Robust Recovery of Temporally Smooth Signals From Under-Determined Multiple Measurements

Zhaofu Chen, Rafael Molina, *Member, IEEE*, and Aggelos K. Katsaggelos, *Fellow, IEEE*

***Abstract*—In this paper, we consider the problem of recovering jointly sparse vectors from underdetermined measurements that are corrupted by both additive noise and outliers. This can be viewed as the robust extension of the Multiple Measurement Vector (MMV) problem. To solve this problem, we propose two general approaches. As a benchmark, the first approach preprocesses the input for outlier removal and then employs state-of-the-art technologies for signal recovery. The second approach, as the main contribution of this paper, is based on formulation of an innovative regularized fitting problem. By solving the regularized fitting problem, we jointly remove outliers and recover the sparse vectors. Furthermore, by exploiting temporal smoothness among the sparse vectors, we improve noise robustness of the proposed approach and avoid the problem of over-fitting. Extensive numerical results are provided to illustrate the excellent performance of the proposed approach.**

***Index Terms*—Signal reconstruction, iterative methods, optimization.**

I. Introduction

CONSIDER the measurement system expressed as follows

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{y} = [y_1, \ldots, y_M]^{\mathrm{T}} \in \mathbb{R}^{M\times 1}$ denotes a measurement vector, $\mathbf{x} = [x_1, \ldots, x_N]^{\mathrm{T}} \in \mathbb{R}^{N\times 1}$ is the underlying signal vector of interest, $\mathbf{\Phi} = [\phi_{ij}] \in \mathbb{R}^{M\times N}$ represents a linear transformation applied on the signal, and $\mathbf{n} = [n_1, \ldots, n_M]^{\mathrm{T}} \in \mathbb{R}^{M\times 1}$ represents additive noise or un-modeled errors. The general system in (1) has been applied to model a wide range of observation or measurement processes. Depending on the application, the matrix $\mathbf{\Phi}$ can be a set of features in machine learning (e.g., feature selection) [1], a dictionary in sparse signal representation problems [2], a measurement system in compressive sensing [3], a degradation process in image restoration/recovery [4], a steering matrix in array signal processing or source localization [5], etc.

Note that in general the matrix $\mathbf{\Phi}$ is wide, i.e., $M \leq N$. For example, in compressive sensing, the number of measurements taken is far less than the number of signal samples. In feature selection, a large set of features serve as a pool of candidates from which the most descriptive ones are selected. In source localization, a dense grid spanning the search region is constructed such that each grid point corresponds to one column in $\mathbf{\Phi}$. In all such applications, the rectangular shape of $\mathbf{\Phi}$ renders the inverse problem of finding $\mathbf{x} \in \mathbb{R}^{N\times 1}$ from $\mathbf{y} \in \mathbb{R}^{M\times 1}$ ill-posed, i.e., there exist infinitely many solutions.

In order to constrain the solution space, prior information about the characteristics of $\mathbf{x}$ must be incorporated. As is well known, most natural signals are sparse either in their native domains, or can be sparsely represented under certain transformations. For example, in source localization problems, the number of true signal sources is far less than the number of scanning grid points; hence most of the entries in $\mathbf{x}$ do not correspond to source locations and are zeros [5]. As another example, most natural images can be sparsely represented under the Discrete Cosine Transform (DCT) or Discrete Wavelet Transform (DWT) by retaining a small subset of coefficients. Without loss of generality, we assume the signal vector $\mathbf{x}$ is sparse, i.e., $\|\mathbf{x}\|_0 \ll N$, where the ℓ_0-(pseudo)norm simply counts the number of nonzeros in $\mathbf{x}$.

The estimation of a sparse $\mathbf{x}$ from a set of under-determined measurements $\mathbf{y}$ is commonly referred to sparse signal recovery with a single measurement vector. This is a well-studied problem. Broadly, the algorithms for solving this type of problems fall into one of the following three categories. Greedy algorithms, represented by Matching Pursuit [6] and Orthogonal Matching Pursuit [7], [8], incrementally seek a subset of the columns in $\mathbf{\Phi}$ that are most correlated with the observation $\mathbf{y}$, and determine the corresponding entries in $\mathbf{x}$ by solving a series of Least Squares (LS) fitting problems. Relaxation-based approaches, such as LASSO [9] and FOCUSS [10], replace the non-convex ℓ_0-(pseudo)norm with convex ℓ_p-norm (with $p \geq 1$, and hence the term "relaxation"), and solve a regularized fitting problem. Last but not least, Bayesian approaches adopt sparsity-promoting priors on $\mathbf{x}$ and approximate the posterior distribution of $\mathbf{x}$ given $\mathbf{y}$ [11], [12].

When the measurement process occurs at T time instances, the system in (1) is extended as follows

$$\mathbf{Y} = \mathbf{\Phi}\mathbf{X} + \mathbf{N}, \tag{2}$$

where $\mathbf{Y} \in \mathbb{R}^{M\times T}$ contains the measurement vectors $\{\mathbf{y}_{\cdot i}\}_{i=1}^{T}$ as its columns, and $\mathbf{X} \in \mathbb{R}^{N\times T}$ contains the underlying signal

Manuscript received May 27, 2014; revised October 03, 2014 and January 21, 2015; accepted January 26, 2015. Date of publication February 12, 2015; date of current version March 02, 2015. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hing Cheung So. This work has been partially supported by a grant from the Department of Energy (DE-NA0000568), the Spanish Ministry of Economy and Competitiveness under project TIN2013-43880-R, the European Regional Development Fund (FEDER), and the CEI BioTic at the Universidad de Granada. (*Corresponding author: A. K. Katsaggelos.*)

Z. Chen and A. K. Katsaggelos are with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA (e-mail: zhaofuchen2014@u.northwestern.edu; aggk@eecs.northwestern.edu).

R. Molina is with the Deptartmento de Ciencias de la Computación e I. A., Universidad de Granada, 18071 Granada, Spain (e-mail: rms@decsai.ugr.es).

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.2015.2403277

vectors $\{\mathbf{x}_{\cdot i}\}_{i=1}^{T}$ as its columns, respectively. Notationally, for a matrix $\mathbf{X}$, we use $\mathbf{x}_{\cdot i}$ to denote its ith column (i.e., the row index is irrelevant and hence is denoted by a dot in the subscript).

Given T measurement vectors in $\mathbf{Y}$, various strategies can be applied depending on the dynamics of $\{\mathbf{x}_{\cdot i}\}_{i=1}^{T}$. On one end, if the signals $\{\mathbf{x}_{\cdot i}\}_{i=1}^{T}$ at different time instances are static, i.e., $\mathbf{x} = \mathbf{x}_{\cdot 1} = \cdots = \mathbf{x}_{\cdot T}$, then the columns of $\mathbf{Y}$ are simply repeated measurements of the same underlying signal $\mathbf{x}$ subject to independent realizations of noise $\mathbf{N} = [\mathbf{n}_{\cdot 1}, \ldots, \mathbf{n}_{\cdot T}]$. In this case, averaging over the measurements yields $\bar{\mathbf{y}} = \sum_{i=1}^{T} \mathbf{y}_{\cdot i}/T$ with improved Signal-to-Noise-Ratio (SNR), which can then be considered as a single measurement. On the other end, if the signals $\{\mathbf{x}_{\cdot i}\}_{i=1}^{T}$ are independent of each other, the problem reduces to T independent instances of single measurement sparse recovery problems, each of which can be solved separately.

Between these two ends, it is most often the case that $\{\mathbf{x}_{\cdot i}\}_{i=1}^{T}$, though not identical replicas of each other, are correlated. In particular, it is common to assume that $\{\mathbf{x}_{\cdot i}\}_{i=1}^{T}$ share a fixed sparsity profile, i.e., the locations of the nonzeros in $\mathbf{x}_{\cdot i}$ do not change with i. This is known as the Multiple Measurement Vector (MMV) problem, where the goal is to recover a row-wise sparse $\mathbf{X}$ from $\mathbf{Y}$. The MMV problem is closely related to other research fields, such as Multi-Task Sparse Learning [13], dynamic compressive sensing [14], etc. MMV is an actively researched problem, and numerous algorithmic solutions exist. Representative examples include greedy algorithm [15], relaxation-based optimization [16], [17], subspace approaches [18], statistical inference based method [19]–[21], and sparse support estimation algorithms [22], [23]. For theoretical analyses of algorithms solving the MMV problem, we refer the readers to [24] and [25].

While the existing algorithms induce row-wise sparsity in $\mathbf{X}$, they often do not fully exploit the information in $\mathbf{X}$, and in particular, the temporal correlation within the nonzero rows of $\mathbf{X}$ had been largely neglected until recently. Utilizing such intra-row structure of $\mathbf{X}$ can potentially lead to improved recovery accuracy. In [26] and [27], temporal correlation has been modeled within a block Sparse Bayesian Learning (SBL) framework, and inference based on the Expectation-Maximization (EM) algorithm has been introduced to determine the posterior distributions. The temporal structure is adaptively learned from the data, and such structure is in turn used to refine the signal estimates. Alternatively, [28] has proposed a hierarchical Bayesian model to characterize the temporal smoothness within the nonzero rows of $\mathbf{X}$, and has adopted both an EM based and a fixed-point iteration based algorithms for inference. In addition, we have proposed two deterministic algorithms in [29] to find temporally smooth and row-wise sparse $\mathbf{X}$ that fits the measurement $\mathbf{Y}$ well. Experimental results confirm that by taking into account the temporal structure of $\mathbf{X}$, better recovery accuracy can be obtained.

Note that one assumption with the model in (2) is that the additive noise $\mathbf{N}$ is generally dense and "small", that is, noise is present in all measurements and its amplitude is small compared with that of the measurements. Based on this assumption, most of the algorithms mentioned above either explicitly (e.g., in relaxation-based and Bayesian approaches) or implicitly (e.g., in greedy approaches) seek $\mathbf{X}$ that both conforms with the prior information and also yields small fitting residual, i.e., $\mathbf{Y} - \mathbf{\Phi X}$ has little energy remaining.

However, this assumption can often be challenged, when anomalies or outliers are present in the measurements. Such outliers can be due to various reasons, such as malfunctioning of the measurement equipments, missing measurement samples, and so on. As a concrete example, consider the use of electroencephalography (EEG) signals for the study of human brain activities. In the ideal case the signal acquisition process can be modeled using (2), where $\mathbf{Y}$ denotes the temporal recordings over multiple channels and $\mathbf{X}$ denotes the underlying cerebral activities, respectively. However, the measurement $\mathbf{Y}$ is very often corrupted by the presence of artifacts, due to movements of the subject under study [30]. For example, eye blinks result in isolated sharp spikes in the channel recordings, which, if not handled properly, will render the subsequent analysis very challenging or even impossible.

To account for the possible presence of anomalies/artifacts, the measurement system is modified as follows

$$\mathbf{Y} = \mathbf{\Phi X} + \mathbf{E} + \mathbf{N}, \tag{3}$$

where $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{N}$ are defined similarly as above in (2), and $\mathbf{E} \in \mathbb{R}^{M \times T}$ is a sparse matrix with arbitrarily large entries. In the aforementioned EEG example, $\mathbf{E}$ accounts for the artifacts due to subject movement, which must be identified and removed to make the analysis of $\mathbf{X}$ possible and meaningful.

We term the measurement system represented in (3) as the "Robust Multiple Measurement Vector" model, or the Robust MMV model for short. Also the problem of finding $\mathbf{X}$ and $\mathbf{E}$ given $\mathbf{Y}$ in (3) is termed the Robust MMV problem. Note that Robust MMV is a newly formulated problem, and hence there is no prior solution to it. In the following sections of the paper, we will develop two types of solutions to the Robust MMV problem. The first type of solutions, presented in Section II, is built upon the *sequential* cascade of a preprocessing module and an MMV signal recovery module. Depending on the specific implementations of these two modules, the sequential approach can have different variations, and hence it essentially represents a solution family. In particular, when we use the state-of-the-art algorithms to implement the preprocessing module and the signal recovery module, the sequential approach can be considered as a benchmark for performance evaluation. The second approach, whose details are presented in Section III, *simultaneously* detects/removes outliers and recovers the underlying signals via regularized minimization. The simultaneous approach, as will be shown with numerical examples, has superior performance compared with the sequential approaches.

The structure of this paper is outlined below with the major contributions highlighted:

- Section I: formulating Robust MMV problem that is more general and practical than the original MMV problem;
- Section II: developing a benchmark sequential approach to the formulated Robust MMV problem utilizing the state-of-the-art technologies;
- Section III: proposing an innovative simultaneous approach that has excellent outlier removal and signal recovery capabilities;
- Section IV: performing extensive numerical evaluation of both the sequential and simultaneous approaches.

Notation: Throughout this paper, matrices and vectors are denoted by uppercase and lowercase boldface letters, respectively.

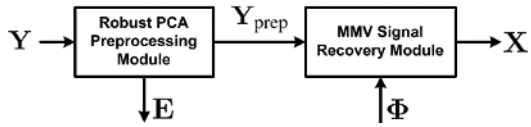


Fig. 1. Block diagram for the sequential approach.

For a matrix $\mathbf{X}$, its ith row, jth column, (i,j)th element are denoted by $\mathbf{x}_{i\cdot}$, $\mathbf{x}_{\cdot j}$, and X_{ij}, respectively. The $M\times N$ all-zero matrix is denoted by $\mathbf{0}_{M\times N}$, and the $M\times M$ identity matrix is denoted by $\mathbf{I}_M$, respectively. $\mathrm{vec}(\cdot)$ is the vectorization operator that stacks the columns of its input matrix into a column vector. The matrix Kronecker product is denoted by $\otimes$. The trace operator is denoted by $\mathrm{tr}(\cdot)$.

II. Sequential Outlier Removal and Signal Recovery for Robust MMV

In this section we present the details of the sequential approach to solving the Robust MMV problem. The block diagram for the sequential approach is shown in Fig. 1. As can be seen, the raw measurement $\mathbf{Y}$ first undergoes a preprocessing step to remove the outliers, and then the preprocessed measurement $\mathbf{Y}_{\mathrm{prep}}$ is fed as input to a core MMV signal recovery module.

One advantage of this sequential approach is the relative independence between the two modules in its structure, as shown in Fig. 1. Thanks to such structural and functional independence, testing and tuning of the two modules can be done separately, which makes system maintenance easy.

Note that both the preprocessing module and the signal recovery module define general functionality rather than specific implementations. In the following paragraphs, we discuss the options of implementing these two modules.

For the outlier-removal preprocessing module, one heuristic implementation is via thresholding, which is used in the analysis of EEG data [31]. In addition to the need of choosing the threshold, one major drawback of simple thresholding is the difficulty in setting the missing values after the outliers are detected. Typically the entire recording from an outlier-corrupted channel will be discarded, which clearly results in inefficient use of data. Alternatively, setting the values at the locations where outliers have been detected to zeros will potentially incur abrupt changes in the measurements, hence lead to degraded recovery accuracy.

Instead of simple thresholding, we present a preprocessing implementation based on matrix decomposition. Denote by K the number of nonzero rows in $\mathbf{X}$, it is clear that

$$\mathrm{rank}(\mathbf{\Phi X}) \le \mathrm{rank}(\mathbf{X}) \le K. \tag{4}$$

From the discussion above, we see the removal of outliers from measurements is essentially a matrix decomposition problem, where the goal is to decompose $\mathbf{Y}$ into a low-rank component $\mathbf{\Phi X}$ and a sparse component $\mathbf{E}$. This is commonly known as the Robust Principle Component Analysis (Robust PCA) problem. Algorithms for the Robust PCA problem usually solve regularized optimization problems with proper convex relaxation to sparsity and rank [32]–[34]. Alternatively, Bayesian approaches model the sparse and low-rank components with appropriate prior structures and employ approximate inference techniques for estimation [35], [36]. In the experiments we adopt two state-of-the-art Robust PCA solvers, namely the Augmented Lagrangian Method (ALM) algorithm proposed in [34] and the Variational Bayesian (VB) algorithm proposed in [36] to implement the preprocessing module. These two algorithms have demonstrated great outlier-removal capabilities, and therefore we expect the preprocessed data from them to be clean enough for the subsequent MMV signal recovery module.

The preprocessed measurement $\mathbf{Y}_{\mathrm{prep}}$, that is $\mathbf{Y}-\mathbf{E}$, is then fed into an MMV signal recovery module that expects outlier-free input. Candidate implementations of this signal recovery module include those algorithms introduced in Section I, e.g., [16], [17], and [26].

Note that we use the state-of-the-art algorithms to implement both the preprocessing module and the MMV signal recovery module. In this way, we expect the performance of the overall system to represent the best obtainable from sequential approaches. Since there is no prior solution to the Robust MMV problem, we use the sequential approach (with state-of-the-art algorithms implementing its modules) as a benchmark for performance evaluation. However, we should point out that, due to the structural simplicity, there is little collaboration between the two cascaded modules, which potentially limits the performance of the sequential approaches. In the next section, we will present a simultaneous approach, in which outlier-removal and signal recovery are performed jointly in an iterative fashion. As will be seen with numerical examples, the simultaneous approach can have better performance than its sequential counterparts.

III. Simultaneous Outlier Removal and Signal Recovery for Robust MMV

In this section we present the development of a simultaneous approach that iteratively removes outliers and recovers the underlying signals. To reiterate, our goal is to find $(\mathbf{X},\mathbf{E})$ that jointly satisfy the following criteria:

1) $\mathbf{X}$ and $\mathbf{E}$ conform well with the measurement $\mathbf{Y}$, i.e., $\|\mathbf{Y}-\mathbf{\Phi X}-\mathbf{E}\|_{\mathrm{F}}^2$ is small, where $\|\cdot\|_{\mathrm{F}}$ denotes the matrix Frobenius norm.
2) Most of the rows in $\mathbf{X}$ are zeros.
3) The nonzero rows in $\mathbf{X}$ are smooth signals.
4) $\mathbf{E}$ is sparse, i.e., most of its entries are zeros.

A. Regularized Fitting Problem

In order to find $(\mathbf{X},\mathbf{E})$ that jointly satisfy the criteria above, we first formulate a regularized fitting problem, where the regularization terms enforces the desired properties of $\mathbf{X}$ and $\mathbf{E}$ set forth above. Specifically, consider the following unconstrained optimization problem

$$\min_{\mathbf{X},\mathbf{E}} \left\{ \frac{1}{2}\|\mathbf{Y}-\mathbf{\Phi X}-\mathbf{E}\|_{\mathrm{F}}^2 + \lambda_1\|\mathbf{X}\|_{1,\mathbf{P}} + \lambda_2\|\mathbf{E}\|_1 \right\}, \tag{5}$$

where

$$\|\mathbf{E}\|_1 = \sum_{i=1}^{M}\sum_{j=1}^{T}|E_{ij}| \tag{6}$$

is a convex relaxation of the $\|\mathbf{E}\|_0$-(pseudo)norm that counts the number of nonzeros in $\mathbf{E}$, and λ_1 and λ_2 are positive regularization parameters. Via relaxation, minimizing $\|\mathbf{E}\|_1$ promotes the sparsity of $\mathbf{E}$.

The (pseudo)norm $\|\mathbf{X}\|_{1,\mathbf{P}}$ is defined as

$$\|\mathbf{X}\|_{1,\mathbf{P}} = \sum_{i=1}^{N} \|\mathbf{x}_{i\cdot}\|_{\mathbf{P}} = \sum_{i=1}^{N} \left(\mathbf{x}_{i\cdot}\mathbf{P}\mathbf{x}_{i\cdot}^{\mathrm{T}}\right)^{1/2}, \tag{7}$$

where $\mathbf{P} \in \mathbb{R}^{T\times T}$ is a symmetric positive semidefinite matrix that models the prior information about the intra-row structure of $\mathbf{x}_{i\cdot}$. As a concrete example, to enforce smoothness of $\mathbf{x}_{i\cdot}$, $\mathbf{P}$ can be constructed as

$$\mathbf{P} = \mathbf{D}^{\mathrm{T}}\mathbf{D}, \tag{8}$$

where $\mathbf{D} \in \mathbb{R}^{T\times T}$ with

$$D_{ij} = \begin{cases} -2, & \text{if } i = j \\ 1, & \text{if } |i-j| = 1 \\ 0, & \text{else} \end{cases} \tag{9}$$

implements a second-order difference operator. With $\mathbf{P}$ defined as above, the $\mathbf{P}$-weighted norm

$$\|\mathbf{x}_{i\cdot}\|_{\mathbf{P}} = \left(\mathbf{x}_{i\cdot}\mathbf{P}\mathbf{x}_{i\cdot}^{\mathrm{T}}\right)^{1/2} \tag{10}$$

extracts the high-frequency components of $\mathbf{x}_{i\cdot}$, the minimization of which results in smoothly varying $\mathbf{x}_{i\cdot}$.

Note according to (7), the norm $\|\mathbf{X}\|_{1,\mathbf{P}}$ can be viewed as the ℓ_1-norm of the vector containing the $\mathbf{P}$-weighted norms of the rows of $\mathbf{X}$, i.e.,

$$\|\mathbf{X}\|_{1,\mathbf{P}} = \|\left[\|\mathbf{x}_{1\cdot}\|_{\mathbf{P}}, \ldots, \|\mathbf{x}_{N\cdot}\|_{\mathbf{P}}\right]\|_1. \tag{11}$$

Clearly this is a generalization of the matrix $\ell_1\ell_2$-norm defined as

$$\|\mathbf{X}\|_{1,2} = \sum_{i=1}^{N} \left(\mathbf{x}_{i\cdot}\mathbf{x}_{i\cdot}^{\mathrm{T}}\right)^{1/2} = \|\left[\|\mathbf{x}_{1\cdot}\|_2, \ldots, \|\mathbf{x}_{N\cdot}\|_2\right]\|_1, \tag{12}$$

which has been used in MMV problems for modeling row-wise sparsity of $\mathbf{X}$. Minimizing $\|\mathbf{X}\|_{1,2}$ leads to $\mathbf{X}$ that has many all-zero rows, but has little influence on the smoothness of the nonzero rows. As a generalization, the norm $\|\mathbf{X}\|_{1,\mathbf{P}}$ inherits from the $\ell_1\ell_2$-norm the capability of enforcing row-wise sparsity, and meanwhile promotes intra-row smoothness when $\mathbf{P}$ is defined as in (8).

Having introduced the regularized optimization in (5), we provide a few comments on its connection with related research fields.

- First of all, we note that the specific form of $\mathbf{P}$ defined above is known as a Laplacian matrix in signal processing, where it is commonly used to extract high-frequency signal features such as image edges. Since we intend to impose general prior knowledge about signal smoothness rather than being specific about the type of signal, we believe the use of a generic Laplacian high-pass operator is appropriate.
- In addition, the weighted-norm regularized minimization in (5) is formally equivalent to the Maximum *A Posteriori* (MAP) inference where $\mathbf{P}$ is the inverse of the prior covariance matrix. As such a result, if it is desirable to restrict the recovered signals to a specific type (rather than letting it be smooth in general), it is possible to use a $\mathbf{P}$ matrix specific for that type. Note that both the formulation in (5) and the algorithmic solutions presented below are valid for any positive semidefinite $\mathbf{P}$.
- Last but not least, note that we use a fixed form of $\mathbf{P}$ in our algorithm, i.e., we apply a classic data-independent high-pass filter to the signal $\mathbf{x}_{i\cdot}$. It is also possible to use a variable $\mathbf{P}$, whose value is updated in an iterative manner. For an example where a variable weighting is used in a similar context, we refer the readers to [26]–[28]. Both options have their respective pros and cons. Using a fixed $\mathbf{P}$ has computational advantage over iteratively estimating a variable $\mathbf{P}$. As is shown with numerical examples in Section IV, the fixed $\mathbf{P}$ is sufficient to yield satisfactory performance in our case.

From the discussion above, we see that by minimizing the cost function in (5) we find $(\mathbf{X}, \mathbf{E})$ that jointly satisfy the criteria set forth at the beginning of this section. In what follows we will present an efficient algorithm for solving this optimization problem.

B. Problem Re-Formulation and Augmented Lagrangian

Note that (5) is a convex optimization problem in $(\mathbf{X}, \mathbf{E})$. The proof is given in Appendix A. For convex problem, any local optimum is also a global optimum.

Since (5) is a convex problem, in principle we can apply generic algorithms such as subgradient method and interior point method, to find a global minimum. However, the presence of $\mathbf{X}$ in both the quadratic fitting term and $\|\mathbf{X}\|_{1,\mathbf{P}}$ makes finding the subgradient challenging. In order to resolve this issue, we can decouple the first two terms in the cost function by introducing an auxiliary primal variable $\mathbf{U} \in \mathbb{R}^{N\times T}$ and the associated constraint as follows

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{E},\mathbf{U}} \left\{\frac{1}{2}\|\mathbf{Y} - \mathbf{\Phi}\mathbf{X} - \mathbf{E}\|_{\mathrm{F}}^2 + \lambda_1\|\mathbf{U}\|_{1,2} + \lambda_2\|\mathbf{E}\|_1\right\} \\ &\text{subject to}: \mathbf{U} - \mathbf{X}\mathbf{Q} = \mathbf{0}_{N\times T}, \end{aligned} \tag{13}$$

where $\mathbf{Q}$ is the positive-semidefinite square root of $\mathbf{P}$. Since $\mathbf{P}$ is symmetric and positive-semidefinite by definition, $\mathbf{Q}$ is both unique and symmetric, i.e., $\mathbf{Q} = \mathbf{Q}^{\mathrm{T}}$ [37].

It is clear by introducing the equality constraint, (13) is equivalent to (5), and hence by solving (13) we are guaranteed to find a solution to (5). In addition, this decoupling procedure leads to a more tractable algorithm, as will be explained shortly.

The problem in (13) can be solved using the iterative Dual Ascent method by forming the Lagrangian. In order to improve the robustness of the Dual Ascent method and alleviate the condition for convergence, we make the cost function in (13) "more quadratic" by *augmenting* it with a penalty term as follows

$$\begin{aligned} &\min_{\mathbf{X},\mathbf{E},\mathbf{U}} \left\{\frac{1}{2}\|\mathbf{Y} - \mathbf{\Phi}\mathbf{X} - \mathbf{E}\|_{\mathrm{F}}^2 + \lambda_1\|\mathbf{U}\|_{1,2} + \lambda_2\|\mathbf{E}\|_1 \right. \\ &\qquad \left. + \frac{\rho}{2}\|\mathbf{U} - \mathbf{X}\mathbf{Q}\|_{\mathrm{F}}^2\right\} \\ &\text{subject to}: \mathbf{U} - \mathbf{X}\mathbf{Q} = \mathbf{0}_{N\times T}, \end{aligned} \tag{14}$$

where $\rho \geq 0$ is the penalty parameter whose selection will be explained shortly.

The augmented constrained problem in (14) is clearly equivalent to (13), and hence to (5), since any feasible $\mathbf{U}$ and $\mathbf{X}$ will make the augmented quadratic penalty equal to 0.

To solve (14), we associate dual variable $\mathbf{G} \in \mathbb{R}^{N\times T}$ with the constraint. The so-called augmented Lagrangian is formed as

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{X},\mathbf{E},\mathbf{U},\mathbf{G}) &= \frac{1}{2}\|\mathbf{Y}-\mathbf{\Phi X}-\mathbf{E}\|_{\mathrm{F}}^2+\lambda_1\|\mathbf{U}\|_{1,2}+\lambda_2\|\mathbf{E}\|_1 \\ &\quad+\frac{\rho}{2}\|\mathbf{U}-\mathbf{XQ}\|_{\mathrm{F}}^2+\langle\mathbf{G},\mathbf{U}-\mathbf{XQ}\rangle, \end{aligned} \tag{15}$$

where $\langle\mathbf{A},\mathbf{B}\rangle$ denotes the inner product of $\mathbf{A}$ and $\mathbf{B}$ of the same dimensions. Note that the augmented Lagrangian $\mathcal{L}_\rho(\mathbf{X},\mathbf{E},\mathbf{U},\mathbf{G})$ can be viewed as the ordinary Lagrangian $\mathcal{L}_0(\mathbf{X},\mathbf{E},\mathbf{U},\mathbf{G})$ of problem (13) augmented with the quadratic penalty term $\frac{\rho}{2}\|\mathbf{U}-\mathbf{XQ}\|_{\mathrm{F}}^2$.

C. Alternating Direction Method of Multipliers

In this section, we apply the Alternating Direction Method of Multipliers (ADMM) framework to find a solution to (14), and hence to (5) thanks to their equivalence. ADMM is a generic primal-dual procedure that is based on augmented Lagrangian, and has been successfully applied to a wide range of problems. For a tutorial on ADMM, the readers are referred to [38].

ADMM has its root in the traditional Dual Ascent optimization method, and has a few features that make it an increasingly popular convex optimization framework. In order to provide motivation for using ADMM, we put it in the context of the current problem and compare it with the Dual Ascent method. Given the augmented Lagrangian in (15), the Dual Ascent method iterates between a primal minimization step and a dual update step as follows

$$\left(\mathbf{X}^{k+1},\mathbf{E}^{k+1},\mathbf{U}^{k+1}\right)=\arg\min_{\mathbf{X},\mathbf{E},\mathbf{U}}\ \mathcal{L}_\rho(\mathbf{X},\mathbf{E},\mathbf{U},\mathbf{G}^k) \tag{16}$$

$$\mathbf{G}^{k+1}=\mathbf{G}^k+\alpha^k(\mathbf{U}^{k+1}-\mathbf{X}^{k+1}\mathbf{Q}), \tag{17}$$

where k is the iteration index and α^k is a step size. This procedure iteratively seeks a saddle point on the augmented Lagrangian, which corresponds to an optimum solution to (14). Note that the primal minimization step in general involves multiple variables, which itself often needs to be solved in an iterative manner until convergence. In our case, the minimization in (16) can be done using the iterative subgradient method as an inner loop (the outer loop being between (16) and (17)). After the inner loop for subgradient method converges, the outer loop moves to the dual update step in (17). Note that the dual update involves a variable step size α^k, whose choice needs to meet certain criteria in order for the Dual Ascent method to converge.

The ADMM approach is "similar" to the Dual Ascent method, in the sense that it also involves primal minimizations and dual update. However, there are important differences which make ADMM more suitable for solving our problem. Specifically, the ADMM algorithm consists of the following iterative steps

$$\mathbf{E}^{k+1}=\arg\min_{\mathbf{E}}\ \mathcal{L}_\rho(\mathbf{X}^k,\mathbf{E},\mathbf{U}^k,\mathbf{G}^k) \tag{18}$$

$$\mathbf{X}^{k+1}=\arg\min_{\mathbf{X}}\ \mathcal{L}_\rho(\mathbf{X},\mathbf{E}^{k+1},\mathbf{U}^k,\mathbf{G}^k) \tag{19}$$

$$\mathbf{U}^{k+1}=\arg\min_{\mathbf{U}}\ \mathcal{L}_\rho(\mathbf{X}^{k+1},\mathbf{E}^{k+1},\mathbf{U},\mathbf{G}^k) \tag{20}$$

$$\mathbf{G}^{k+1}=\mathbf{G}^k+\rho(\mathbf{U}^{k+1}-\mathbf{X}^{k+1}\mathbf{Q}). \tag{21}$$

Two notable differences between the Dual Ascent iterations and the ADMM iterations are observed. First of all, the joint minimization with respect to all primal variables $(\mathbf{X},\mathbf{E},\mathbf{U})$ in (16) is replaced by the sequential or alternating minimizations from (18) to (20) carried with respect to each primal variable while the others are held constant. One major advantage of the alternating minimizations is that they are much more tractable than the joint minimization. Also note that in ADMM, one pass from (18) to (20) is carried out in each iteration, instead of multiple passes until convergence as would be required for the inner loop of the Dual Ascent method. The second difference between these two set of iterations is that α^k in (17) is replaced by (and fixed at) ρ in (21), which eliminates the need of choosing the step size at each iteration. We will present later that, an optional update on ρ can be carried out at the end of each ADMM iteration to expedite the convergence.

In the following we present solutions to the alternating minimizations from (18) to (20).

1) Update of $\mathbf{E}$*:* To solve (18), denote $\mathbf{F}=\mathbf{Y}-\mathbf{\Phi X}^k$, and it follows from (15) that

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{X}^k,\mathbf{E},\mathbf{U}^k,\mathbf{G}^k) &= \frac{1}{2}\|\mathbf{Y}-\mathbf{\Phi X}^k-\mathbf{E}\|_{\mathrm{F}}^2+\lambda_2\|\mathbf{E}\|_1+C \\ &= \frac{1}{2}\|\mathbf{E}-\mathbf{F}\|_{\mathrm{F}}^2+\lambda_2\|\mathbf{E}\|_1+C \\ &= \sum_{i=1}^{M}\sum_{j=1}^{T}\left\{\frac{1}{2}(E_{ij}-F_{ij})^2+\lambda_2|E_{ij}|\right\}+C, \end{aligned} \tag{22}$$

where C is a constant that is independent of the variable of current interest, that is, $\mathbf{E}$ in this case. Since $\mathcal{L}_\rho(\mathbf{X}^k,\mathbf{E},\mathbf{U}^k,\mathbf{G}^k)$ is separable as shown in (22), the problem in (18) reduces to MT minimizations in $\{E_{ij}\}$. The solution to each of the MT minimizations is given by the soft-thresholding operator as follows

$$E_{ij}=\begin{cases}F_{ij}-\lambda_2, & \text{if } F_{ij}>\lambda_2\\ F_{ij}+\lambda_2, & \text{if } F_{ij}<-\lambda_2\\ 0, & \text{else}\end{cases}. \tag{23}$$

2) Update of $\mathbf{X}$*:* To solve (19), we first note that $\mathcal{L}_\rho(\mathbf{X},\mathbf{E}^{k+1},\mathbf{U}^k,\mathbf{G}^k)$, as the sum of quadratic convex functions in the rows and columns of $\mathbf{X}$ plus a linear function in $\mathbf{X}$, is quadratic and convex in $\mathbf{X}$. Therefore, a solution can be found by setting the gradient of $\mathcal{L}_\rho(\mathbf{X},\mathbf{E}^{k+1},\mathbf{U}^k,\mathbf{G}^k)$ with respect to $\mathbf{X}$ to zero. Regarding the implementation, there are two options, which we present below.

Utilizing matrix algebra, we see from (15) that

$$\begin{aligned}&\mathcal{L}_\rho(\mathbf{X},\mathbf{E}^{k+1},\mathbf{U}^k,\mathbf{G}^k)\\ &\quad=\frac{1}{2}\mathrm{tr}\left(\mathbf{X}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi X}-2(\mathbf{Y}-\mathbf{E}^{k+1})^{\mathrm{T}}\mathbf{\Phi X}\right)\\ &\qquad-\mathrm{tr}\left((\mathbf{G}^k)^{\mathrm{T}}\mathbf{XQ}\right)+\frac{\rho}{2}\mathrm{tr}\left(\mathbf{XPX}^{\mathrm{T}}-2\mathbf{QX}^{\mathrm{T}}\mathbf{U}^k\right)+C,\end{aligned} \tag{24}$$

where constant independent of $\mathbf{X}$ has been absorbed in C. Setting the gradient of (24) with respect to $\mathbf{X}$ to $\mathbf{0}_{N\times T}$, we have

$$\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{X}+\rho\mathbf{X}\mathbf{P}=\mathbf{\Phi}^{\mathrm{T}}(\mathbf{Y}-\mathbf{E}^{k+1})+\mathbf{G}^{k}\mathbf{Q}+\rho\mathbf{U}^{k}\mathbf{Q}, \quad (25)$$

which is known as a Sylvester equation in the control theory. The classical method for numerically solving the Sylvester equation is the Bartels-Stewart algorithm [39], whose computational complexity is $\mathcal{O}(N^3)$ for the general complex-valued case. Based on the Bartels-Stewart algorithm, there exist improved Sylvester equation solvers with reduced computational complexity and memory requirement. For instance, [40] presents numerical approaches that solve real-valued Sylvester equations in $\mathcal{O}(N^2)$ time.

As an alternative to solving the matrix equation in (25), we can also work on the vectorized version. Specifically, let $\mathbf{y}$, $\mathbf{x}$, $\mathbf{e}$, and $\mathbf{u}$ be the vectorized versions of $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{E}$, and $\mathbf{U}$, respectively. Also define $\mathbf{s}=\mathrm{vec}(\mathbf{G}^{k}\mathbf{Q}^{\mathrm{T}})$, $\mathbf{R}=\mathbf{Q}\otimes\mathbf{I}_N$, and $\mathbf{\Psi}=\mathbf{I}_T\otimes\Phi$. With the above definitions, it follows that

$$\begin{aligned}&\mathcal{L}_{\rho}(\mathbf{X},\mathbf{E},\mathbf{U},\mathbf{G})\\&\quad=\frac{1}{2}\|\mathbf{y}-\mathbf{e}-\mathbf{\Psi}\mathbf{x}\|_2^2+\frac{\rho}{2}\|\mathbf{u}-\mathbf{R}\mathbf{x}\|_2^2-\mathbf{s}^{\mathrm{T}}\mathbf{x}+C, \quad (26)\end{aligned}$$

where C is a constant independent of $\mathbf{X}$ (or $\mathbf{x}$). Setting the gradient of (26) with respect to $\mathbf{x}$ to zero leads to the following system of equations

$$\left(\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi}+\rho\mathbf{R}^{\mathrm{T}}\mathbf{R}\right)\mathbf{x}=\mathbf{\Psi}^{\mathrm{T}}(\mathbf{y}-\mathbf{e})+\mathbf{s}+\rho\mathbf{R}^{\mathrm{T}}\mathbf{u}. \quad (27)$$

Since $\mathbf{\Psi}^{\mathrm{T}}\mathbf{\Psi}+\rho\mathbf{R}^{\mathrm{T}}\mathbf{R}$ does not change over iterations, we can compute it off-line and cache its inverse, such that solving the system of equations in (27) can be done in $\mathcal{O}(N^2T^2)$.

We have experimentally compared the computational complexities of solving the matrix-version and the vector-version of $\mathbf{X}$ update. For the matrix version, the MATLAB implementation of the classical Bartels-Stewart algorithm (i.e., *lyap.m*) was used thanks to its tested stability and availability. For the vector version, we have programmed the update ourselves. Empirical evidence suggests that solving the matrix-version is computationally preferable than solving the equivalent vector version. Additionally, note that for applications where computational complexity or memory consumption is a concern, more efficient approaches, e.g., [40], are available.

3) Update of $\mathbf{U}$*:* To find a solution to (20), denote $\mathbf{V}=\mathbf{X}^{k+1}\mathbf{Q}-\rho^{-1}\mathbf{G}^{k}$, and it follows that

$$\begin{aligned}&\mathcal{L}_{\rho}(\mathbf{X}^{k+1},\mathbf{E}^{k+1},\mathbf{U},\mathbf{G}^{k})\\&\quad=\lambda_1\|\mathbf{U}\|_{1,2}+\langle\mathbf{G}^{k},\mathbf{U}\rangle+\frac{\rho}{2}\|\mathbf{U}-\mathbf{X}^{k+1}\mathbf{Q}\|_{\mathrm{F}}^2+C\\&\quad=\lambda_1\|\mathbf{U}\|_{1,2}+\frac{\rho}{2}\|\mathbf{U}-\left(\mathbf{X}^{k+1}\mathbf{Q}-\rho^{-1}\mathbf{G}^{k}\right)\|_{\mathrm{F}}^2+C\\&\quad=\lambda_1\|\mathbf{U}\|_{1,2}+\frac{\rho}{2}\|\mathbf{U}-\mathbf{V}\|_{\mathrm{F}}^2+C\\&\quad=\sum_{i=1}^{N}\left\{\lambda_1\|\mathbf{u}_{i\cdot}\|_2+\frac{\rho}{2}\|\mathbf{u}_{i\cdot}-\mathbf{v}_{i\cdot}\|_2^2\right\}+C. \quad (28)\end{aligned}$$

From (28) it is clear that the minimization in (20) is separable over $\{\mathbf{u}_{i\cdot}\}_{i=1}^{N}$. For each $\mathbf{u}_{i\cdot}$, it is clear that

$$f(\mathbf{u}_{i\cdot})=\lambda_1\|\mathbf{u}_{i\cdot}\|_2+\frac{\rho}{2}\|\mathbf{u}_{i\cdot}-\mathbf{v}_{i\cdot}\|_2^2 \quad (29)$$

is convex in $\mathbf{u}_{i\cdot}$ because it is the sum of its ℓ_2-norm and a quadratic with positive definite Hessian.

To determine the minimum value of $f(\mathbf{u}_{i\cdot})$, two cases need to be considered:

1) If $\mathbf{u}_{i\cdot}=\mathbf{0}_{1\times T}$, $f(\mathbf{u}_{i\cdot})=\rho\|\mathbf{v}_{i\cdot}\|_2^2/2$.
2) If $\mathbf{u}_{i\cdot}\neq\mathbf{0}_{1\times T}$, $f(\mathbf{u}_{i\cdot})$ is convex and smooth in $\mathbf{u}_{i\cdot}$. Setting the gradient of $f(\mathbf{u}_{i\cdot})$ with respect to $\mathbf{u}_{i\cdot}$ to zero, we have

$$\nabla_{\mathbf{u}_{i\cdot}}f(\mathbf{u}_{i\cdot})=\frac{\lambda_1\mathbf{u}_{i\cdot}}{\|\mathbf{u}_{i\cdot}\|_2}+\rho(\mathbf{u}_{i\cdot}-\mathbf{v}_{i\cdot})=\mathbf{0}_{1\times T}, \quad (30)$$

from which it follows that $\|\mathbf{u}_{i\cdot}\|_2=\|\mathbf{v}_{i\cdot}\|_2-\lambda_1/\rho$ and

$$\mathbf{u}_{i\cdot}=\left(1-\frac{\lambda_1}{\rho\|\mathbf{v}_{i\cdot}\|_2}\right)\mathbf{v}_{i\cdot}. \quad (31)$$

The corresponding value of $f(\mathbf{u}_{i\cdot})$ in this case is $f(\mathbf{u}_{i\cdot})=\lambda_1\|\mathbf{v}_{i\cdot}\|_2-\lambda_1^2/(2\rho)$.

Comparing the minimum values of $f(\mathbf{u}_{i\cdot})$ in the two cases above, we can determine the optimal $\mathbf{u}_{i\cdot}$ as follows

$$\mathbf{u}_{i\cdot}=\begin{cases}\left(1-\dfrac{\lambda_1}{\rho\|\mathbf{v}_{i\cdot}\|_2}\right)\mathbf{v}_{i\cdot}, & \text{if } \|\mathbf{v}_{i\cdot}\|_2\geq\frac{\lambda_1}{\rho}\\ \mathbf{0}_{1\times T}, & \text{else}\end{cases}. \quad (32)$$

From (32) we see that the update rule for $\mathbf{U}$ is simply soft-thresholding applied on its rows. This is analogous to the soft-thresholding applied on the entries of $\mathbf{E}$, as is shown in (23).

D. Convergence of ADMM and Optimality Conditions

According to [38], there are two conditions sufficient for the convergence of the ADMM iterations presented above. The first condition requires the cost function in (13) to be closed, proper, and convex, which is clearly satisfied. The second condition requires that the ordinary Lagrangian $\mathcal{L}_0(\mathbf{X},\mathbf{E},\mathbf{U},\mathbf{G})$ of the problem in (13) has a saddle point. Since (13) is a convex optimization problem and there exists at least one feasible solution, it follows that strong duality holds for (13). This implies the existence of a saddle point of $\mathcal{L}_0(\mathbf{X},\mathbf{E},\mathbf{U},\mathbf{G})$, and consequently, the convergence of the ADMM iterations.

Now we investigate the optimality conditions of the problem in (13). For $(\mathbf{X}^{\star},\mathbf{E}^{\star},\mathbf{U}^{\star},\mathbf{G}^{\star})$ to be optimal, they have to satisfy the following necessary conditions

1) Feasibility, i.e.,

$$\mathbf{0}_{N\times T}=\mathbf{U}^{\star}-\mathbf{X}^{\star}\mathbf{Q}. \quad (33)$$

2) Gradients (Subgradients) with respect to primal variables vanish (include the origin), i.e.,

$$\mathbf{0}_{N\times T}\in\partial\lambda_1\|\mathbf{U}^{\star}\|_{1,2}+\mathbf{G}^{\star} \quad (34)$$

$$\mathbf{0}_{N\times T}=\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{X}^{\star}-\mathbf{\Phi}^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{E}^{\star}\right)-\mathbf{G}^{\star}\mathbf{Q}^{\mathrm{T}} \quad (35)$$

$$\mathbf{0}_{N\times T}\in\partial\lambda_2\|\mathbf{E}^{\star}\|_1+\mathbf{E}^{\star}-\left(\mathbf{Y}-\mathbf{\Phi}\mathbf{X}^{\star}\right). \quad (36)$$

At each iteration, since $\mathbf{U}^{k+1}$ minimizes $\mathcal{L}_{\rho}(\mathbf{X}^{k+1},\mathbf{E}^{k+1},\mathbf{U},\mathbf{G}^{k})$, it follows that the corresponding subgradient includes the origin, i.e.,

$$\begin{aligned}\mathbf{0}_{N\times T}&\in\partial\lambda_1\|\mathbf{U}^{k+1}\|_{1,2}+\mathbf{G}^{k}+\rho\left(\mathbf{U}^{k+1}-\mathbf{X}^{k+1}\mathbf{Q}\right)\\&=\partial\lambda_1\|\mathbf{U}^{k+1}\|_{1,2}+\mathbf{G}^{k+1}. \quad (37)\end{aligned}$$

Therefore, the condition in (34) is always satisfied at the end of each iteration.

Since $\mathbf{X}^{k+1}$ minimizes $\mathcal{L}_\rho(\mathbf{X}, \mathbf{E}^{k+1}, \mathbf{U}^k, \mathbf{G}^k)$, it follows that the corresponding gradient vanishes, i.e.,

$$\begin{aligned}\mathbf{0}_{N\times T} &= \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{X}^{k+1} - \mathbf{\Phi}^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{E}^{k+1}\right) - \mathbf{G}^k\mathbf{Q}^{\mathrm{T}} \\ &\quad + \rho\mathbf{X}^{k+1}\mathbf{P} - \rho\mathbf{U}^k\mathbf{Q} \\ &= \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{X}^{k+1} - \mathbf{\Phi}^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{E}^{k+1}\right) - \mathbf{G}^{k+1}\mathbf{Q}^{\mathrm{T}} \\ &\quad + \rho\left(\mathbf{U}^{k+1} - \mathbf{U}^k\right)\mathbf{Q}. \end{aligned} \tag{38}$$

From (38) we see that the last term $\rho(\mathbf{U}^{k+1} - \mathbf{U}^k)\mathbf{Q}$ denotes the quantity by which the optimality condition (35) is violated, and therefore it can be viewed as a residual.

Denote by

$$\mathbf{D}_{\mathrm{p}}^{k+1} = \mathbf{U}^{k+1} - \mathbf{X}^{k+1}\mathbf{Q} \tag{39}$$

and

$$\mathbf{D}_{\mathrm{d}}^{k+1} = \rho\left(\mathbf{U}^{k+1} - \mathbf{U}^k\right)\mathbf{Q} \tag{40}$$

the primal residual and dual residual, respectively. As the iterations proceed, both the primal and dual residuals approach zero, and the ADMM algorithm is guaranteed to converge for the convex optimization problem as in (13) [38].

Although the ADMM iterations converge for any fixed penalty parameter $\rho \geq 0$, it is in practice possible to adjust the value of ρ along the iterations, with the goal of improving the convergence and making the performance less dependent on the initial choice of ρ. An intuitive and popular strategy for adjusting ρ is presented as follows [41].

On one hand, note that the update rule (21) suggests that a larger ρ will enforce the feasibility $\mathbf{U} = \mathbf{X}\mathbf{Q}$ more strongly, and hence produce a smaller primal residual $\mathbf{D}_{\mathrm{p}}$. On the other hand, the definition of $\mathbf{D}_{\mathrm{d}}$ in (40) suggests that a smaller ρ will yield smaller dual residual. Since for convergence we need both primal and dual residuals to be small, it makes sense that the value of ρ be adjusted along the iterations. Specifically, the following update rule is commonly used

$$\rho^{k+1} = \begin{cases} \beta\rho^k, & \text{if } \|\mathbf{D}_{\mathrm{p}}\|_{\mathrm{F}}/\|\mathbf{D}_{\mathrm{d}}\|_{\mathrm{F}} > \tau \\ \beta^{-1}\rho^k, & \text{if } \|\mathbf{D}_{\mathrm{p}}\|_{\mathrm{F}}/\|\mathbf{D}_{\mathrm{d}}\|_{\mathrm{F}} < \tau^{-1} \\ \rho^k, & \text{else} \end{cases}, \tag{41}$$

where $\beta > 1$ and $\tau > 1$ are parameters. In the numerical examples below, we use $\beta = 2$ and $\tau = 10$, which are typical values.

The ADMM algorithm for solving the constrained problem in (13) is summarized in Algorithm 1.

Algorithm 1: ADMM Solver for the Regularized Fitting Problem in (13)

1: Inputs: $\mathbf{Y} \in \mathbb{R}^{M\times T}$, $\mathbf{\Phi} \in \mathbb{R}^{M\times N}$, $\lambda_1 > 0$, $\lambda_2 > 0$, (optional) β, (optional) τ
2: Outputs: $\mathbf{X} \in \mathbb{R}^{N\times T}$, $\mathbf{E} \in \mathbb{R}^{M\times T}$
3: Initialize: $\mathbf{X}^0 = \mathbf{U}^0 = \mathbf{G}^0 = \mathbf{0}_{N\times T}$, $\mathbf{E}^0 = \mathbf{0}_{M\times T}$, $\rho^0 = 1$, $k = 0$
4: **while** not converged **do**
5: Update $\mathbf{E}^{k+1}$ using (23)
6: Update $\mathbf{X}^{k+1}$ by solving either (25) or (27)
7: Update $\mathbf{U}^{k+1}$ using (32)
8: Update $\mathbf{G}^{k+1}$ using (21)
9: (Optional) Update ρ^{k+1} using (41)
10: $k \leftarrow k + 1$
11: **end while**
12: Set $\mathbf{X} = \mathbf{X}^k$ and $\mathbf{E} = \mathbf{E}^k$.

IV. Numerical Examples

In this section we demonstrate the performance of the proposed algorithm with experimental results. Specifically, we consider two scenarios: outlier-free case and outlier-present case. In both cases, the performance of the proposed algorithm is validated and compared with that obtained from state-of-the-art approaches. For all algorithms considered in this paper, their parameters were empirically tuned to yield the best results. Specifically, we have adopted a greedy search procedure for parameter tuning, which is implemented as a coarse logarithmic search followed by a refined linear search. The empirically optimal parameters selected via this search were used to generate the results reported herein. Based on our experiments, the various algorithms considered in the numerical examples are robust to the selection of the parameters.

A. Outlier-Free Case

When outliers $\mathbf{E}$ are not present, the measurement model in (3) reduces to the conventional Multiple Measurement Vector model in (2). Specifically, in this experiment, we considered problems with measurement size M varying from 30 to 100 in the increment of 5, while the dimensions of the latent signal $\mathbf{X}$ were fixed at $N = T = 200$. Denote by K the number of nonzero rows in $\mathbf{X}$. Two sparsity levels of $\mathbf{X}$ were considered, i.e., either $K = 5$ or $K = 10$. Each nonzero row of $\mathbf{X}$ was generated as a Hanning-window tapered sinusoid, where the number of periods was uniformly drawn among 1, 2, and 3, and the phase was uniformly distributed between 0 and π. The transformation $\mathbf{\Phi}$ was generated according to a uniform spherical ensemble, i.e., each $\boldsymbol{\phi}_{\cdot i}$ was independently drawn from a uniform distribution on the M-sphere with radius 1. Independent and identically distributed Gaussian noise $\mathbf{N}$ was added to the measurement, resulting in an SNR at 10 dB.

The proposed algorithm, termed "ADMM" herein, can be adapted in a straightforward manner to this model by setting $\lambda_2 = \infty$ (or a large value such as 10^{16} in practice), which enforces $\mathbf{E} = \mathbf{0}_{M\times T}$ in the result. To model the smoothness of the nonzero rows of $\mathbf{X}$, we used $\mathbf{P} = \mathbf{D}^{\mathrm{T}}\mathbf{D}$ with $\mathbf{D}$ defined as in (9).

For comparison, we also include in this case the following algorithms: MFOCUSS [16] (with $p = 0.8$), MBP [17], MSBL [19], and TMSBL [26]. Among these existing algorithms, TMSBL is the only one that takes temporal correlation into account to the best of our knowledge, while MFOCUSS is usually reported to yield the best recovery performance among the techniques that do not utilize temporal correlation. For MFOCUSS and MBP, implementations from the Multiple-Spars Toolbox [42] were used, and for MSBL and TMSBL the implementations obtained from the authors' website was used.

Denote by $\mathbf{X}$ and $\hat{\mathbf{X}}$ the ground truth and the reconstructed signal, respectively. The reconstruction error is quantified as

$$\epsilon_X = \|\mathbf{X} - \hat{\mathbf{X}}\|_{\mathrm{F}}^2/\|\mathbf{X}\|_{\mathrm{F}}^2. \tag{42}$$

Results averaged over 100 random runs are shown in Fig. 2, where the lines show the average reconstruction errors while the vertical bars indicate the standard deviations. As we see, the proposed ADMM algorithm outperforms the existing approaches by a margin. By penalizing non-smoothness in the reconstructed

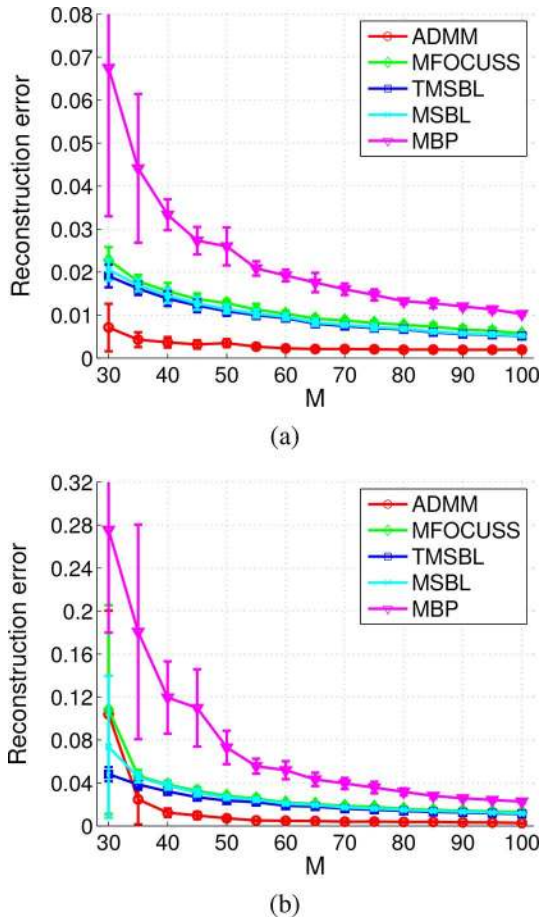


Fig. 2. Reconstruction error versus the number of measurements M in the outlier-free case. (a) Results for 5 nonzero rows in $\mathbf{X}$, (b) results for 10 nonzero rows in $\mathbf{X}$.

signals, the proposed algorithm is less susceptible to the presence of noise.

B. Outlier-Present Case

In this subsection we consider the case where the measurements are contaminated by both outliers and noise. The sequential benchmark approach presented in Section II and the simultaneous approach presented in Section III will be investigated in terms of their capabilities to remove outliers $\mathbf{E}$ and recover the underlying signal $\mathbf{X}$.

As has been explained in Section II, the preprocessing module in the sequential approach solves a Robust PCA problem. To this end, we consider two state-of-the-art Robust PCA solvers, namely, the Augmented Lagrangian Method presented in [34], and the Variational Bayesian method presented in [36]. The output from the ALM or VB preprocessing module is considered outlier-free and therefore can be used as input to the subsequent signal recovery module. For signal recovery, we consider two options: MFOCUSS and TMSBL. These two algorithms yield the best results as is shown in the previous subsection. Moreover, TMSBL is a representative approach that takes into account temporal correlation, while MFOCUSS is a representative that does not take such information into account. Combining the preprocessing and the signal recovery modules, we have four configurations, termed as "ALM_MFOCUSS", "ALM_TMSBL", "VB_MFOCUSS", and "VB_TMSBL", respectively, which we believe are among the best in terms of reconstruction accuracy.

1) Proof of Concept: Before providing details on the numerical analysis of the various approaches to solving the Robust MMV problem, we first present a simple example to demonstrate the need of special algorithms, such as those developed in this paper, in order to recovery signals from outlier-corrupted measurement. In this experiment, we set $M = 100$, $N = 200$, $T = 100$, and set $\mathbf{X}$ to have $K = 3$ nonzero rows generated as random sinusoids tapered with a Hann window. As a proof of concept experiment, we set $\mathbf{E}$ to contain only a single outlier, whose amplitude is several times greater than the signal amplitude. In attempt to recover the underlying signals, we applied the MFOCUSS and TMSBL algorithms (without applying ALM or VB for outlier-removal preprocessing), and compare their results with that obtained from ADMM. The reconstruction error obtained from the three algorithms are as follows: (1) $\epsilon_{\text{ADMM}} = 3.65 \times 10^{-3}$, (2) $\epsilon_{\text{MFOCUSS}} = 1.83 \times 10^{-1}$, (3) $\epsilon_{\text{TMSBL}} = 1.45$. Unsurprisingly, the MFOCUSS and TMSBL algorithms, both of which expect outlier-free measurements as input, failed to recover the underlying signals even for this fairly easy test case. The ADMM algorithm, thanks to its outlier-removal capability, accurately recovered the signals as expected.

2) Analysis on Computational Complexity: It is instructive to examine the computational complexity of the five approaches. We employ the *Big-*$\mathcal{O}$ notation to analyze the growth of computational complexity with the problem dimensions M, N and T.

For each iteration of ADMM, the update of $\mathbf{E}$ takes $\mathcal{O}(MNT)$; the update of $\mathbf{X}$ takes $\mathcal{O}(N^3)$ (assuming (25) is solved); and the update of $\mathbf{U}$ takes $\mathcal{O}(NT^2)$. Therefore, the computational complexity of the ADMM approach is $\mathcal{O}(MNT + N^3 + NT^2)$ per iteration.

For the ALM preprocessing step, the bulk of computation in each iteration is the Singular Value Decomposition (SVD) of an $M \times T$ matrix, which takes $\mathcal{O}(MT^2 + M^2T)$. For the VB preprocessing step, the computation in each iteration is centered on updating the various covariance matrices, which takes $\mathcal{O}(M^3 + T^3)$ in general. For MFOCUSS, each iteration takes $\mathcal{O}(M^2N)$, and for TMSBL, each iteration takes $\mathcal{O}(N^3)$.

The computational complexity of the approaches considered herein is summarized in Table I. Inspecting the overall complexity, we see there is no clear winner among these approaches. Therefore, we consider the "marginal" complexity when only N or T varies (M is smaller than N by definition and hence is not considered). From the last two columns in Table I, we see that the complexity of ADMM is comparable with that of ALM_TMSBL. However, we will see shortly that ADMM has significantly better results in terms of recovery accuracy than its counterparts.

3) Reconstruction Accuracy vs. Problem Scale: Before getting into the details of numerical analysis, we make a clarification on the experimental setup followed herein. Since the Robust MMV problem involves multiple variables, e.g., problem scale, compression ratio M/N, number of signal components K, outlier density $\|\mathbf{E}\|_0$, etc., the problem space grows combinatorially with the number of variables. As such a result, it is infeasible to allow all variables to vary together and explore the entire problem space. In order to meaningfully analyze the algorithmic performance, we have applied the "separation of

TABLE I
COMPARISON OF COMPUTATIONAL COMPLEXITY

Approach	Overall	Varying N	Varying T
ADMM	$\mathcal{O}(MNT + N^3 + NT^2)$	$\mathcal{O}(N^3)$	$\mathcal{O}(T^2)$
ALM_MFOCUSS	$\mathcal{O}(MT^2 + M^2T + M^2N)$	$\mathcal{O}(N)$	$\mathcal{O}(T^2)$
ALM_TMSBL	$\mathcal{O}(MT^2 + M^2T + N^3)$	$\mathcal{O}(N^3)$	$\mathcal{O}(T^2)$
VB_MFOCUSS	$\mathcal{O}(M^3 + T^3 + M^2N)$	$\mathcal{O}(N)$	$\mathcal{O}(T^3)$
VB_TMSBL	$\mathcal{O}(M^3 + T^3 + N^3)$	$\mathcal{O}(N^3)$	$\mathcal{O}(T^3)$

TABLE II
COMPARISON OF RECOVERY ACCURACY FOR VARYING PROBLEM SCALES

N	K	$\|\mathbf{E}\|_0$	SNR (dB)	ϵ_X from ADMM	ϵ_X from ALM_MFOCUSS	ϵ_X from ALM_TMSBL	ϵ_X from VB_MFOCUSS	ϵ_X from VB_TMSBL	Performance Improvement
100	3	250	10	$\mathbf{6.18 \times 10^{-3}}$	1.02×10^{-2}	8.15×10^{-3}	1.54×10^{-2}	1.20×10^{-2}	24.2%
200	5	1000	10	$\mathbf{2.58 \times 10^{-3}}$	7.01×10^{-3}	6.10×10^{-3}	1.02×10^{-2}	9.00×10^{-3}	57.7%
400	10	4000	10	$\mathbf{1.08 \times 10^{-3}}$	5.93×10^{-3}	5.35×10^{-3}	7.57×10^{-3}	6.76×10^{-3}	79.8%
600	15	9000	10	$\mathbf{8.45 \times 10^{-4}}$	5.74×10^{-3}	5.18×10^{-3}	6.47×10^{-3}	5.56×10^{-3}	83.7%
1000	25	25000	10	$\mathbf{7.63 \times 10^{-4}}$	5.63×10^{-3}	5.11×10^{-3}	6.21×10^{-3}	5.48×10^{-3}	85.1%
100	3	250	20	$\mathbf{5.93 \times 10^{-4}}$	9.27×10^{-4}	7.47×10^{-4}	1.17×10^{-3}	1.14×10^{-3}	20.6%
200	5	1000	20	$\mathbf{2.79 \times 10^{-4}}$	7.16×10^{-4}	5.76×10^{-4}	9.24×10^{-4}	9.02×10^{-4}	51.6%
400	10	4000	20	$\mathbf{1.48 \times 10^{-4}}$	5.80×10^{-4}	5.01×10^{-4}	6.95×10^{-4}	6.90×10^{-4}	70.5%
600	15	9000	20	$\mathbf{1.09 \times 10^{-4}}$	6.01×10^{-4}	5.16×10^{-4}	5.62×10^{-4}	5.50×10^{-4}	78.9%
1000	25	25000	20	$\mathbf{9.87 \times 10^{-5}}$	5.91×10^{-4}	5.03×10^{-4}	5.86×10^{-4}	5.65×10^{-4}	80.4%
Overall Performance Improvement									**63.2%**

variable" principle in designing the experiments. Specifically, in each set of experiments, we varied only one primary variable, while either fixing the others or making them dependent on the primary variable. In this way we can clearly analyze the algorithmic performance from different experimental perspectives.

In this experiment, we consider problems of varying scales $N \in \{100, 200, 400, 600, 1000\}$. For each scale, we set $M = N/2$ and $T = N$. For signal $\mathbf{X}$ the number of nonzero rows K increases linearly with N, and the nonzero rows were tapered sinusoids generated in a manner that is similar to the previous experiment. In all cases, we had $K \ll N = T$, rendering $\mathbf{X}$ of low rank, and therefore justifying our use of a Robust PCA approach for preprocessing. For $\mathbf{E}$, the density of the outliers was fixed at $\|\mathbf{E}\|_0 = 5\%MT$, and the nonzeros of $\mathbf{E}$ were independently drawn from the uniform $\mathcal{U}(-10, 10)$ distribution. Gaussian noise $\mathbf{N}$ was added to the measurement yielding SNR at 10 dB and 20 dB, respectively.

Table II summarizes the experimental results for all the test cases introduced above. For each test case, the reconstruction errors averaged over 100 independent runs are shown. The performance improvement is defined as the percentage reduction between the lowest reconstruction error and the second lowest reconstruction error.

Four observations follow from the results in Table II. Firstly, in all test cases the ADMM simultaneous approach outperforms its sequential counterpart by a significant margin. In overall, ADMM reduces reconstruction error by over 63% in comparison with ALM_TMSBL (the second best approach). In certain cases, the error reduction is over 80%.

Secondly, the performance of all the algorithms considered herein generally improves when scale of the problem increases. From the last column of the table, we see that the ADMM approach enjoys more significant performance improvement than the other four approaches. The price to pay for this more rapid performance improvement is computational complexity, which has been examined previously.

Thirdly, if we are interested in estimating K, i.e., the number of nonzero rows in $\mathbf{X}$, we can simply count the number of rows in $\hat{\mathbf{X}}$ that have norms exceeding a threshold. Such a threshold can be set dynamically, for instance, to be a small fraction of the largest row-wise norm in $\hat{\mathbf{X}}$. Experimental results confirm that with this thresholding step all the approaches considered herein are able to correctly identify the nonzero rows.

Finally, we note that the use of ALM for preprocessing yields generally similar, yet slightly better results than the use of VB. Therefore, in the following experiments we use only ALM for preprocessing for the clarity of comparison.

4) Robustness to Lack of Measurements: In this experiment we investigate how the performance of the various algorithms considered herein is affected by the availability of measurements. For this purpose, we fixed the dimensions at $N = T = 200$, and varied M from 30 to 100 in the increment of 5. The number of nonzero rows in $\mathbf{X}$ was fixed at $K = 5$, and the density of outliers in $\mathbf{E}$ was set at $\|\mathbf{E}\|_0 = 5\%MT$. The signal and outliers were generated in a similar manner as above. Gaussian noise was added to yield an SNR at 10 dB.

Fig. 3 plots the results averaged over 100 independent runs with error bars indicating the standard deviations. It is clear from the figure that the ADMM approach consistently outperform the two sequential approaches by a large margin. Averaged across all numbers of measurements, ADMM yields more than 60% performance improvement over ALM_TMSBL, which represents the best performance in sequential approaches. Moreover, by examining the slopes of the curves in Fig. 3, we observe that the performance of ADMM degrades much more gracefully

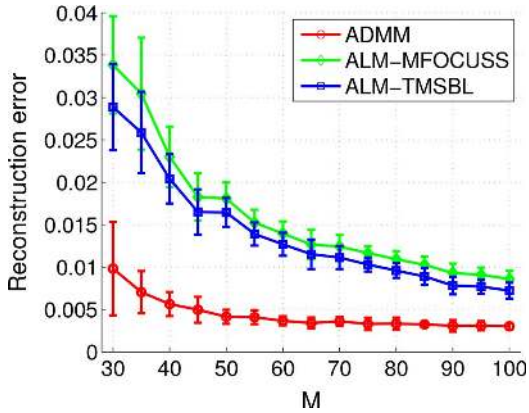


Fig. 3. Reconstruction error versus the number of measurements M when outliers are present. Results for 5 nonzero rows in $\mathbf{X}$ and SNR at 10 dB. Average reduction in reconstruction error between ADMM and ALM_TMSBL is 62.5%.

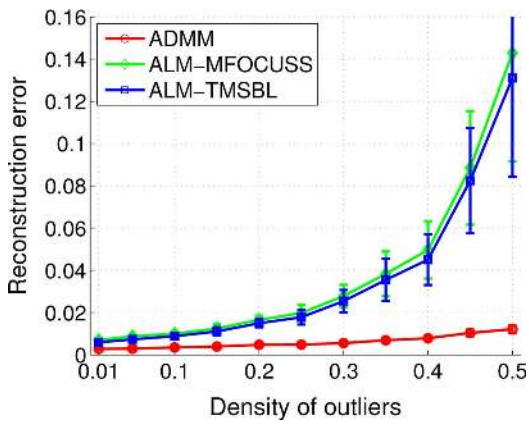


Fig. 4. Reconstruction error versus density of outliers in $\mathbf{E}$. Average reduction in reconstruction error between ADMM and ALM_TMSBL is 72.7%.

than that of the other two approaches when measurements become increasingly scarce.

5) Robustness to Density of Outliers: In this experiment we illustrate the performance of the various algorithms when different levels of outliers are present in the measurement. Specifically, the dimensions of the problem were fixed at $M = 100$, $N = T = 200$, and the number of nonzero rows in $\mathbf{X}$ was fixed at $K = 5$. The density of outliers was varied over a wide range from 1% to 50%. The generation of signals and outliers was similar as above. Gaussian noise was added to the measurement resulting in an SNR at 10 dB.

The effectiveness of the ADMM approach in handling outliers is clearly illustrated in Fig. 4. Note that the ADMM approach consistently outperforms the sequential approaches by a significant margin. Moreover, over the wide range of density levels considered herein, the performance of the ADMM approach barely degrades, while in contrast, the reconstruction error by the sequential approaches has increased by over 20 times. The excellent performance of the ADMM approach is attributed to the joint estimation of $\mathbf{E}$ and $\mathbf{X}$, where iterative refinements are made.

6) Reconstruction Accuracy vs. Number of Nonzero Signals: In this experiment we fix the dimensions of the problem, and examine how the number of temporal signals in $\mathbf{X}$ affects the algorithmic performance. Intuitively, when more signals are present,

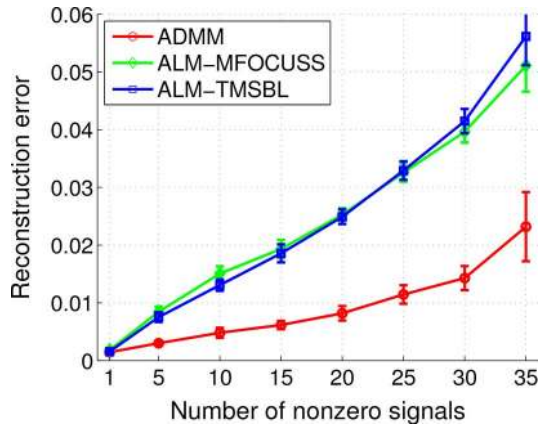


Fig. 5. Reconstruction error versus the number of nonzero signals in $\mathbf{X}$ when 5% of measurements were corrupted by outliers. SNR $= 10$ dB. Average reduction in reconstruction error between ADMM and ALM_TMSBL is 56.7%.

i.e., when $\mathbf{X}$ has more nonzero rows, the recovery problem becomes more difficult. To quantitatively analyze this trend, we set up the experiment as follows. The dimensions of the problem were fixed at $M = 100$, $N = T = 200$, and the density of outliers was fixed at $\|\mathbf{E}\|_0 = 5\%MT$. The number of nonzero rows in $\mathbf{X}$ was selected from $K \in \{1, 5, 10, 15, 20, 25, 30, 35\}$. The generation of signals and outliers was done in a similar fashion as above. Gaussian noise was added to the measurement resulting in an SNR at 10 dB.

The experimental results are plotted in Fig. 5. The curves confirm that the algorithmic performance degrades as the number of signals increases. Despite the similar trend, ADMM yields significantly lower reconstruction error than its sequential counterparts, and the average performance improvement is close to 60%.

7) Robustness to Noise: In this experiment we investigate how noise in the measurement affects the performance of the algorithms. Before making quantitative comparison, we first examine in Figs. 6 and 7 the typical waveforms obtained from the various algorithms, which provide us with more insight into the effect of noise.

In Fig. 6, the nonzero rows of $\mathbf{X}$ consist of 5 tapered sinusoids generated in a similar fashion as above. The parameters of the problem were fixed at $M = 100$, $N = 200$, and $T = 200$, $\|\mathbf{E}\|_0 = 5\%MT$, respectively. Noise was added to yield an SNR at 10 dB. Fig. 6(a) shows a typical realization of the nonzero rows in $\mathbf{X}$. In Figs. 6(b)-(d), we use colored curves to denote the rows of $\hat{\mathbf{X}}$ corresponding to the nonzero rows of $\mathbf{X}$, and use black curves to denote the remaining rows. As is evidenced by the figures, all of the algorithms can accurately identify the nonzero rows of $\mathbf{X}$. In addition, it is clear that by using smoothness-promoting regularizations in (7), the ADMM approach yields more accurate recovery of the signals, while the results of the other approaches contain spurious variations in the recovered signals. The spurious variations are due to over-fitting the noisy measurements. Similar observations follow when we examine Fig. 7, where the nonzero signals were obtained as realizations of EEG waveforms.

In the following, we quantitatively evaluate the robustness of the algorithms to noise. For this purpose, we set $M = 100$, $N = T = 200$, $K = 5$, and $\|\mathbf{E}\|_0 = 5\%MT$. The generation of

TABLE III
COMPARISON OF ALGORITHMIC PERFORMANCE UNDER CHALLENGING TEST CONDITIONS

ID	Experimental Setting	ϵ from ADMM	ϵ from ALM_MFOCUSS	ϵ from ALM_TMSBL	Description
1	$N = 200, M = 20, T = 100, K = 3$	$\mathbf{1.45 \times 10^{-2}}$	6.98×10^{-2}	6.19×10^{-2}	heavy compression
2	$N = 200, M = 100, T = 100, K = 50$	$\mathbf{1.71 \times 10^{-1}}$	2.51×10^{-1}	2.09×10^{-1}	strong nonzero signal presence
3	$N = 200, M = 20, T = 30, K = 3$	$\mathbf{6.75 \times 10^{-2}}$	9.58×10^{-1}	9.31×10^{-1}	heavy compression + short temporal sequence
4	$N = 200, M = 20, T = 100, K = 10$	$\mathbf{8.06 \times 10^{-1}}$	9.22×10^{-1}	1.03	combination of 1 and 2

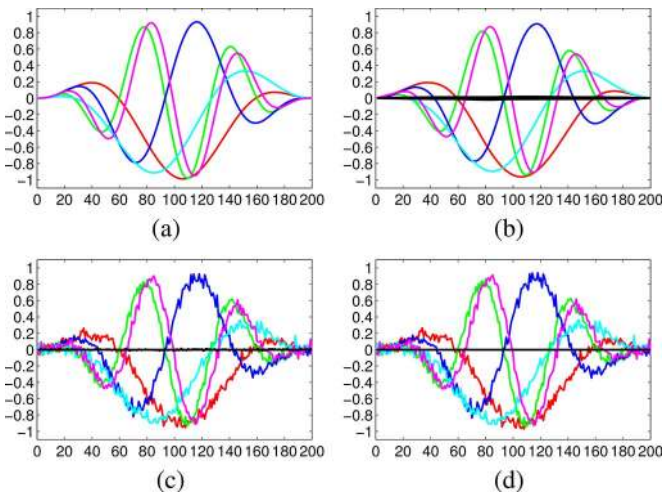


Fig. 6. Reconstruction of smooth signals (sinusoidal waveforms). (a) Original signals, (b) reconstructed from ADMM ($\epsilon_X = 4.18 \times 10^{-3}$), (c) reconstructed from ALM_MFOCUSS ($\epsilon_X = 1.37 \times 10^{-2}$), (d) reconstructed from ALM_TMSBL ($\epsilon_X = 1.24 \times 10^{-2}$).

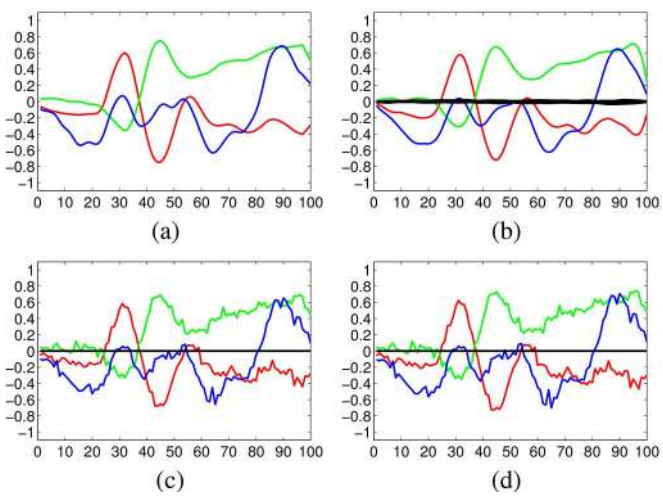


Fig. 7. Reconstruction of smooth signals (EEG waveforms). (a) Original signals, (b) reconstructed from ADMM ($\epsilon_X = 1.23 \times 10^{-2}$), (c) reconstructed from ALM_MFOCUSS ($\epsilon_X = 2.01 \times 10^{-2}$), (d) reconstructed from ALM_TMSBL ($\epsilon_X = 1.54 \times 10^{-2}$).

signals and outliers was similar to above. Noise level was varied to yield a range of SNR values from 5 dB to 40 dB, which is of practical interest for applications including EEG/MEG signal processing [43], [44].

As can be seen in Fig. 8, the ADMM approach yields significantly lower reconstruction error than the sequential approaches

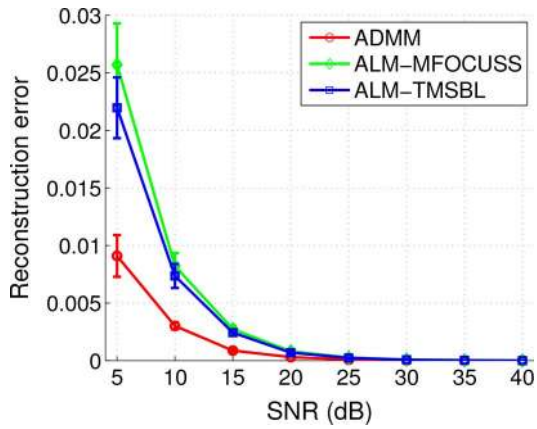


Fig. 8. Reconstruction error vs. SNR. Average reduction in reconstruction error between ADMM and ALM_TMSBL is 57.3%.

across all noise levels examined, especially at low SNR conditions. The robustness to noise is attributed to the incorporation of prior smoothness knowledge.

8) Additional Numerical Examples: In the previous subsections we have analyzed the performance of various algorithms from different problem perspectives. In addition to the comprehensive analysis above, we present in this subsection some additional numerical examples, which are considered relatively "tough". The objective is to provide a thorough investigation into the algorithmic performance under challenging experimental conditions.

The reconstruction errors from the various algorithms under the challenging conditions are summarized in Table III. Each row in the table corresponds to one test case. For example, in test case 1 we consider heavily compressed measurement data with $N/M = 10$; in test case 2 the row-wise sparsity of $\mathbf{X}$ is as high as $K/N = 0.25$; test case 3 is similar to test case 1 but with shorter measurement sequences; and finally test case 4 is a combination of the first two test cases. As is evidenced by the results in the table, the performance of all algorithms under investigation degrades due to the difficulty of the test cases. However, it is clear that the performance of the ADMM algorithm degrades much more gracefully than its sequential counterparts, demonstrating its robustness even under relatively challenging experimental conditions.

In summary, we have examined the algorithmic performance of the approaches discussed in this paper. Extensive experiments covering a wide range of conditions confirm that the ADMM approach is highly effective in recovering signals from corrupted measurements.

V. Conclusion

In this paper we considered the problem of recovering jointly sparse vectors from under-determined measurements that are corrupted by both additive noise and outliers. This was presented as the robust extension of the MMV problem. To solve this problem, we proposed two general frameworks, with the first being a sequential approach based on preprocessing and the state-of-the-art technologies, and the second being based on the formulation of an innovative regularized fitting problem. We proposed an algorithmic solution based on the ADMM procedure to solve this regularized fitting problem. The approach based on ADMM has excellent robustness to outliers, and yields significantly lower reconstruction error than its sequential counterpart.

Appendix A

In this appendix, we show that (5) is a convex optimization problem in $(\mathbf{X}, \mathbf{E})$. First, define $\mathbf{Z} = [\mathbf{X}^{\mathrm{T}}, \mathbf{E}^{\mathrm{T}}]^{\mathrm{T}}$ as optimization variable. With this definition, the cost function in (5) can be written as

$$\frac{1}{2}\|\mathbf{Y} - \mathbf{\Phi X} - \mathbf{E}\|_{\mathrm{F}}^2 + \lambda_1\|\mathbf{X}\|_{1,\mathbf{P}} + \lambda_2\|\mathbf{E}\|_1 = \frac{1}{2}\|\mathbf{Y} - \mathbf{AZ}\|_{\mathrm{F}}^2 + \lambda_1\|\mathbf{BZ}\|_{1,\mathbf{P}} + \lambda_2\|\mathbf{CZ}\|_1, \quad \text{(A.1)}$$

where $\mathbf{A} = [\mathbf{\Phi}, \mathbf{I}_M]$, $\mathbf{B} = [\mathbf{I}_N, \mathbf{0}_{N\times M}]$, and $\mathbf{C} = [\mathbf{0}_{M\times N}, \mathbf{I}_M]$, respectively.

The first term in (A.1) is separable across $\{\mathbf{z}_{\cdot i}\}_{i=1}^T$, where each term is a quadratic function of $\mathbf{z}_{\cdot i}$, i.e.,

$$\frac{1}{2}\|\mathbf{Y} - \mathbf{AZ}\|_{\mathrm{F}}^2 = \sum_{i=1}^{T}\left\{\frac{1}{2}\mathbf{z}_{\cdot i}^{\mathrm{T}}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)\mathbf{z}_{\cdot i} - \mathbf{y}_{\cdot i}^{\mathrm{T}}\mathbf{A}\mathbf{z}_{\cdot i} + \frac{1}{2}\|\mathbf{y}_{\cdot i}\|_2^2\right\}. \quad \text{(A.2)}$$

Since $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is positive semidefinite, it follows that each of quadratic terms in (A.2) is a convex function in $\mathbf{z}_{\cdot i}$, and therefore the first term in (A.1) is convex in $\mathbf{Z}$.

To see the second term in (A.1) is convex in $\mathbf{Z}$, let $\mathbf{Z}_1 = [\mathbf{X}_1^{\mathrm{T}}, \mathbf{E}_1^{\mathrm{T}}]^{\mathrm{T}}$ and $\mathbf{Z}_2 = [\mathbf{X}_2^{\mathrm{T}}, \mathbf{E}_2^{\mathrm{T}}]^{\mathrm{T}}$, respectively. For any $\alpha \in [0, 1]$, it is clear that

$$\begin{aligned}\lambda_1\|\mathbf{B}\left(\alpha\mathbf{Z}_1 + (1-\alpha)\mathbf{Z}_2\right)\|_{1,\mathbf{P}} &= \lambda_1\|\alpha\mathbf{X}_1 + (1-\alpha)\mathbf{X}_2\|_{1,\mathbf{P}} \\ &\le \alpha\lambda_1\|\mathbf{X}_1\|_{1,\mathbf{P}} + (1-\alpha)\lambda_1\|\mathbf{X}_2\|_{1,\mathbf{P}} \\ &= \alpha\lambda_1\|\mathbf{BZ}_1\|_{1,\mathbf{P}} + (1-\alpha)\lambda_1\|\mathbf{BZ}_2\|_{1,\mathbf{P}}, \quad \text{(A.3)}\end{aligned}$$

where the inequality follows from the homogeneity and subadditivity properties of norms. By definition of convexity, we see $\lambda_1\|\mathbf{BZ}\|_{1,\mathbf{P}}$ is convex in $\mathbf{Z}$.

In a similar fashion we can show the third term in (A.1) is convex in $\mathbf{Z}$. Therefore, (A.1), as the sum of convex functions in $\mathbf{Z}$, is itself convex in $\mathbf{Z}$. Since (5) is unconstrained, it follows that it is a convex optimization in $\mathbf{Z} = [\mathbf{X}^{\mathrm{T}}, \mathbf{E}^{\mathrm{T}}]^{\mathrm{T}}$.

References

[1] C. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.

[2] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.

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

[4] A. K. Katsaggelos, *Digital Image Restoration*, M. R. Schroeder, T. Kohonen, and T. S. Huang, Eds. Secaucus, NJ, USA: Springer-Verlag/New York, Inc., 1991.

[5] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.

[6] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[7] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Conf. Rec. 27th Asilomar Conf. Signals, Syst., Comput.*, Nov. 1993, vol. 1, pp. 40–44.

[8] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[9] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Series B (Methodol.)*, vol. 58, no. 1, pp. 267–288, 1996.

[10] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[11] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J.. Mach. Learn. Res.*, vol. 1, pp. 211–244, Sep. 2001.

[12] S. D. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 53–63, 2010.

[13] T. Zhang, B. Ghanem, S. Liu, and N. Ahuja, "Robust visual tracking via multi-task sparse learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2042–2049.

[14] N. Vaswani, "Kalman filtered compressed sensing," in *Proc. 15th IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 893–896.

[15] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006.

[16] 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, Jul. 2005.

[17] J. A. Tropp, "Algorithms for simultaneous sparse approximation. Part II: Convex relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, 2006.

[18] L. Kiryung, Y. Bresler, and M. Junge, "Subspace methods for joint sparse recovery," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3613–3641, Jun. 2012.

[19] D. P. Wipf and B. D. Rao, "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3704–3716, Jul. 2007.

[20] J. Kim, W. Chang, B. Jung, D. Baron, and J. C. Ye, "Belief propagation for joint sparse recovery," Feb. 2011, arXiv:1102.3289 [Online]. Available: http://arxiv.org/abs/1102.3289

[21] J. Ziniel and P. Schniter, "Efficient high-dimensional inference in the multiple measurement vector problem," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 340–354, Jan. 2013.

[22] P. Pal and P. Vaidyanathan, "On application of LASSO for sparse support recovery with imperfect correlation awareness," in *Conf. Rec. 46th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2012, pp. 958–962.

[23] P. Pal and P. P. Vaidyanathan, "Correlation-aware sparse support recovery: Gaussian sources," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2013, pp. 5880–5884.

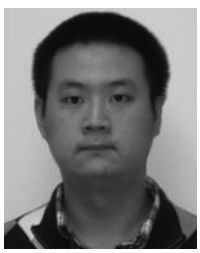
[24] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst, "Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms," *J. Fourier Anal. Appl.*, vol. 14, no. 5–6, pp. 655–687, 2008.

[25] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 505–519, Jan. 2010.

[26] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 5, pp. 912–926, Sep. 2011.

[27] Z. Zhang and B. D. Rao, "Iterative reweighted algorithms for sparse signal recovery with temporally correlated source vectors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2011, pp. 3932–3935.

[28] M. Luessi, D. S. Babacan, R. Molina, and A. K. Katsaggelos, "Bayesian simultaneous sparse approximation with smooth signals," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5716–5729, Nov. 2013.
[29] Z. Chen, R. Molina, and A. K. Katsaggelos, "Recovery of correlated sparse signals from under-sampled measurements," in *Proc. 22nd Eur. Signal Process. Conf.*, Lisbon, Portugal, 2014.
[30] R. J. Croft and R. J. Barry, "Removal of ocular artifact from the EEG: A review," *Neurophysiologie Clinique*, vol. 30, no. 1, pp. 5–19, 2000.
[31] Statistical Parametric Mapping, 2014 [Online]. Available: http://www.fil.ion.ucl.ac.uk/spm/
[32] J. Cai, E. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, Jan. 2010.
[33] A. Ganesh, Z. Lin, J. Wright, L. Wu, M. Chen, and Y. Ma, ""Fast algorithms for recovering a corrupted low-rank matrix," in *Proc. 3rd IEEE Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Dec. 13–16, 2009, pp. 213–216.
[34] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," Univ. of Illinois at Urbana-Champaign, IL, USA, Tech. Rep. #UILU-ENG-09-2215, 2009.
[35] X. Ding, L. He, and L. Carin, "Bayesian robust principal component analysis," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3419–3430, 2011.
[36] S. Babacan, M. Luessi, R. Molina, and A. Katsaggelos, "Sparse Bayesian methods for low-rank matrix estimation," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 3964–3977, Aug. 2012.
[37] R. A. Horn and C. R. Johnso*n, Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[38] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.
[39] R. H. Bartels and G. W. Stewart, "Solution of the matrix equation ax + xb = c," *Commun. ACM*, vol. 15, no. 9, pp. 820–826, Sep. 1972.
[40] D. C. Sorensen and Y. Zhou, "Direct methods for matrix Sylvester and Lyapunov equations," *J. Appl. Math.*, vol. 2003, no. 6, pp. 277–303, 2003.
[41] B. S. He, H. Yang, and S. L. Wang, "Alternating direction method with self-adaptive penalty parameters for monotone variational inequalities," *J. Optim. Theory Appl.*, vol. 106, no. 2, pp. 337–356, 2000.
[42] Multiple-Spars Toolbox, 2014 [Online]. Available: http://asi.insa-rouen.fr/enseignants/~arakoto/code/SSAindex.html
[43] K. Friston, L. Harrison, J. Daunizeau, S. Kiebel, C. Phillips, N. Trujillo-Barreto, R. Henson, G. Flandin, and J. Mattout, "Multiple sparse priors for the M/EEG inverse problem," *NeuroImage*, vol. 39, no. 3, pp. 1104–1120, 2008.
[44] C. Phillips, J. Mattout, M. D. Rugg, P. Maquet, and K. J. Friston, "An empirical Bayesian solution to the source reconstruction problem in EEG," *NeuroImage*, vol. 24, no. 4, pp. 997–1011, 2005.



Zhaofu Chen received the B.Sc. degree in information science and engineering from Zhejiang University, Hangzhou, China, in 2008, the M.Sc. degree in electrical and computer engineering from University of Florida, Gainesville, FL, in 2010, and the Ph.D. degree in electrical engineering and computer science from Northwestern University, Evanston, IL, in 2014. He is currently at Google, Inc.

His research interests include signal processing, sparse and low-rank problems, statistical approaches to inverse problems.

Rafael Molina (M'88) was born in 1957. He received the Degree in mathematics (statistics) in 1979 and the Ph.D. degree in optimal design in linear models in 1983. He became a Professor of Computer Science and Artificial Intelligence with University of Granada, Spain in 2000. He was a Former Dean of the Computer Engineering School, University of Granada (1992–2002), and the Head of the Computer Science and Artificial Intelligence Department, University of Granada (2005–2007).

His current research interests include using Bayesian modeling and inference in problems like image restoration (applications to astronomy and medicine), super resolution of images and video, blind deconvolution, computational photography, source recovery in medicine, compressive sensing, low rank matrix decomposition, active learning, and classification.

Dr. Molina serves the IEEE and other Professional Societies, including Applied Signal Processing. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING (2005–2007), an Associate Editor of *Progress in Artificial Intelligence* (2010), an Associate Editor of *Digital Signal Processing* (2011), and an Area Editor (2011). He is a recipient of the IEEE ICIP Papaer Award in 2007, the ISPA Best Paper Award in 2009, and the EUSIPCO Best Student Paper Award in 2013. He co-authored a paper that received the Runner-Up Prize at Reception for early-stage researchers at the House of Commons.

Aggelos K. Katsaggelos (F'98) received the Diploma degree in electrical and mechanical engineering from the Aristotelian University of Thessaloniki, Greece, in 1979, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, in 1981 and 1985, respectively.

In 1985, he joined the Department of Electrical Engineering and Computer Science at Northwestern University, where he is currently a Professor holder of the AT&T chair. He was previously the holder of the Ameritech Chair of Information Technology (1997–2003). He is also the Director of the Motorola Center for Seamless Communications, a member of the Academic Staff, NorthShore University Health System, an affiliated faculty at the Department of Linguistics and he has an appointment with the Argonne National Laboratory.

He has published extensively in the areas of multimedia signal processing and communications (over 200 journal papers, 500 conference papers and 40 book chapters) and he is the holder of 25 international patents. He has supervised 46 Ph.D. theses.

Among his many professional activities Prof. Katsaggelos was Editor-in-Chief of the IEEE Signal Processing Magazine (1997–2002), a BOG Member of the IEEE Signal Processing Society (1999–2001), and a member of the Publication Board of the IEEE PROCEEDINGS (2003–2007). He is a Fellow of the IEEE (1998) and SPIE (2009) and the recipient of the IEEE Third Millennium Medal (2000), the IEEE Signal Processing Society Meritorious Service Award (2001), the IEEE Signal Processing Society Technical Achievement Award (2010), an IEEE Signal Processing Society Best Paper Award (2001), an IEEE ICME Paper Award (2006), an IEEE ICIP Paper Award (2007), an ISPA Paper Award (2009), and a EUSIPCO Paper Award (2013). He was a Distinguished Lecturer of the IEEE Signal Processing Society (2007–2008).