Robust Jointly-Sparse Signal Recovery Based on Minimax Concave Loss Function

Kyohei Suzuki*, Yukawa Masahiro*
*Department of Electronics and Electrical Engineering, Keio University, Japan

***Abstract*—We propose a robust approach to recovering the jointly-sparse signals in the presence of outliers. We formulate the recovering task as a minimization problem involving three terms: (i) the minimax concave (MC) loss function, (ii) the MC penalty function, and (iii) the squared Frobenius norm. The MC-based loss and penalty functions enhance robustness and group sparsity, respectively, while the squared Frobenius norm induces the convexity. The problem is solved, via reformulation, by the primal-dual splitting method, for which the convergence condition is derived. Numerical examples show that the proposed approach enjoys remarkable outlier robustness.**

***Index Terms*—robustness, minimax concave function, jointly-sparse signals, multiple measurement vector problem, feature selection**

I. Introduction

Outlier, or impulsive noise, is one of the major causes of performance degradation in signal processing, machine learning, data analysis, *etc*. In this paper, we present a robust approach with provable global convergence using a certain "nonconvex" loss function, for recovering jointly-sparse vectors (sharing a common support). In fact, the jointly-sparse recovery problem has widely been studied in several different contexts. We briefly review the history of multiple measurement vector (MMV) problem and feature selection.

The MMV problem aims to recover multiple sparse vectors sharing the same support from measurements, encountered in many applications [1]–[3]. A typical formulation for the MMV problem is given as follows:

$$\min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{X}\|_{2,0} \text{ s.t. } \boldsymbol{B}=\boldsymbol{X}\boldsymbol{A}, \tag{1}$$

where $\boldsymbol{X} := [\boldsymbol{x}_1 \dots \boldsymbol{x}_d] \in \mathcal{X} := \mathbb{R}^{n\times d}$, $\boldsymbol{B} \in \mathcal{Y} := \mathbb{R}^{n\times m}$, $\boldsymbol{A} \in \mathbb{R}^{d\times m}$ for $m < d$, and $\|\boldsymbol{X}\|_{2,0} := |(\{i \in \{1,2,\dots,d\} \mid \boldsymbol{x}_i \neq \boldsymbol{0}\})|$ ($|\cdot|$ denotes the cardinality of a set). In the particular case of $n = 1$, the MMV problem reduces to single measurement vector (SMV) problem, which is an ordinary problem in compressive sensing. While the recovery results of MMV have no advantage over SMV in the worst-case scenario, it performs much better in an average-case analysis [4], [5]. There are many techniques to solve the MMV problem [5]–[8].

Without the condition $m < d$, the MMV problem in (1) also represents the feature selection problem, which is a problem of selecting an important subset of features from input data, thereby enhancing the performance. Feature selection plays an important role in such applications that involve high dimensional data [9], [10]. In bioinformatics, for instance, robustness is of particular importance [9]. Robust feature selection (RFS) [11] is based on the $\ell_{2,1}$ loss minimization (see Section II), which leads to robust recovery since it is insensitive to large errors. Despite its fast convergence, the scalability of RFS to high dimensional data is rather limited due to the computation of matrix inversion.

In this paper, we propose an efficient robust approach to recovering jointly-sparse vectors in the presence of outliers. To attain robust estimates against outliers, we introduce a nonconvex loss based on the minimax concave (MC) function [12]–[14], as well as the MC-based penalty to promote group sparsity. Since the MC loss returns a constant value for those errors exceeding a certain threshold, it prevents from being affected by outliers severely. While the MC function is nonconvex, the convexity of the whole cost function is maintained due to the introduction of the squared Frobenius norm. The reformulated problem can be solved by the primal-dual splitting method [15], for which the convergence condition for the current specific case is derived with a Lipschitz constant of the gradient used in the method. Numerical simulations show that the proposed approach can recover the support of jointly-sparse signals robustly even in huge outliers scenarios; the $\ell_{2,1}$ norm loss function fails in this case.

II. Preliminaries

We first introduce notations and definitions. We then state the problem addressed in the present study.

A. Notation and definition

Throughout this paper, matrices, vectors, and linear operators are denoted by boldface uppercase letters, boldface lowercase letters, and uppercase letters, respectively. For any matrix $\boldsymbol{A}$, the ith column is denoted by $\boldsymbol{a}_i$. Let $\boldsymbol{I}_n$ denote the $n \times n$ identity matrix. The ℓ_p norm of any $\boldsymbol{v} \in \mathbb{R}^m$ is defined as $\|\boldsymbol{v}\|_p := (\sum_{i=1}^m |v_i|^p)^{1/p}$ for $p \geq 1$. The $\ell_{2,1}$ and Frobenius norms of any $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ are defined as $\|\boldsymbol{A}\|_{2,1} := \sum_{i=1}^m \|\boldsymbol{a}_i\|_2$ and $\|\boldsymbol{A}\|_{\mathrm{F}} := (\sum_{i=1}^n \sum_{j=1}^m A_{i,j}^2)^{1/2}$, respectively. For any matrix $\boldsymbol{A}$, $\lambda_{\max}(\boldsymbol{A})$ denotes the maximal eigenvalue of $\boldsymbol{A}$.

Let $\mathcal{X} := \mathbb{R}^{n\times d}$ and $\mathcal{Y} := \mathbb{R}^{n\times m}$. For any Hilbert spaces $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ and $(\mathcal{Y}, \langle\cdot,\cdot\rangle_{\mathcal{Y}})$, $\mathcal{B}(\mathcal{X},\mathcal{Y})$ denotes the set of bounded linear operators from $\mathcal{X}$ to $\mathcal{Y}$. If $\langle L\boldsymbol{Y}, \boldsymbol{Y}\rangle_{\mathcal{Y}} \geq 0$ for any $\boldsymbol{Y} \in \mathcal{Y}$, L is called a positive semidefinite operator. For any linear operator $L \in \mathcal{B}(\mathcal{Y},\mathcal{Y})$, the adjoint operator $L^* \in \mathcal{B}(\mathcal{Y},\mathcal{Y})$ is defined as the operator satisfying

This work was supported by KAKENHI Grant Number 18H01446.

$\langle L\boldsymbol{B}, \boldsymbol{Y}\rangle_{\mathcal{Y}} = \langle \boldsymbol{B}, L^*\boldsymbol{Y}\rangle_{\mathcal{Y}}$ for any $\boldsymbol{B}, \boldsymbol{Y} \in \mathcal{Y}$. If $L^* = L$, L is called a self-adjoint operator. A square root of a positive semidefinite self-adjoint operator L is denoted by $L^{1/2}$, and is defined as an self-adjoint operator Λ satisfying $\Lambda^2 = L$. Let O denote a self-mapping (defined on an arbitrary space) that maps any point to the null vector.

Let $\Gamma_0(\mathcal{X})$ denote the set of proper lower semicontinuous[1] convex functions from $\mathcal{X}$ to $(-\infty, +\infty]$. The infimal convolution of any functions $f, g : \mathcal{Y} \to (-\infty, +\infty]$ is given by $(f \boxdot g) : \mathcal{Y} \to [-\infty, +\infty] : \boldsymbol{Z} \mapsto \inf_{\boldsymbol{Y}\in\mathcal{Y}}\{f(\boldsymbol{Y}) + g(\boldsymbol{Z}-\boldsymbol{Y})\}$, when the minimizer exists at every point of its domain. The Moreau envelope of f is defined as $f \boxdot q$, where $q := \frac{1}{2}\|\cdot\|_{\mathrm{F}}^2$. For any $f \in \Gamma_0(\mathcal{X})$, the proximity operator is defined as $\mathrm{prox}_f(\boldsymbol{X}) := \arg\min_{\boldsymbol{\Xi}\in\mathcal{X}} \left(f(\boldsymbol{\Xi}) + \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{\Xi}\|_{\mathcal{X}}^2\right)$, where $\|\boldsymbol{X}\|_{\mathcal{X}} := \sqrt{\langle \boldsymbol{X}, \boldsymbol{X}\rangle_{\mathcal{X}}}$.

B. Problem statement

Given a known matrix $\boldsymbol{A} \in \mathbb{R}^{d\times m}$, we consider the following model:

$$\boldsymbol{B} = \boldsymbol{X}^*\boldsymbol{A} + \boldsymbol{E} + \boldsymbol{O} \in \mathcal{Y}\,(:= \mathbb{R}^{n\times m}) \tag{2}$$

where $\boldsymbol{X}^* \in \mathcal{X}\,(:= \mathbb{R}^{n\times d})$, and $\boldsymbol{E} \in \mathcal{Y}$ and $\boldsymbol{O} \in \mathcal{Y}$ are the noise and outlier matrices, respectively. Here, $\boldsymbol{X}^*$ and $\boldsymbol{O}$ are assumed to be column sparse; this assumption is also used implicitly in [11], [16], [17]. The problem addressed in this paper is stated as follows: recover $\boldsymbol{X}^*$ in (2) from the known/measurable matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ (with $\boldsymbol{E}$ and $\boldsymbol{O}$ unknown). Due to the presence of outliers, the classical regularized least square regression approach

$$\min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{B} - \boldsymbol{X}\boldsymbol{A}\|_{\mathrm{F}}^2 + \lambda R(\boldsymbol{X}) \tag{3}$$

is known to fail, where $R(\boldsymbol{X})$ is the regularization term and $\lambda > 0$. To attain robustness against outliers, Problem (P_0) has been considered in the context of feature selection [11]:

$$(\mathrm{P}_0) \quad \min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{B} - \boldsymbol{X}\boldsymbol{A}\|_{2,1} + \lambda\|\boldsymbol{X}\|_{2,1},$$

which leads to outlier robustness compared to the classical approach. For feature selection, d, m, and n are the dimension of data, the number of data, and the dimension of observations (the number of classes in the classification case), respectively, while in the context of the MMV problem, those are the dimension of the vectors to be recovered, the dimension of measurements, the number of measurements, respectively.

III. Proposed Approach

We first present the proposed problem formulation to further enhance the robustness compared to RFS. We then show that the newly posed problem can be solved by the primal-dual splitting method [15] under an appropriate reformulation, and present closed-form expressions of the operators used in the algorithm. We finally present convergence analysis. All the results will be presented without proofs. An extended version of the present work including all the proofs will be presented elsewhere.

[1]A function $f : \mathcal{X} \to (-\infty, +\infty]$ is proper if the domain $\mathrm{dom} f := \{\boldsymbol{X} \in \mathcal{X} \mid f(\boldsymbol{X}) < +\infty\} \neq \emptyset$. A function $f : \mathcal{X} \to (-\infty, +\infty]$ is lower semicontinuous at $\boldsymbol{X} \in \mathcal{X}$ if, for every $(\boldsymbol{X}_k)_{k=1}^{\infty} \subset \mathcal{X}$, $\boldsymbol{X}_k \to \boldsymbol{X} \Rightarrow f(\boldsymbol{X}) \leq \liminf_{k\to\infty} f(\boldsymbol{X}_k)$.

A. Proposed formulation

We formulate the robust jointly-sparse signal recovery problem as follows:

$$(\mathrm{P}_1) \quad \min_{\boldsymbol{X}\in\mathcal{X}} \left(\Phi_L(\boldsymbol{B} - \boldsymbol{X}\boldsymbol{A}) + \lambda_1\Phi_M(\boldsymbol{X}) + \frac{\lambda_2}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2\right),$$

where $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, the MC functions $\Phi_L : \mathcal{Y} \to \mathbb{R}$, $\Phi_M : \mathcal{X} \to \mathbb{R}$, and the linear operators $L \in \mathcal{B}(\mathcal{Y}, \mathcal{Y})$, $M \in \mathcal{B}(\mathcal{X}, \mathcal{X})$ are defined as follows:

$$\Phi_L(\boldsymbol{Y}) := \|\boldsymbol{Y}\|_{2,1} - \min_{\boldsymbol{Z}\in\mathcal{Y}} \left(\|\boldsymbol{Z}\|_{2,1} + \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{Z}\|_L^2\right), \tag{4}$$

$$\Phi_M(\boldsymbol{X}) := \|\boldsymbol{X}\|_{2,1} - \min_{\boldsymbol{\Xi}\in\mathcal{X}} \left(\|\boldsymbol{\Xi}\|_{2,1} + \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{\Xi}\|_M^2\right), \tag{5}$$

$$L\boldsymbol{B} := \boldsymbol{B}\mathrm{diag}(l_1, \ldots, l_m),\ l_i > 0,\ \forall i = 1, \ldots, m, \tag{6}$$

$$M\boldsymbol{X} := \boldsymbol{X}\mathrm{diag}(\mu_1, \ldots, \mu_n),\ \mu_j > 0,\ \forall j = 1, \ldots, n. \tag{7}$$

Note here that $L = L^*$ and $M = M^*$.

For $L = O$, $M = O$ and $\lambda_2 = 0$, problem (P_1) reduces to (P_0). Since the MC functions used in both loss and penalty terms in (P_1) are nonconvex, the third term is necessary to obtain the convexity of the whole cost function (see Proposition 1 below and its following discussions). Indeed, the use of $\Phi_L(\boldsymbol{B} - \boldsymbol{X}\boldsymbol{A})$ makes the outliers less important than using the $\ell_{2,1}$ norm. We show in the next section that (P_1) is solved by using the primal-dual splitting method [15].

B. Algorithm to solve (P_1)

The following lemma is used for the reformulation of Problem (P_1).

Lemma 1: For any $\boldsymbol{X} \in \mathcal{X}\,(:= \mathbb{R}^{n\times d})$,

$$\Phi_L(\boldsymbol{B} - \boldsymbol{X}\boldsymbol{A}) = \|\boldsymbol{B} - \boldsymbol{X}\boldsymbol{A}\|_{2,1} - \frac{1}{2}\|\boldsymbol{B} - \boldsymbol{X}\boldsymbol{A}\|_L^2 + \left(\imath_C \circ L^{1/2} \boxdot q\right)(L^{1/2}(\boldsymbol{B} - \boldsymbol{X}\boldsymbol{A})),$$

$$\Phi_M(\boldsymbol{X}) = \|\boldsymbol{X}\|_{2,1} - \frac{1}{2}\|\boldsymbol{X}\|_M^2 + \left(\imath_C \circ M^{1/2} \boxdot q\right)(M^{1/2}\boldsymbol{X}),$$

where $C := \mathrm{lev}_{\leq 1}\|\cdot\|_{2,\infty} := \{\boldsymbol{X} \in \mathcal{X} \mid \|\boldsymbol{X}\|_{2,\infty} \leq 1\}$, $\|\boldsymbol{A}\|_{2,\infty} := \max\{\|\boldsymbol{a}_1\|_2, \ldots, \|\boldsymbol{a}_m\|_2\}$, and the indicator function is defined as $\Gamma_0(\mathcal{Y}) \ni \imath_C : \mathcal{Y} \to [0, +\infty] : \boldsymbol{Y} \mapsto \begin{cases} 0, & \text{if } \boldsymbol{Y} \in C, \\ +\infty, & \text{otherwise.} \end{cases}$

By Lemma 1, Problem (P_1) is reformulated as follows:

$$(\mathrm{P}_1') \quad \min_{\boldsymbol{X}\in\mathcal{X}} [F(\boldsymbol{X}) + G(\boldsymbol{X}) + H(L_1\boldsymbol{X})]. \tag{8}$$

Here, the linear operator L_1 and the functions F, G, and H are defined as follows:

$$\mathcal{B}(\mathcal{X}, \mathcal{Y}) \ni\ L_1 : \boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{A}, \tag{9}$$

TABLE I
COMPUTATIONAL COMPLEXITY (q: THE NUMBER OF ITERATIONS, p: THE NUMBER OF POWER ITERATIONS).

Proposed	$\mathcal{O}(\max\{qnmd, \min\{d^2, m^2\}\max\{d, m, p\}\})$
RFS	$\mathcal{O}(qm(m+d)\max\{m+d, n\})$

$$\Gamma_0(\mathcal{X}) \ni F := \frac{\lambda_2}{2}\|\cdot\|_{\mathrm{F}}^2 - \frac{1}{2}\|L_1\cdot\|_L^2 - \frac{\lambda_1}{2}\|\cdot\|_M^2 + \langle L_1\cdot, \boldsymbol{B}\rangle_L + \Big(\imath_C \circ L^{1/2} \boxdot q\Big)(L^{1/2}(\boldsymbol{B} - L_1\cdot)) + \lambda_1\Big(\imath_C \circ M^{1/2} \boxdot q\Big)(M^{1/2}\cdot), \tag{10}$$

$$\Gamma_0(\mathcal{X}) \ni G := \lambda_1\|\cdot\|_{2,1}, \tag{11}$$

$$\Gamma_0(\mathcal{Y}) \ni H := \|\boldsymbol{B} - \cdot\|_{2,1}, \tag{12}$$

where the inner product $\langle\cdot,\cdot\rangle_L$ is defined as $\langle \boldsymbol{Y}, \boldsymbol{B}\rangle_L := \langle L\boldsymbol{Y}, \boldsymbol{B}\rangle_{\mathrm{F}} := \mathrm{Trace}((L\boldsymbol{Y})^{\mathsf{T}}\boldsymbol{B})$ for any $\boldsymbol{Y}, \boldsymbol{B} \in \mathcal{Y}$. The function F is strictly convex under some condition as shown by Proposition 1 below (F is also smooth as shown in Proposition 5 in Section III-D).

Proposition 1: The function F is convex if and only if

$$\lambda_2 \geq \lambda_{\max}\{\boldsymbol{A}\mathrm{diag}(l_1, \ldots, l_m)\boldsymbol{A}^{\mathsf{T}} + \lambda_1\mathrm{diag}(\mu_1, \ldots, \mu_n)\}. \tag{13}$$

In particular, F is strictly convex if and only if inequality (13) holds with strict inequality.

Strict convexity of the entire function in (P_1') is verified by Proposition 1 with the convexity of $G(\boldsymbol{X})$ and $H(L_1\boldsymbol{X})$. The coercivity of the entire function in (P_1') is verified with the following proposition.

Proposition 2: The function $\Phi_L(\boldsymbol{B} - L_1\cdot) + \lambda_1\Phi_M(\cdot) + \frac{\lambda_2}{2}\|\cdot\|_{\mathrm{F}}^2$ is coercive. Here, a function $f : \mathcal{X} \to (-\infty, +\infty]$ is coercive if $\lim_{\|\boldsymbol{X}\|_{\mathcal{X}}\to+\infty} f(\boldsymbol{X}) = +\infty$.

By Propositions 1 and 2, the cost function of (P_1') is strictly convex and coercive if $\lambda_2 > \lambda_{\max}\{\boldsymbol{A}\mathrm{diag}(l_1, \ldots, l_m)\boldsymbol{A}^{\mathsf{T}} + \lambda_1\mathrm{diag}(\mu_1, \ldots, \mu_n)\}$, and hence the uniqueness and existence of the solution of (P_1) are guaranteed in this case [18, p. 159].

For initial $\boldsymbol{X}_0 \in \mathcal{X}$, $\boldsymbol{Y}_0 \in \mathcal{Y}$ and $(\rho_t)_{t\in\mathbb{N}} \subset (0, \delta)$ for $\delta := 2 - \frac{\beta}{2}\left(\frac{1}{\tau} - \sigma\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})\right)^{-1} \in [1, 2)$, the primal-dual splitting method [15] to solve problem (P_1') is given as follows:[2]

$$\begin{aligned}
&\tilde{\boldsymbol{X}}_{t+1} = \mathrm{prox}_{\tau G}(\boldsymbol{X}_t - \tau(\nabla F(\boldsymbol{X}_t) + L_1^*\boldsymbol{Y}_t)),\\
&\tilde{\boldsymbol{Y}}_{t+1} = \mathrm{prox}_{\sigma H^*}(\boldsymbol{Y}_t + \sigma L_1(2\tilde{\boldsymbol{X}}_{t+1} - \boldsymbol{X}_t)),\\
&(\boldsymbol{X}_{t+1}, \boldsymbol{Y}_{t+1}) = \rho_t(\tilde{\boldsymbol{X}}_{t+1}, \tilde{\boldsymbol{Y}}_{t+1}) + (1-\rho_t)(\boldsymbol{X}_t, \boldsymbol{Y}_t).
\end{aligned} \tag{14}$$

Here, the convex conjugate $H^* : \mathcal{Y} \to [-\infty, +\infty]$ is defined as $H^*(\boldsymbol{Y}) := \sup_{\boldsymbol{Z}\in\mathcal{Y}}(\langle \boldsymbol{Z}, \boldsymbol{Y}\rangle_{\mathcal{Y}} - H(\boldsymbol{Z}))$, and β will be given explicitly in Section III-D below. The complexities of Algorithm (14) and RFS are summarized in Table I.

Remark 1: We discuss the possibility of applying the primal-dual splitting method or the alternating direction method of multipliers (ADMM) directly (i.e., without the reformulation) to Problem (P_1). We first mention that, in order to consider the first and third terms as a single function of $\boldsymbol{X}\boldsymbol{A}$, one needs the additional condition $\mathrm{rank}\boldsymbol{A} = d$, which strictly limits the applicability. A possible approach would therefore be to consider the second and third terms as a single function of $\boldsymbol{X}$ and then use the firm shrinkage operators studied in [21]. Let us first consider the primal-dual splitting methods of Chambolle-Pock or Condat. Both methods use the proximity operator of the conjugate function of the first term $\Phi_L(\boldsymbol{B} - \cdot)$ of (P_1). Due to the fact that the Fenchel conjugate of a given function coincides with that of its lower-semicontinuous convex envelope [18, Proposition 13.14], one can readily verify that the proximity operator of Φ_L^* becomes the zero mapping O. As a result, the proximity operator of $[\Phi_L \circ (\boldsymbol{B} - \cdot)]^*$ becomes the zero mapping as well (one can use the basic properties of conjugate function and proximity operator [18] to verify this). This means that the first term gives no impact on the algorithm output, and thus there is no hope to obtain a solution P_1. In contrast, the proximity operator of the conjugate function of Φ_L does not appear explicitly in the ADMM iterate. However, its convergence analysis is nontrivial in this case.[3] For instance, the approach by Eckstein and Bertsekas [22] applies the Douglas-Rachford splitting method to the dual problem. The conjugate function appearing in the dual problem is replaced by the original function due essentially to the Moreau decomposition, which only holds in the convex case.

C. Closed-form expressions for the operators

We present below closed-form expressions of the operators used in (14) below.

Proposition 3: Closed-form expressions for ∇F, L_1^*, $\mathrm{prox}_{\tau G}$, and $\mathrm{prox}_{\sigma H^*}$ are given as follows:

1) $\nabla F(\boldsymbol{X}) = \lambda_2\boldsymbol{X} - \lambda_1\sum_{i=1}^{d}\min\left\{\frac{1}{\|\boldsymbol{x}_i\|_2}, \mu_i\right\}\boldsymbol{x}_i\boldsymbol{e}_{d,i}^{\mathsf{T}} + L_1^*\sum_{i=1}^{m}\min\left\{\frac{1}{\|[\boldsymbol{B} - L_1\boldsymbol{X}]_i\|_2}, l_i\right\}[\boldsymbol{B} - L_1\boldsymbol{X}]_i\boldsymbol{e}_{m,i}^{\mathsf{T}}$,
2) $L_1^*\boldsymbol{Y} = \boldsymbol{Y}\boldsymbol{A}^{\mathsf{T}}$,
3) $\mathrm{prox}_{\tau G}(\boldsymbol{X}) = \sum_{i=1}^{m}\max\left\{1 - \frac{\tau\lambda_1}{\|\boldsymbol{x}_i\|_2}, 0\right\}\boldsymbol{x}_i\boldsymbol{e}_{m,i}^{\mathsf{T}}$,
4) $\mathrm{prox}_{\sigma H^*}(\boldsymbol{Y}) = \sum_{i=1}^{m}\min\left\{\frac{1}{\|\boldsymbol{y}_i - \sigma\boldsymbol{b}_i\|_2}, 1\right\}(\boldsymbol{y}_i - \sigma\boldsymbol{b}_i)\,\boldsymbol{e}_{m,i}^{\mathsf{T}}$.

Here, $\{\boldsymbol{e}_{r,i}\}_{i=1}^r$ denotes the standard basis of $\mathbb{R}^r$ for any dimension r.

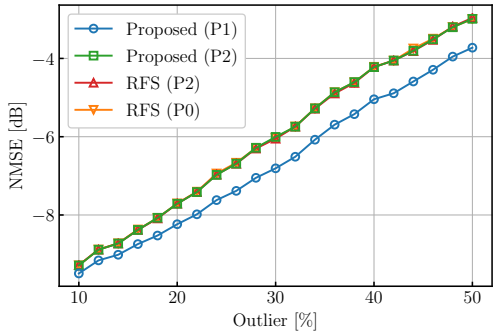
D. Convergence condition

Assuming that (13) holds, the convergence of the algorithm in (14) is ensured by [15, Theorem 3.1] as soon as ∇F is β-Lipschitz continuous and the parameters σ and τ satisfy the inequality

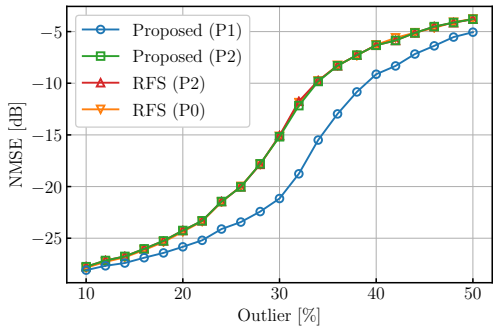
$$\frac{1}{\tau} - \sigma\|L_1\|_{\mathrm{F}}^2 \geq \frac{\beta}{2} > 0, \tag{15}$$

[2]Solving (P_1') by the primal-dual splitting method of Chambolle and Pock [19] or the alternating direction method of multipliers [20] may cause significant increases of complexity and memory requirements or an inner loop.

[3]Although it is shown in [21] that a rescaled firm-shrinkage operator is firmly nonexpansive, its application to the present case is not straightforward.

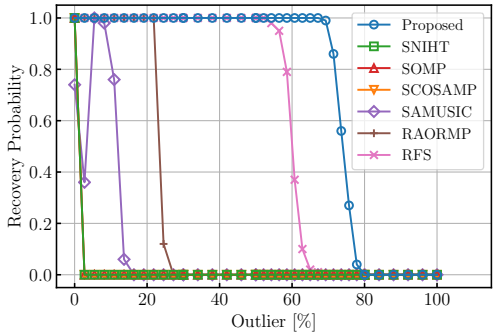


(a) SNR 10 dB

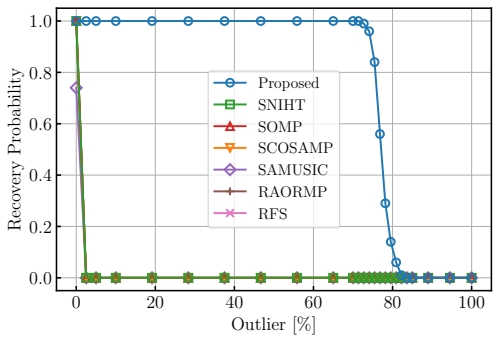


(b) SNR 30 dB

Fig. 1. NMSE for different column-sparsity of outlier matrix under $d = 128$, $m = 256$, and $n = 128$.



(a) $\mathrm{SNR_O}$ −30 dB



(b) $\mathrm{SNR_O}$ −3000 dB

Fig. 2. Recovery probability for different rate of outliers under $d = 256$, $m = 128$, $n = 32$, $k = 16$, and SNR 30 dB.

where a mapping $T : \mathcal{X} \to \mathcal{X}$ is Lipschitz continuous with constant $\beta > 0$ if $\|T(\boldsymbol{X}) - T(\boldsymbol{\Xi})\|_{\mathcal{X}} \leq \beta\|\boldsymbol{X} - \boldsymbol{\Xi}\|_{\mathcal{X}}$ for every $(\boldsymbol{X}, \boldsymbol{\Xi}) \in \mathcal{X}^2$, and $\|L\|_{\mathrm{F}} := \sup_{\|\boldsymbol{Y}\|_{\mathrm{F}}=1} \|L\boldsymbol{Y}\|_{\mathrm{F}}$ is the operator norm induced by the Frobenius norm. In the following propositions, we provide the constant β as well as an admissible choice for σ and τ.

Proposition 4: The proximity parameter $\sigma := 1/\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})(1/\tau - \beta/2)$ satisfies (15) for any $\tau < 2/\beta$ ($\Leftrightarrow 1/\tau - \beta/2 > 0$).

Proposition 5: The gradient operator ∇F is Lipschitz continuous with constant

$$\beta = \lambda_{\max}(\lambda_2 \boldsymbol{I}_d - \boldsymbol{A}\mathrm{diag}(l_1, \ldots, l_m)\boldsymbol{A}^{\mathsf{T}} - \mathrm{diag}(\mu_1, \ldots, \mu_n)) + \lambda_{\max}(\boldsymbol{A}\mathrm{diag}(l_1, \ldots, l_m)\boldsymbol{A}^{\mathsf{T}}) + \lambda_1 \max\{\mu_1, \ldots, \mu_n\}. \quad (16)$$

IV. Numerical Examples

We show the robustness of the proposed approach against outliers and its performance in support recovery.

A. Robustness

Matrices $\boldsymbol{X}^* \in \mathcal{X}\,(:= \mathbb{R}^{n\times d})$ and $\boldsymbol{A} \in \mathbb{R}^{d\times m}$ are generated with $d = 128$, $m = 256$, and $n = 128$ from the i.i.d. normal distribution $\mathcal{N}(0, 1)$. Here, we consider dense $\boldsymbol{X}^*$ to show the pure effects of the robustification (excluding the sparsification effect). Signal-to-noise ratio (SNR) of $\boldsymbol{E}$ is set to 10 dB and 30 dB. The outlier matrix $\boldsymbol{O}$ is column sparse, and its non-zero elements are drawn from $\mathcal{N}(0, 1)$ and are then multiplied by the factor 100. To measure the accuracy of the recovered signals, normalized mean squared errors (NMSE) given by $\mathrm{NMSE} := \|\boldsymbol{X}^* - \hat{\boldsymbol{X}}\|_{\mathrm{F}}^2/\|\boldsymbol{X}^*\|_{\mathrm{F}}^2$ are used. RFS [11] is considered for comparison. In [11], it is mentioned that one can easily extend RFS to those problems with different penalties, such as

$$(\mathrm{P}_2) \quad \min_{\boldsymbol{X}\in\mathcal{X}} \|\boldsymbol{B} - \boldsymbol{X}\boldsymbol{A}\|_{2,1} + \frac{\lambda_2}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^2, \quad (17)$$

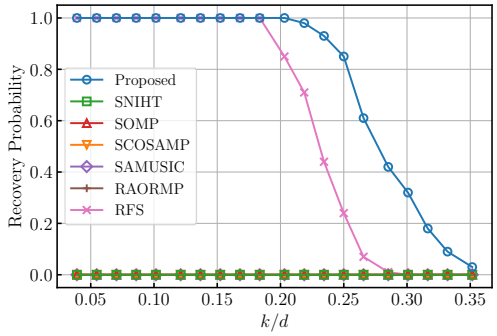
which is a special case of (P_1) for $L = O$ and $\lambda_1 = 0$. Figure 1 depicts the performance when the rate of outliers (the ratio between m and the number of nonzero columns of $\boldsymbol{O}$) changes between 10% and 50%. It is seen that the proposed approach is more effective for a denser outlier matrix.

We mention that, once recovering the support, one can remove those data corresponding to the off-support components and solve another (smaller size) regression problem that involves a smaller number of variables than that of the original problem. In that respect, the present setting of $d < m$ is a reasonable choice.

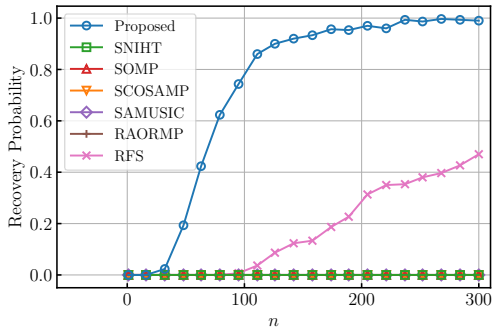
B. Support recovery

We demonstrate some properties of the proposed approach for support recovery. We compare the proposed approach to RFS and the state-of-the-art MMV algorithms: SNIHT [7], SOMP [6], SCOSAMP [7], SAMUSIC [8], and RAORMP [5].

First we investigate the robustness of the proposed approach for sparse signals under SNR 30 dB for $d = 256$, $m = 128$, and $n = 32$. We generate matrices $\boldsymbol{X}^* \in \mathcal{X}\,(:= \mathbb{R}^{n\times d})$ and $\boldsymbol{A} \in \mathbb{R}^{d\times m}$, both of which obey the i.i.d. normal distribution



(a) $d = 256$, $m = 128$, $n = 32$.



(b) $d = 256$, $m = 128$, $k = 0.4d$.

Fig. 3. Recovery probability as a function of k/d and n under SNR 30 dB, outlier 30%, and $\mathrm{SNR_O}$ -30 dB.

$\mathcal{N}(0, 1)$, and set $d - k$ column vectors of $\boldsymbol{X}$ to zero vectors, where k is called block sparsity. We define the signal to outlier-noise ratio as

$$\mathrm{SNR_O} := 10 \log_{10} \frac{\|\boldsymbol{X}^* \boldsymbol{A}\|_{\mathrm{F}}^2 / m}{\|\boldsymbol{O}\|_{\mathrm{F}}^2 / k'}, \tag{18}$$

where k' denotes the number of non-zero column vectors in $\boldsymbol{O}$ (we let $\mathrm{SNR_O} = -30, -3000$ dB). Figure 2 shows the support recovery probability under 100 trials for different rate of outliers. While the existing MMV algorithms fail in the presence of outliers, the proposed approach and RFS are robust. The robustness of the proposed approach significantly outperforms RFS in the case of $\mathrm{SNR_O}$ -30 dB. It should be remarked that the proposed approach has sufficient robustness even when $\mathrm{SNR_O}$ is -3000 dB.

Figure 3 plots the recovery probability as a function of k/d and n, under SNR 30 dB, outlier 30%, $\mathrm{SNR_O}$ -30 dB, and 300 trials. The proposed approach achieves higher recovery probability than RFS due to its remarkable robustness property coming from the use of the MC loss function.

V. Conclusion

We proposed a robust approach to recovering jointly-sparse signals in the presence of outliers. The main result is that the MC loss function leads to the remarkable robustness to outliers. The problem is solved by the primal-dual splitting method, for which the convergence condition for the current specific case is derived with a Lipschitz constant of the gradient used in the method. The numerical results showed that the proposed algorithm outperformed the RFS in terms of robustness against outliers.

References

[1] S. Erickson and C. Sabatti, "Empirical Bayes estimation of a sparse vector of gene expression changes," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

[2] I. F. Gorodnitsky, J. S. George, and B. D. Rao, "Neuromagnetic source imaging with FOCUSS: a recursive weighted minimum norm algorithm," *Electroencephalography and Clinical Neurophysiology*, vol. 95, no. 4, pp. 231–251, 1995.

[3] D. M. Malioutov, M. Cetin, and A. S. Willsky, "Source localization by enforcing sparsity through a Laplacian prior: an SVD-based approach," in *IEEE Workshop on Statistical Signal Processing, 2003*. IEEE, 2003, pp. 573–576.

[4] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Transactions on Information Theory*, vol. 56, no. 1, pp. 505–519, 2009.

[5] M. E. Davies and Y. C. Eldar, "Rank awareness in joint sparse recovery," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1135–1146, 2012.

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

[7] J. D. Blanchard, M. Cermak, D. Hanle, and Y. Jing, "Greedy algorithms for joint sparse recovery," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1694–1704, 2014.

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

[9] M. Banerjee, S. Mitra, and H. Banka, "Evolutionary rough feature selection in gene expression data," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 4, pp. 622–632, 2007.

[10] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 313–325.

[11] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Advances in Neural Information Processing Systems*, 2010, pp. 1813–1821.

[12] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[13] C.-H. Zhang *et al.*, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[14] I. Selesnick, "Sparse regularization via convex analysis," *IEEE Transactions on Signal Processing*, vol. 65, no. 17, pp. 4481–4494, 2017.

[15] L. Condat, "A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460–479, 2013.

[16] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.

[17] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011, pp. 1324–1329.

[18] H. H. Bauschke, P. L. Combettes *et al.*, *Convex analysis and monotone operator theory in Hilbert spaces*, 1st ed. Springer, 2011, vol. 408.

[19] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[20] D. Gabay, "Chapter IX Applications of the Method of Multipliers to Variational Inequalities," in *Studies in Mathematics and Its Applications*. Elsevier, 1983, vol. 15, pp. 299–331.

[21] I. Bayram, "On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty," *IEEE Transactions on Signal Processing*, vol. 64, no. 6, pp. 1597–1608, 2015.

[22] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1-3, pp. 293–318, 1992.