \{1, \ldots, p\} \tag{26}$$

$$|\omega_i - \omega_j| > \nu \quad \forall i, j \in \{1, \ldots, K\} \tag{27}$$

otherwise the two frequencies, say $\hat{\theta}_i$ and $\hat{\theta}_j$, can be merged. With the above assumptions, we have

$$\boldsymbol{A}\left(\widehat{\boldsymbol{\theta}}\right)\widehat{\boldsymbol{z}} - \boldsymbol{A}\left(\overline{\boldsymbol{\theta}}_0\right)\overline{\boldsymbol{z}}_0 = \sum_{i=1}^{p} \hat{z}_i \boldsymbol{a}\left(\hat{\theta}_i\right) + \sum_{i=p+1}^{J} \hat{z}_i \boldsymbol{a}\left(\hat{\theta}_i\right)$$

$$- \sum_{i=1}^{K} \alpha_i \boldsymbol{a}\left(\omega_i\right) = \sum_{i=1}^{p} \hat{z}_i \boldsymbol{a}\left(\hat{\theta}_i\right) + \sum_{i=1}^{K} \eta_i \boldsymbol{a}\left(\omega_i\right) \tag{28}$$

in which

$$\eta_i \triangleq \begin{cases} \hat{z}_{p+i} - \alpha_i & p + i \leq J \\ -\alpha_i & \text{otherwise} \end{cases} \tag{29}$$

Define $\eta_{\max} \triangleq \max_i |\eta_i|$ and assume

$$\eta_{\max} > \tau \tag{30}$$

where τ is a very small value, say 10^{-6}, to ensure that $\{\hat{z}_i\}_{i=p+1}^{J}$ and $\{\alpha_i\}$ are differentiable, otherwise $\{\hat{z}_i\}_{i=p+1}^{J}$ can be considered essentially equivalent to $\{\alpha_i\}$.

In the following, we will show that we can always find a sufficiently small ϵ such that the inequality $L(\widehat{\boldsymbol{z}}) > L(\boldsymbol{z}_0)$ holds. The result is summarized as follows.

Theorem 1: Let $\{\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{z}}\}$ denote a *feasible solution* of (21) and satisfy the assumptions made from (22) to (28). Also, let $\overline{\boldsymbol{\theta}}_0 \triangleq [\omega_1 \ \ldots \ \omega_K]$ and $\overline{\boldsymbol{z}}_0 \triangleq [\alpha_1 \ \ldots \ \alpha_K]$ denote the true frequencies and complex amplitudes, respectively; $\boldsymbol{\theta}_0$ and $\boldsymbol{z}_0$ denote the associated augmented N-dimensional vectors. Define $\alpha_{\max} \triangleq \max_i |\alpha_i|$, and assume $M \geq 2K$. If

$$\epsilon < \min\left\{ \left(\frac{C^{M-K+1}}{2^K \alpha_{\max}^{2K}} \right)^{\frac{1}{M-2K+1}}, \alpha_{\max}^2 \right\} \tag{31}$$

then we have $L(\widehat{\boldsymbol{z}}) > L(\boldsymbol{z}_0)$, where

$$C \triangleq \frac{1}{M(N-M+K)} C_1 \tau^2 \tag{32}$$

$$C_1 \triangleq \frac{(M-1)^{(M-1)/2}}{M^{(M-1)}} \left| 2 \sin \frac{\nu}{2} \right|^{M(M-1)/2} \tag{33}$$

where ν and τ are defined in (25) and (30), respectively.

Proof: See Appendix B. ■

Theorem 1 suggests that for any feasible solution $\{\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{z}}\} \neq \{\boldsymbol{\theta}_0, \boldsymbol{z}_0\}$, we can always find a sufficiently small ϵ such that $L(\widehat{\boldsymbol{z}}) > L(\boldsymbol{z}_0)$. In other words, by setting $\epsilon \to 0$, we can ensure that the global minimum of (21) approaches the true solution. Note that the frequency minimum separation condition required by atomic-norm minimization methods (e.g., [6], [7], [9]) to ensure exact recovery is no longer needed in our analysis. This suggests that the proposed algorithm has the potential to resolve more closely-spaced frequency components.

When noise is present, the recovery performance will certainly degrade. A theoretical analysis of the recovery accuracy of the proposed method in the noisy case, however, seems very challenging, and will be a topic of our future investigation. Nevertheless, our experiments suggest that the frequency estimates are less sensitive to the observation noise. Even with a moderate signal-to-noise ratio, our proposed method can still super-resolve the frequency components with a decent success rate. We note that some theoretical analyses for the noisy case are now available for the atomic-norm minimization method, and it was shown [9] that the frequency estimation error is bounded and tends to zero as the number of samples grows.

VI. Extension to the MMV Model

In some practical applications such as EEG/MEG source localization and DOA estimation, multiple measurements $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L\}$ of a time series process may be available. Also, is was shown [35] that for the line spectral estimation problem, exploitation of the covariance matrix of the received signal can help remove the minimum frequency separation requirement for exact recovery. This motivates us to consider the super-resolution compressed sensing problem in a multiple measurement vector (MMV) framework [36]

$$\boldsymbol{Y} = \boldsymbol{A}\left(\boldsymbol{\theta}\right)\boldsymbol{Z} + \boldsymbol{E} \tag{34}$$

where $\boldsymbol{Y} \triangleq [\boldsymbol{y}_1 \ \boldsymbol{y}_2 \cdots \boldsymbol{y}_L]$ is an observation matrix consisting of L observed vectors, $\boldsymbol{Z} \triangleq [\boldsymbol{z}_1 \ \boldsymbol{z}_2 \cdots \boldsymbol{z}_L]$ is a sparse matrix with each row representing a possible source or frequency component, and $\boldsymbol{E}$ denotes the noise matrix. Note that in the MMV model, we assume that the support of the sparse signal remains unchanged over time, that is, the matrix $\boldsymbol{Z}$ has a common row sparsity pattern. This is a reasonable assumption in many applications where the variations of locations or frequencies are slow compared to the sampling rate. The problem of joint dictionary

parameter learning and sparse signal recovery can be formulated as follows

$$\min_{\boldsymbol{Z},\boldsymbol{\theta}} \quad \| \boldsymbol{u} \|_0 \qquad \text{s.t.} \quad \| \boldsymbol{Y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{Z} \|_F \leq \xi \tag{35}$$

where $\| \boldsymbol{X} \|_F$ denotes the Frobenius norm of the matrix $\boldsymbol{X}$, and $\boldsymbol{u}$ is a column vector with its entry u_n defined as

$$u_n \triangleq \| \boldsymbol{z}_{n\cdot} \|_2 \quad \forall n = 1, \ldots, N$$

in which $\boldsymbol{z}_{n\cdot}$ represents the nth row of $\boldsymbol{Z}$. Thus $\| \boldsymbol{u} \|_0$ equals to the number of nonzero rows in $\boldsymbol{Z}$. Clearly, the optimization (35) aims to search for a set of unknown parameters $\{\theta_n\}$ with which the observed matrix $\boldsymbol{Y}$ can be represented by as few atoms as possible with a specified error tolerance. Again, to make the problem (35) tractable, the ℓ_0-norm can be replaced with the log-sum functional, which leads to the following optimization

$$\min_{\boldsymbol{Z},\boldsymbol{\theta}} \quad L(\boldsymbol{Z}) = \sum_{n=1}^{N} \log\left(\| \boldsymbol{z}_{n\cdot} \|_2^2 + \epsilon \right) \qquad \text{s.t.} \quad \|\boldsymbol{Y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{Z}\|_F \leq \xi \tag{36}$$

The above optimization can be reformulated as

$$\min_{\boldsymbol{Z},\boldsymbol{\theta}} \quad G(\boldsymbol{Z},\boldsymbol{\theta}) \triangleq L(\boldsymbol{Z}) + \lambda \|\boldsymbol{Y} - \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{Z}\|_F^2 \tag{37}$$

Again, we can resort to the majorization-minimization (MM) approach to solve (37). Details of the derivations are omitted here since they are similar to the development for the single measurement vector case.

VII. Simulation Results

We now carry out experiments to illustrate the performance of our proposed super-resolution iterative reweighted algorithm (referred to as SURE-IR)[2]. In our simulations, the initial value of λ and the pruning threshold τ are set equal to $\lambda^{(0)} = 100$ and $\tau = 0.05$, respectively. Also, to improve the stability of our proposed algorithm, the initial value of λ is kept unchanged and the frequency components are unpruned during the first few iterations. The parameter d used in (19) to update λ is set to $d = 0.2$. We compare our proposed algorithm with other existing state-of-the-art super-resolution compressed sensing methods, namely, the sparse Bayesian learning with dictionary refinement algorithm (denoted as DicRefCS) [12], the sparse Bayesian learning with dictionary estimation (denoted as SBL-DE) [13], the atomic norm minimization via the semi-definite programming (SDP) approach [8], and the off-grid sparse Bayesian inference (OGSBI) algorithm [2]. The work [3] is not included for comparison because the method [3] requires to place a non-negativity constraint on the sparse signal. This assumption is valid for covariance-based estimation problems but inconsistent with the setup in our simulations. Among these methods, the SURE-IR, DicRefCS, SBL-DE, and the OGSBI methods require to prespecify the initial grid points. In our experiments, the initial grid points are set to be $\boldsymbol{\theta}^{(0)} = (2\pi/N)[0 \ \ldots \ N-1]^T$, where we choose $N = 64$ for the SURE-IR, the DicRefCS and the SBL-DE methods. While for the OGSBI method, a much finer grid ($N = 200$) is used to improve the Taylor approximation accuracy and the recovery performance.

In our experiments, the signal $\boldsymbol{y}_T \triangleq [y_1 \ \ldots \ y_T]^T$ is a mixture of K complex sinusoids corrupted by independent and identically distributed (i.i.d.) zero-mean Gaussian noise, i.e.,

$$y_l = \sum_{k=1}^{K} \alpha_k e^{-j\omega_k l} + e_l \qquad l = 1, \ldots, T$$

where the frequencies $\{\omega_k\}$ are uniformly generated over $[0, 2\pi)$ and the amplitudes $\{\alpha_k\}$ are uniformly distributed on the unit circle. The measurements $\boldsymbol{y}$ are obtained by randomly selecting M entries from T elements of $\boldsymbol{y}_T$. The observation quality is measured by the peak-signal-to-noise ratio (PSNR) which is defined as PSNR $\triangleq 10\log_{10}(1/\sigma^2)$, where σ^2 denotes the noise variance.

We introduce two metrics to evaluate the recovery performance of respective algorithms, namely, the reconstruction signal-to-noise ratio (RSNR) and the success rate. The RSNR measures the accuracy of reconstructing the original signal $\boldsymbol{y}_T$ from the partial observations $\boldsymbol{y}$, and is defined as

$$\text{RSNR} = 20\log_{10}\left(\frac{\| \boldsymbol{y}_T \|_2}{\| \boldsymbol{y}_T - \widehat{\boldsymbol{y}}_T \|_2} \right)$$

The other metric evaluates the success rate of exactly resolving the K frequency components $\{\omega_k\}$. The success rate is computed as the ratio of the number of successful trials to the total number of independent runs, where $\{\alpha_k\}, \{\omega_k\}$ and the sampling indices (used to obtain $\boldsymbol{y}$) are randomly generated for each run. A trial is considered successful if the number of frequency components is estimated correctly and the estimation error between the estimated frequencies $\{\hat{\omega}_k\}$ and the true parameters $\{\omega_k\}$ is smaller than 10^{-3}, i.e., $\frac{1}{2\pi}\| \boldsymbol{\omega} - \widehat{\boldsymbol{\omega}} \|_2 \leq 10^{-3}$. Note that the SDP methods requires the knowledge of the number of complex sinusoids, K, which is assumed perfectly known to it. The OGSBI method may result in an overestimated solution which may contain multiple peaks around each true frequency component. To compute the success rate for the OGSBI, we only keep those K frequency components associated with the first K largest coefficients.

In the following, we examine the behavior of respective algorithms under different scenarios. In Fig. 1, we plot the average RSNRs and success rates of respective algorithms as a function of the number of measurements M, where we set $T = 64, K = 3$, PSNR $= 25$ dB. Results are averaged over 10^3 independent runs, with $\{\alpha_k\}, \{\omega_k\}$ and the sampling indices (used to obtain $\boldsymbol{y}$ from $\boldsymbol{y}_T$) randomly generated for each run. We see that the proposed method is superior to all other four methods in terms of both the RSNR and success rate. In particular, it is worth mentioning that the proposed method outperforms the SDP method which is guaranteed to find the global solution. This is probably because the log-sum penalty functional adopted by our algorithm is more sparse-encouraging than the atomic norm that is considered as the continuous analog to the ℓ_1 norm for discrete signals. We also see that the OGSBI method,

[2]Codes are available at http://www.junfang-uestc.net/codes/Sure-IR.rar.

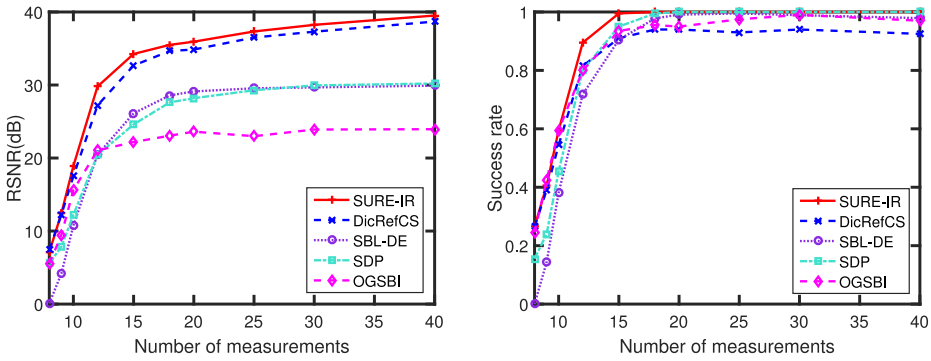

Fig. 1. RSNRs and success rates of respective algorithms vs. $M, T = 64, K = 3$, and PSNR $= 25$ dB. (a) RSNRs vs. M. (b) Success rates vs. M.

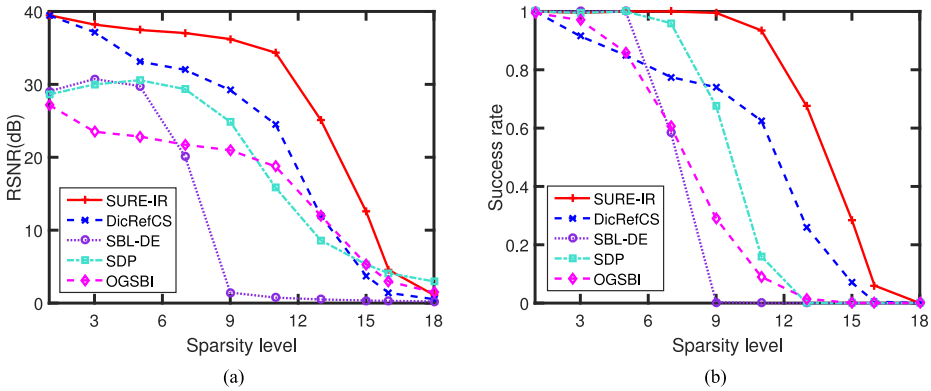

Fig. 2. RSNRs and success rates of respective algorithms vs. $K, T = 64, M = 30$, and PSNR $= 25$ dB. (a) RSNRs vs. K. (b) Success rates vs. K.

though using a very fine grid, still achieves performance inferior to our proposed SURE-IR method. In Fig. 2, we depict the RSNRs and success rates of respective algorithms vs. the number of complex sinusoids, K, where we set $T = 64, M = 30$, and PSNR $= 25$ dB. It can be observed that our proposed SURE-IR algorithm outperforms other methods by a big margin for a moderately large number of complex sinusoids K. For example, when $K = 10$, a gain of about 10 dB in RSNR can be achieved by our algorithm as compared with the DicRefCS and the SDP methods. This advantage makes our algorithm the most attractive for scenarios consisting of a moderate or large number of sinusoid components. On the other hand, we also noticed that the SBL-DE method incurs a considerable performance performance loss as K increases. Although the rationale behind [12] and [13] are similar, we see that the behaviors of these two algorithms are quite different, possibly because they have different inference schemes for updating the dictionary.

To better illustrate the performance, we plot the phase transition curve for each algorithm. Set $T = 64$ and PSNR $= 25$ dB. We vary the sparsity level $K = 3 : 3 : 51$ and the number of measurements $M = 3 : 3 : 51$. For each point (M, K), we conduct 100 independent trials and compute the success rate of exactly resolving the K frequency components, with frequencies randomly generated (the minimum frequency separation is ensured to be greater than $2\pi/N$) and amplitudes uniformly distributed on the unit circle for each trial. The definition of a successful trial is the same as described earlier. In the phase transition plot, the grey value of each point represents the success rate, with white corresponding to perfect recovery while black corresponding to complete failure. We can see from Fig. 3 that our proposed algorithm has a sharper transition boundary than other methods. We also notice that our proposed method provides a cleaner area below the transition boundary as compared with other competing methods, which implies higher success rates are achieved for points in the area.

We now examine the ability of respective algorithms in resolving closely-spaced frequency components. The signal $\boldsymbol{y}$ is assumed a mixture of K complex sinusoids with the frequency spacing $d_f \triangleq \frac{1}{2\pi}(\omega_1 - \omega_2) = \mu/T$, where μ is the frequency spacing coefficient ranging from 0.4 to 2. Fig. 4 shows success rates of respective algorithms vs. the frequency spacing coefficient μ, where we set $T = 64, M = 25$, and PSNR $= 15$ dB. Results are averaged over 10^3 independent runs. We observe that when the frequency components are very close to each other, the SDP can hardly identify the true frequency parameters, whereas the nonconvex optimziation-based methods, including the SURE-IR, the DicRefCS and the SBL-DE are still capable of resolving these closely-spaced components with decent success rates. The OGSBI method does not perform as well as the other three nonconvex optimziation-based methods in resolving closely-spaced frequency components, possibly because the Taylor expansion-based approximation is not accurate enough to distinguish closely located complex sinusoids.

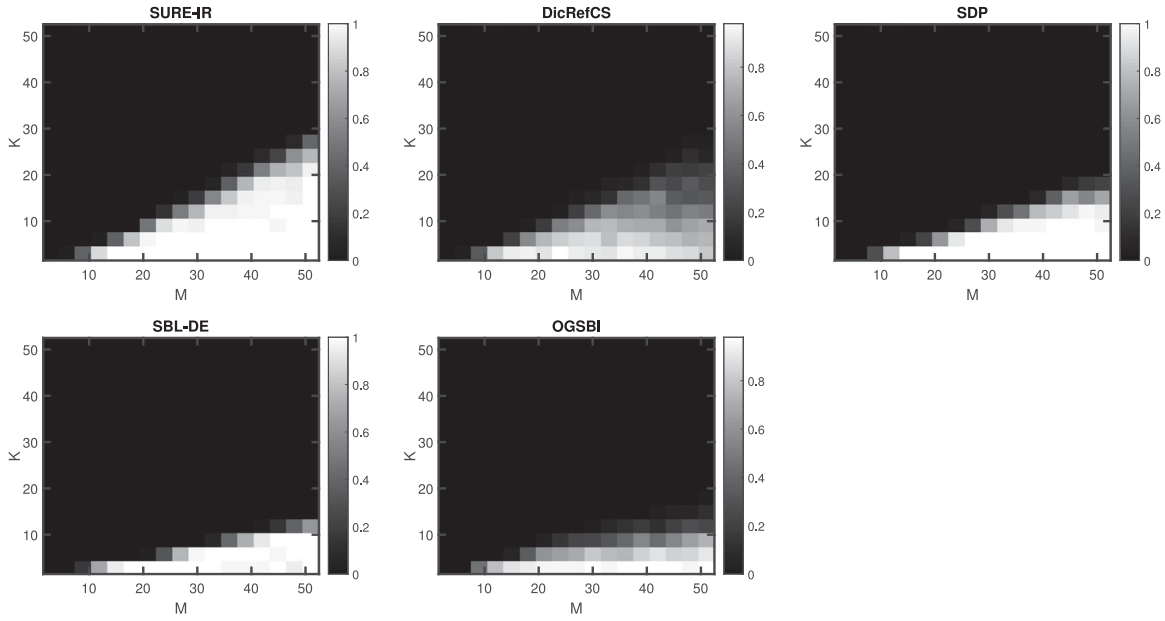

Fig. 3. Phase transitions of respective algorithms, $T = 64$, and PSRN $= 25$ dB.

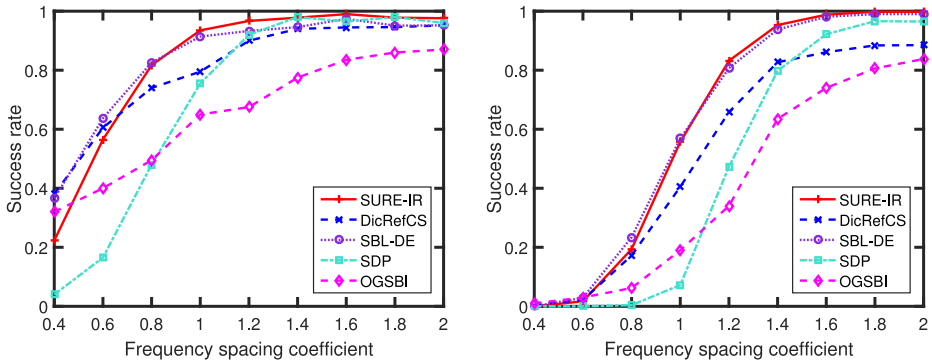

Fig. 4. Success rates of respective algorithms vs. the frequency spacing coefficient μ, $T = 64$, $M = 25$, PSNR $= 15$ dB. (a) $K = 2$. (b) $K = 3$.

Our last experiment tests the recovery performance of respective algorithms using a real-world amplitude modulated (AM) signal [5], [37] that encodes the message appearing in the top left corner of Fig. 6. The signal was transmitted from a communication device using carrier frequency $\omega_c = 8.2$ kHz, and the received signal was sampled by an analog-digital converter (ADC) at a rate of 32 kHz. The sampled signal has a total number of 32768 samples. For the sake of computational efficiency, in our experiment, the AM signal is divided into a number of short-time segments, each consisting of $T = 1024$ data samples. For each segment, we randomly select M data samples, based on which we use respective algorithms to recover the whole segment. After all segments are reconstructed, we perform AM demodulation on the recovered signal to reconstruct the original message. The RSNR is then computed using the reconstructed message and the true message. Fig. 5 plots the RSNRs of respective algorithms vs. the ratio M/T (the SDP method was not included in this experiment due to its prohibitive computational complexity when the signal dimension is large). We see that our proposed SURE-IR method offers the best performance and presents a significant performance advantage over the other algorithms for a small ratio M/T, where data acquisition is more beneficial due to high compression rates. In particular, when $M/T = 0.02$, all

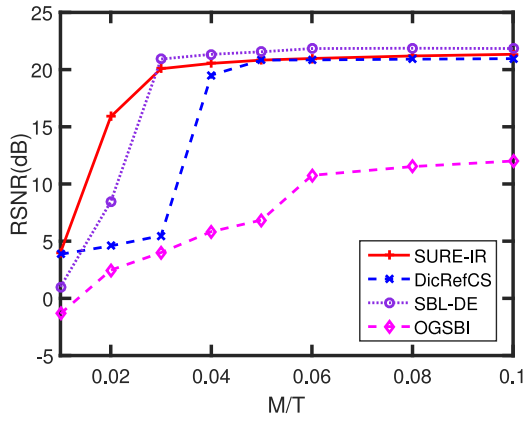

Fig. 5. RSNRs of respective algorithms vs. the ratio M/T.

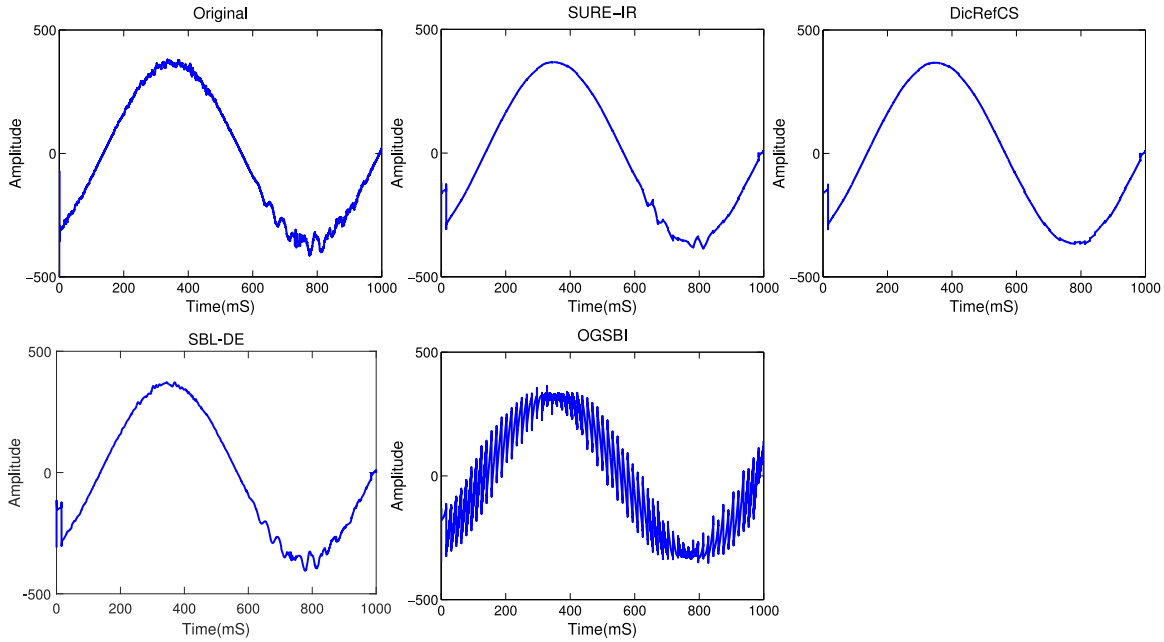

Fig. 6. The true message and the messages reconstructed by respective algorithms, $M = 100$.

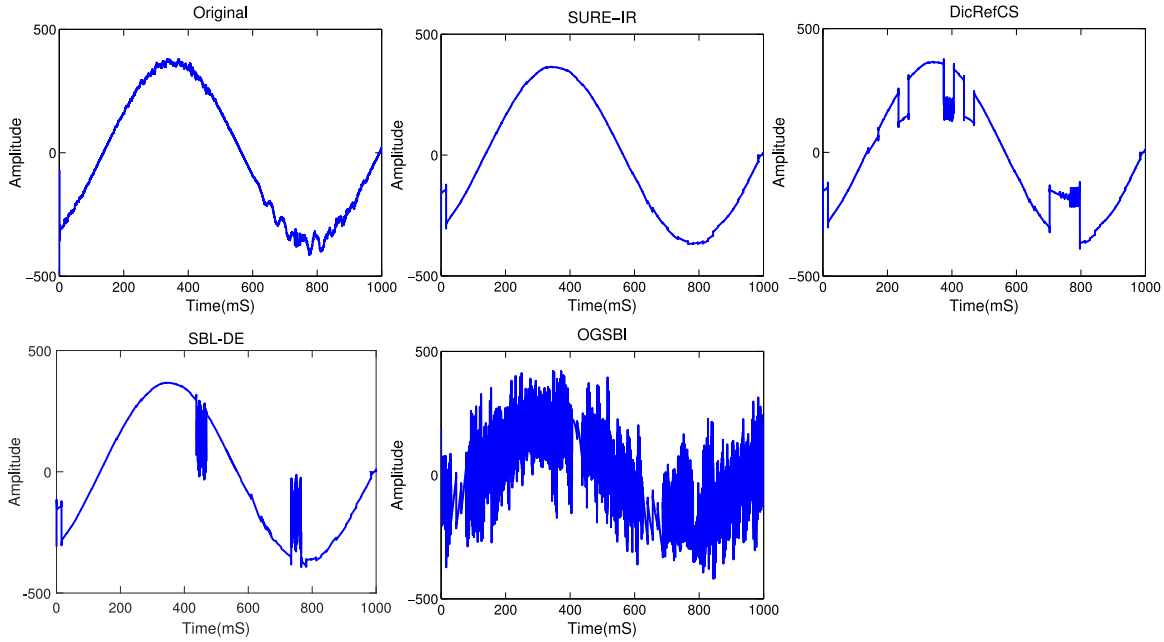

Fig. 7. The true message and the messages reconstructed by respective algorithms, $M = 20$.

the other methods (DicRefCS, SBL-DE and OGSBI) fail to provide an accurate reconstruction, while our proposed algorithm still renders a decent recovery accuracy. Figs. 6 and 7 show the true message and the messages recovered by respective algorithms, where M is set to 20 and 100, respectively. It can be seen that our proposed algorithm can obtain a fairly accurate reconstruction of the original signal even with as few as $M = 20$ measurements, whereas the message reconstructed by the other three methods, particularly the OGSBI, is highly smeared/distorted. The average running times of respective algorithms are also provided in Table I. We see that the SBL-DE is the most computationally efficient algorithm, whereas our proposed method has a higher computational complexity than the other three methods. This is because in the second step of our proposed algorithm, the dictionary parameters $\{\theta_i\}$ are refined in a sequential

TABLE I
AVERAGE RUNNING TIMES OF RESPECTIVE ALGORITHMS (SEC)

Algorithm	$M = 20$	$M = 60$	$M = 100$
SURE-IR	118.78	124.02	140.16
DicRefCS	11.95	12.55	15.92
SBL-DE	3.034	3.761	4.867
OGSBI	28.13	29.24	33.34

manner. This sequential refinement leads to a more stable estimate of $\{\theta_i\}$, but meanwhile resulting in an increased computational complexity.

VIII. Conclusion

This paper studied the super-resolution compressed sensing problem where the sparsifying dictionary is characterized by a set of unknown parameters in a continuous domain. Such a problem arises in many practical applications such as direction-of-arrival estimation and line spectral estimation. By resorting to the majorization-minimization approach, we developed a generalized iterative reweighted ℓ_2 algorithm for joint dictionary parameter learning and sparse signal recovery. The proposed algorithm iteratively decreases a surrogate function majorizing a given objective function, leading to a gradual and interweaved iterative process to refine the unknown parameters and the sparse signal. Simulation results show that our proposed algorithm effectively overcomes the grid mismatch problem and achieves a super-resolution accuracy in resolving the unknown frequency parameters. The proposed algorithm also demonstrates superiority over several existing super-resolution compressed sensing methods in resolving the unknown parameters and reconstructing the original signal.

Appendix A
Derivative of $f(\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$

Define

$$\boldsymbol{X} \triangleq \boldsymbol{A}\left(\boldsymbol{\theta}\right)\left(\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)\boldsymbol{A}\left(\boldsymbol{\theta}\right)+\lambda^{-1}\boldsymbol{D}^{(t)}\right)^{-1}\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)$$

Using the chain rule, the first derivative of $f(\boldsymbol{\theta})$ with respect to $\theta_i, \forall i$ can be computed as

$$\frac{\partial f\left(\boldsymbol{\theta}\right)}{\partial\theta_i}=\mathrm{tr}\left\{\left(\frac{\partial f\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{X}}\right)^T\frac{\partial\boldsymbol{X}}{\partial\theta_i}\right\}+\mathrm{tr}\left\{\left(\frac{\partial f\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{X}^*}\right)^T\frac{\partial\boldsymbol{X}^*}{\partial\theta_i}\right\}$$

where $\boldsymbol{X}^*$ donates the conjugate of the complex matrix $\boldsymbol{X}$, and

$$\frac{\partial f\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{X}}=\frac{\partial}{\partial\boldsymbol{X}}\mathrm{tr}\left\{-\boldsymbol{y}\boldsymbol{y}^H\boldsymbol{X}\right\}=-\left(\boldsymbol{y}\boldsymbol{y}^H\right)^T$$

$$\frac{\partial f\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{X}^*}=\frac{\partial}{\partial\boldsymbol{X}^*}\mathrm{tr}\left\{-\boldsymbol{y}\boldsymbol{y}^H\boldsymbol{X}\right\}=\boldsymbol{0}$$

$$\begin{aligned}\frac{\partial\boldsymbol{X}}{\partial\theta_i}&=\frac{\partial}{\partial\theta_i}\left(\boldsymbol{A}\left(\boldsymbol{\theta}\right)\left(\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)\boldsymbol{A}\left(\boldsymbol{\theta}\right)+\lambda^{-1}\boldsymbol{D}^{(t)}\right)^{-1}\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)\right)\\&=\frac{\partial\boldsymbol{A}\left(\boldsymbol{\theta}\right)}{\partial\theta_i}\left(\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)\boldsymbol{A}\left(\boldsymbol{\theta}\right)+\lambda^{-1}\boldsymbol{D}^{(t)}\right)^{-1}\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)\\&\quad+\boldsymbol{A}\left(\boldsymbol{\theta}\right)\left(\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)\boldsymbol{A}\left(\boldsymbol{\theta}\right)+\lambda^{-1}\boldsymbol{D}^{(t)}\right)^{-1}\frac{\partial\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)}{\partial\theta_i}\\&\quad-\boldsymbol{A}\left(\boldsymbol{\theta}\right)\left(\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)\boldsymbol{A}\left(\boldsymbol{\theta}\right)+\lambda^{-1}\boldsymbol{D}^{(t)}\right)^{-1}\left(\frac{\partial\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)}{\partial\theta_i}\boldsymbol{A}\left(\boldsymbol{\theta}\right)\right.\\&\quad\left.+\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)\frac{\partial\boldsymbol{A}\left(\boldsymbol{\theta}\right)}{\partial\theta_i}\right)\left(\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)\boldsymbol{A}\left(\boldsymbol{\theta}\right)+\lambda^{-1}\boldsymbol{D}^{(t)}\right)^{-1}\boldsymbol{A}^H\left(\boldsymbol{\theta}\right)\end{aligned}$$

Appendix B
Proof of Theorem 1

From (28), we have

$$\sum_{i=1}^{p}\hat{z}_i\boldsymbol{a}\left(\hat{\theta}_i\right)+\sum_{i=1}^{K}\eta_i\boldsymbol{a}\left(\omega_i\right)=0 \tag{38}$$

The above equation can be rewritten as

$$-\sum_{i=M-K+1}^{p}\hat{z}_i\boldsymbol{a}\left(\hat{\theta}_i\right)=\sum_{i=1}^{M-K}\hat{z}_i\boldsymbol{a}\left(\hat{\theta}_i\right)+\sum_{i=1}^{K}\eta_i\boldsymbol{a}\left(\omega_i\right) \tag{39}$$

Taking the norm of both sides of the above equation, we obtain

$$\left\|\sum_{i=M-K+1}^{p}\hat{z}_i\boldsymbol{a}\left(\hat{\theta}_i\right)\right\|_2^2=\|\boldsymbol{B}\boldsymbol{x}_0\|_2^2 \tag{40}$$

where $\boldsymbol{B}$ is an $M\times M$ square matrix defined as

$$\boldsymbol{B}\triangleq\left[\boldsymbol{a}\left(\omega_1\right)\ \ldots\ \boldsymbol{a}\left(\omega_K\right)\ \boldsymbol{a}\left(\hat{\theta}_1\right)\ \ldots\ \boldsymbol{a}\left(\hat{\theta}_{M-K}\right)\right]$$

and

$$\boldsymbol{x}_0\triangleq\left[\eta_1\ \ldots\ \eta_K\ \hat{z}_1\ \ldots\ \hat{z}_{M-K}\right]^T$$

From (40), we would like to obtain a lower bound on $|\hat{z}_{M-K+1}|$. To this objective, we first seek a lower bound for $\|\boldsymbol{B}\boldsymbol{x}_0\|_2$. Clearly we have

$$\frac{\|\boldsymbol{B}\boldsymbol{x}_0\|_2^2}{\|\boldsymbol{x}_0\|_2^2}\geq\min_{\boldsymbol{x}}\frac{\|\boldsymbol{B}\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}=\sigma_{\min}\left(\boldsymbol{B}\right) \tag{41}$$

The following lemma provides a lower bound on the minimum singular value of any square matrix.

Lemma 1: Let $\boldsymbol{X}\in\mathbb{C}^{M\times M}$, and $\sigma_{\min}(\boldsymbol{X})$ denote the minimum singular value of $\boldsymbol{X}$. Then

$$\begin{aligned}\sigma_{\min}\left(\boldsymbol{X}\right)\ \geq\ &\left(\frac{M-1}{M}\right)^{(M-1)/2}\times\left|\det\left(\boldsymbol{X}\right)\right|\\&\times\max\left\{\frac{c_{\min}\left(\boldsymbol{X}\right)}{\prod_{i=1}^{M}c_i\left(\boldsymbol{X}\right)},\frac{r_{\min}\left(\boldsymbol{X}\right)}{\prod_{i=1}^{M}r_i\left(\boldsymbol{X}\right)}\right\}\end{aligned} \tag{42}$$

where $r_i(\boldsymbol{X}), c_i(\boldsymbol{X})$ denote the 2-norm of the ith row and column of $\boldsymbol{X}$, respectively, and $r_{\min}(\boldsymbol{X}), c_{\min}(\boldsymbol{X})$ denote the minimum of $r_i(\boldsymbol{X})$ and $c_i(\boldsymbol{X})$, respectively.

Proof: Please refer to [38], [39] for a rigorous proof. ■

Note that for the matrix $\boldsymbol{B}$ defined in (40), we can easily verify that $r_i(\boldsymbol{B})=c_i(\boldsymbol{B})=\sqrt{M},\forall i$, and since $\boldsymbol{B}$ is a square Vandermonde matrix, its determinant is given by

$$\det\left(\boldsymbol{B}\right)=\prod_{1\leq l<m\leq M}\left(e^{j\varphi_l}-e^{j\varphi_m}\right) \tag{43}$$

where $\{\varphi_i\}_{i=1}^{M} \triangleq S_\omega \cup \{\widehat{\theta}_i\}_{i=1}^{M-K}$. By utilizing Lemma 1, we have

$$\begin{aligned}\sigma_{\min}(\boldsymbol{B}) &\geq \left(\frac{\sqrt{M-1}}{M}\right)^{M-1} \left|\prod_{1\leq l<m\leq M}\left(e^{j\varphi_l}-e^{j\varphi_m}\right)\right| \\ &\overset{(a)}{\geq} \left(\frac{\sqrt{M-1}}{M}\right)^{M-1} |2\sin(\nu/2)|^{\frac{M(M-1)}{2}} \triangleq C_1\end{aligned} \tag{44}$$

where (a) comes from $|e^{j\varphi_l}-e^{j\varphi_m}| \geq 2|\sin(\nu/2)|$. From (41) and (44), we have

$$\|\boldsymbol{B}\boldsymbol{x}_0\|_2^2 \geq C_1 \|\boldsymbol{x}_0\|_2^2 \geq C_1\eta_{\max}^2 > C_1\tau^2 \tag{45}$$

On the other hand, the term on the left-hand side of (40) can be upper bounded by

$$\begin{aligned}\left\|\sum_{i=M-K+1}^{p}\hat{z}_i\boldsymbol{a}\left(\hat{\theta}_i\right)\right\|_2^2 &\leq \sum_{i=M-K+1}^{p}\left\|\hat{z}_i\boldsymbol{a}\left(\hat{\theta}_i\right)\right\|_2^2 \\ &= \sum_{i=M-K+1}^{p} M|\hat{z}_i|^2 \\ &\leq M(p-M+K)|\hat{z}_{M-K+1}|^2\end{aligned} \tag{46}$$

where the last inequality follows from the fact that entries $\{\hat{z}_i\}_{i=1}^{p}$ are sorted in a descending order. Combining (40), (45) and (46), we arrive at $|\hat{z}_{M-K+1}|$ is lower bounded by

$$|\hat{z}_{M-K+1}|^2 \geq \frac{C_1\tau^2}{M(p-M+K)} \geq \frac{C_1\tau^2}{M(N-M+K)} \triangleq C \tag{47}$$

Since entries $\{\hat{z}_i\}_{i=1}^{p}$ are sorted in a descending order, we have

$$|\hat{z}_i|^2 \geq C \quad i=1,\ldots,M-K+1 \tag{48}$$

Based on (48), a lower bound on $L(\widehat{\boldsymbol{z}})$ can be readily obtained as

$$\begin{aligned}L(\widehat{\boldsymbol{z}}) &= \sum_{i=1}^{N}\log\left(|\hat{z}_i|^2+\epsilon\right) \\ &= \sum_{i=1}^{p}\log\left(|\hat{z}_i|^2+\epsilon\right) \\ &\quad + \sum_{i=1}^{J-p}\log\left(|\hat{z}_{p+i}|^2+\epsilon\right) + (N-J)\log\epsilon \\ &\geq (M-K+1)\log C + [N-(M-K+1)]\log\epsilon \\ &\triangleq g_{\mathrm{LB}}\end{aligned} \tag{49}$$

On the other hand, $L(\boldsymbol{z}_0)$ is upper bounded by

$$\begin{aligned}L(\boldsymbol{z}_0) &= \sum_{i=1}^{K}\log\left(|\alpha_i|^2+\epsilon\right) + (N-K)\log\epsilon \\ &\leq K\log\left(2\alpha_{\max}^2\right) + (N-K)\log\epsilon \triangleq g_{\mathrm{UB}}\end{aligned} \tag{50}$$

When the condition (31) is satisfied, we have $g_{\mathrm{LB}} \geq g_{\mathrm{UB}}$. As a consequence, the inequality $L(\widehat{\boldsymbol{z}}) > L(\boldsymbol{z}_0)$ holds valid from (49)–(50). This completes our proof.

REFERENCES

[1] Y. Chi, L. L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2182–2195, May 2011.

[2] Z. Yang, L. Xie, and C. Zhang, "Off-grid direction of arrival estimation using sparse Bayesian inference," *IEEE Trans. Signal Process.*, vol. 61, no. 1, pp. 38–42, Jan. 2013.

[3] Z. Tan, P. Yang, and A. Nehorai, "Joint sparse recovery method for compressed sensing with structured dictionary mismatches," *IEEE Trans. Signal Process.*, vol. 62, no. 19, pp. 4997–5008, Oct. 2014.

[4] A. Fannjiang and W. Liao, "Coherence pattern-guided compressive sensing with unresolved grids," *SIAM J. Imag. Sci.*, vol. 5, no. 1, pp. 179–202, 2012.

[5] M. F. Duarte and R. G. Baraniuk, "Spectral compressive sensing," *Appl. Comput. Harmon. Anal.*, vol. 35, pp. 111–129, 2013.

[6] E. Candès and C. Fernandez-Granda, "Towards a mathematical theory of super-resolution," *Commun. Pure Appl. Math.*, vol. 67, no. 6, pp. 906–956, Jun. 2014.

[7] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht, "Compressed sensing off the grid," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7465–7490, Nov. 2013.

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

[9] G. Tang, B. N. Bhaskar, and B. Recht, "Near minimax line spectral estimation," *IEEE Trans. Inf. Theory*, vol. 61, no. 1, pp. 499–512, Jan. 2015.

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

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

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

[13] T. L. Hansen, M. A. Badiu, B. H. Fleury, and B. D. Rao, "A sparse Bayesian learning algorithm with dictionary parameter estimation," presented at the IEEE 8th Sensor Array Multichannel Signal Process. Workshop (SAM), A Coruña, 2014.

[14] M. Yaghoobi, L. Daudet, and M. E. Davies, "Parametric dictionary design for sparse coding," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4800–4810, Dec. 2009.

[15] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstructions from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 699–616, Mar. 1997.

[16] B. D. Rao and K. Kreutz-Delgado, "An affine scaling methodology for best basis selection," *IEEE Trans. Signal Process.*, vol. 47, no. 1, pp. 187–200, Jan. 1999.

[17] E. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877–905, Dec. 2008.

[18] R. Chartrand and W. Yin, "Iterative reweighted algorithm for compressive sensing," presented at the IEEE Int. Conf. Acoust., Speech, Signal Process., Las Vegas, 2008.

[19] D. Wipf and S. Nagarajan, "Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 317–329, Apr. 2010.

[20] 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 Process. Lett.*, vol. 21, no. 6, pp. 761–765, Jun. 2014.

[21] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, no. 3, pp. 276–280, Mar. 1986.

[22] R. Roy and T. Kailath, "ESPRIT-estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[23] Y. Hua and T. K. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 5, pp. 814–824, May 1990.

[24] B. Friedlander, "An efficient parametric technique for Doppler-delay estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3953–3963, Aug. 2012.
[25] Z. Tan, Y. C. Eldar, and A. Nehorai, "Direction of arrival estimation using co-prime arrays: A super resolution viewpoint," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5565–5576, Nov. 2014.
[26] Y. Shen, J. Fang, and H. Li, "Exact reconstruction analysis of log-sum minimization for compressed sensing," *IEEE Signal Process. Lett.*, vol. 20, pp. 1223–1226, Dec. 2013.
[27] M. Ataee, H. Zayyani, M. Babaie-Zadeh, and C. Jutten, "Parametric dictionary learning using steepest descent," presented at the IEEE Int. Conf. Acoust., Speech, Signal Process., Dallas, 2010.
[28] K. Lange, D. Hunter, and I. Yang, "Optimization transfer using surrogate objective functions," *J. Comput. Graph. Statist.*, vol. 9, no. 1, pp. 1–20, Mar. 2000.
[29] I. Daubechies, R. Devore, M. Fornasier, and C. S. Gunturk, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math.*, vol. LXIII, pp. 1–38, 2010.
[30] D. Ba, B. Babadi, P. L. Purdon, and E. N. Brown, "Convergence and stability of iteratively re-weighted least squares algorithms," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 183–195, Jan. 2014.
[31] F. Bauer and M. A. Lukas, "Comparing parameter choice methods for regularization of ill-posed problems," *Math. Comput. Simul.*, vol. 81, no. 9, pp. 1795–1841, May 2011.
[32] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.
[33] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 2008.
[34] J. Fang, Y. Shen, H. Li, and P. Wang, "Pattern-coupled sparse Bayesian learning for recovery of block-sparse signals," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 360–372, Jan. 2015.
[35] Y. Li and Y. Chi, "Off-the-grid line spectrum denoising and estimation with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 64, no. 5, pp. 1257–1269, Mar. 2016.
[36] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, July 2005.
[37] J. A. Tropp, J. N. Laska, M. F. Duarte, J. K. Romberg, and R. G. Baraniuk, "Beyond Nyquist: Efficient sampling of sparse bandlimited signals," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 520–544, Jan. 2010.
[38] Y.-S. Yu and D.-H. Gu, "A note on a lower bound for the smallest singular value," *Linear Algebra Appl.*, vol. 253, no. 1, pp. 25–38, 1997.
[39] Y. P. Hong and C.-T. Pan, "A lower bound for the smallest singular value," *Linear Algebra Appl.*, vol. 172, pp. 27–32, 1992.

Jun Fang (M'08) received the B.S. and M.S. degrees from the Xidian University, Xi'an, China, in 1998 and 2001, respectively, and the Ph.D. degree from the National University of Singapore, Singapore, in 2006, all in electrical engineering.

During 2006, he was a postdoctoral research associate in the Department of Electrical and Computer Engineering, Duke University. From January 2007 to December 2010, he was a research associate with the Department of Electrical and Computer Engineering, Stevens Institute of Technology. Since January 2011, he has been with the University of Electronic of Science and Technology of China. His current research interests include sparse theory and compressed sensing, and Bayesian inference for data analysis.

Dr. Fang received the IEEE Jack Neubauer Memorial Award in 2013 for the best systems paper published in the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He is an Associate Technical Editor for *IEEE Communications Magazine*, and an Associate Editor for IEEE SIGNAL PROCESSING LETTERS.

Feiyu Wang received his B.Sc. degree from the Zhejiang Ocean University in 2011, and his M.Sc. degree from Ningbo University in 2014. Since August 2014, he has been working towards his Ph.D. degree with the University of Electronic of Science and Technology of China. His current research interests include compressed sensing, sparse theory and statistical learning.

Yanning Shen received her B.Sc. and M.Sc. degrees in electrical engineering from University of Electronic Science and Technology of China, Chengdu, in 2011 and 2014, respectively. Since August 2014, she has been working towards her Ph.D. degree with the Department of Electrical and Computer Engineering, U University of Minnesota. Her research interests include sparse signal processing and machine learning.

Hongbin Li (M'99–SM'08) received the B.S. and M.S. degrees from the University of Electronic Science and Technology of China, in 1991 and 1994, respectively, and the Ph.D. degree from the University of Florida, Gainesville, FL, in 1999, all in electrical engineering.

From July 1996 to May 1999, he was a Research Assistant in the Department of Electrical and Computer Engineering at the University of Florida. Since July 1999, he has been with the Department of Electrical and Computer Engineering, S Stevens Institute of Technology, Hoboken, NJ, where he is a Professor. He was a Summer Visiting Faculty Member at the Air Force Research Laboratory in the summers of 2003, 2004 and 2009. His general research interests include statistical signal processing, wireless communications, and radars.

Dr. Li received the IEEE Jack Neubauer Memorial Award in 2013 for the best systems paper published in the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, the Outstanding Paper Award from the IEEE AFICON Conference in 2011, the Harvey N. Davis Teaching Award in 2003 and the Jess H. Davis Memorial Award for excellence in research in 2001 from Stevens Institute of Technology, and the Sigma Xi Graduate Research Award from the University of Florida in 1999. He has been a member of the IEEE SPS Signal Processing Theory and Methods (2011 to now) Technical Committee (TC) and the IEEE SPS Sensor Array and Multichannel TC (2006 to 2012). He has been an Associate Editor for *Signal Processing* (Elsevier) (since 2013), the IEEE TRANSACTIONS ON SIGNAL PROCESSING (2006 to 2009 and since 2014), IEEE SIGNAL PROCESSING LETTERS (2005 to 2006), and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (2003 to 2006), as well as a Guest Editor for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, and the *EURASIP Journal on Applied Signal Processing*. He has been involved in various conference organization activities, including serving as a General Co-Chair for the 7th IEEE Sensor Array and Multichannel Signal Processing (SAM) Workshop, Hoboken, NJ, June 17–20, 2012. Dr. Li is a member of Tau Beta Pi and Phi Kappa Phi.

Rick S. Blum (S'83–M'84–SM'94–F'05) received a B.S. in electrical engineering from the Pennsylvania State University in 1984 and his M.S. and Ph.D. in electrical engineering from the University of Pennsylvania i in 1987 and 1991. From 1984 to 1991 he was a member of technical staff at General Electric Aerospace in Valley Forge, Pennsylvania, and he graduated from GE's Advanced Course in Engineering. Since 1991, he has been with the Electrical and Computer Engineering Department at Lehigh University in Bethlehem, Pennsylvania, where he is currently a Professor and holds the Robert W. Wieseman Chaired Research Professorship i in Electrical Engineering. His research interests include signal processing f for smart grid, communications, sensor networking, radar and sensor processing. He is on the editorial board for the *Journal of Advances in Information Fusion* of the International Society of Information Fusion. He was an associate editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and for IEEE COMMUNICATIONS LETTERS. He has edited special issues for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, and the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He is a member of the SAM Technical Committee (TC) of the IEEE Signal Processing Society. He was a member of the Signal Processing for Communications TC of the IEEE Signal Processing Society and is a member of the Communications Theory TC of the IEEE Communication Society. He was on the awards Committee of the IEEE Communication Society.

Dr. Blum is a former IEEE Signal Processing Society Distinguished Lecturer, an IEEE Third Millennium Medal winner, a member of Eta Kappa Nu and Sigma Xi, and holds several patents. He was awarded an ONR Young Investigator Award and an NSF Research Initiation Award. His IEEE Fellow Citation "for scientific contributions to detection, data fusion and signal processing with multiple sensors" acknowledges contributions to the field of sensor networking.