Block-Sparse Recovery with Optimal Block Partition

Hiroki Kuroda, *Member, IEEE,* and Daichi Kitahara, *Member, IEEE*

***Abstract*—This paper presents a convex recovery method for block-sparse signals whose block partitions are unknown a priori. We first introduce a nonconvex penalty function, where the block partition is adapted for the signal of interest by minimizing the mixed ℓ_2/ℓ_1 norm over all possible block partitions. Then, by exploiting a variational representation of the ℓ_2 norm, we derive the proposed penalty function as a suitable convex relaxation of the nonconvex one. For a block-sparse recovery model designed with the proposed penalty, we develop an iterative algorithm which is guaranteed to converge to a globally optimal solution. Numerical experiments demonstrate the effectiveness of the proposed method.**

***Index Terms*—Block-sparsity, unknown partition, penalty function, convex optimization, proximal splitting algorithm.**

I. Introduction

Sparsity has been widely used in signal processing and machine learning by the ℓ_1 norm, i.e., the best convex approximation of the ℓ_0 pseudo-norm [1]–[8]. At the same time, many efforts have also been devoted to improve the performance by better approximating the ℓ_0 pseudo-norm, e.g., [9]–[16]. Another important direction for the performance improvement is to exploit the underlying structure of nonzero components. Block-sparsity is a typical instance of such structured sparsity, where nonzero components are clustered in blocks. For recovery of a block-sparse signal whose block partition is *known a priori*, extensive researches, e.g., [17]–[26], show the remarkable performance improvements by the mixed ℓ_2/ℓ_1 norm using the known block partition, compared to the non-structured sparse methods. However, in fact, the information on the block partition is not available in many applications, e.g., acoustic signal recovery [27]–[29], image restoration [30]–[32], change detection [33]–[35], and radar signal processing [36]–[39]. For instance, the target signal of the phased array weather radar [37]–[41] is block-sparse in the Fourier domain due to the narrow bandwidth of the power spectrum, but the block partition is unknown because it depends on the unknown mean and standard deviation of the Doppler frequency. In such situations, the estimation accuracy of the mixed ℓ_2/ℓ_1 norm often degrades due to a pre-fixed block partition used instead of the ideal one.

Motivated by these observations, we consider the estimation problem of $\boldsymbol{x}^\star \in \mathbb{C}^N$ which is block-sparse across unknown non-overlapping blocks $\mathcal{B}_1^\star, \ldots, \mathcal{B}_{K^\star}^\star$ (see Fig. 1 for an illustration). More precisely, we suppose that the subvector $\boldsymbol{x}_{\mathcal{B}_k^\star}^\star := (x_n^\star)_{n \in \mathcal{B}_k^\star}$ contains only (nearly) zero components for many k, i.e.,

$$\boldsymbol{x}_{\mathcal{B}_k^\star}^\star \approx \mathbf{0} \text{ for many } k \in \{1, \ldots, K^\star\}.$$

We use the term *block-sparse* in a strict sense, i.e., $\mathcal{B}_k^\star$ consists of only consecutive indices for each $k = 1, \ldots, K^\star$:

$$\mathcal{B}_k^\star = \{n \in \{1, \ldots, N\} \mid n_k^\star \leq n \leq m_k^\star\},$$

with some $n_k^\star, m_k^\star \in \{1, \ldots, N\}$. We also suppose that $\mathcal{B}_1^\star, \ldots, \mathcal{B}_{K^\star}^\star$ form a partition of $\{1, \ldots, N\}$: $\bigcup_{k=1}^{K^\star} \mathcal{B}_k^\star = \{1, \ldots, N\}$, $\mathcal{B}_k^\star \neq \varnothing \quad (k = 1, \ldots, K^\star)$, and $\mathcal{B}_k^\star \cap \mathcal{B}_{k'}^\star = \varnothing \quad (k \neq k')$. Note that the non-overlapping condition $\mathcal{B}_k^\star \cap \mathcal{B}_{k'}^\star = \varnothing$ is not restrictive since the sizes of $\mathcal{B}_1^\star, \ldots, \mathcal{B}_{K^\star}^\star$ can be different.

Several attempts have been made to deal with the unknown blocks $\mathcal{B}_1^\star, \ldots, \mathcal{B}_{K^\star}^\star$. A commonly used modification is to use overlapping blocks in the mixed ℓ_2/ℓ_1 norm, e.g., [42]–[49]. Among them, the latent group lasso (LGL) method [48], [49] presents a sophisticated approach to select relevant blocks from pre-defined overlapping blocks. The LGL penalty is defined as the minimum of a convex function, and thus the corresponding regularization model can be solved as a convex optimization problem (see Appendix A for detail). However, since the problem size grows with the number of candidate blocks, the LGL method has a restriction on the number of candidate blocks, which leads to the degradation of the estimation accuracy. Nonconvex frameworks are also developed, but they have computational difficulties. The greedy methods [50], [51] have to manage candidate blocks explicitly, and thus are not suitable for the situation of unknown block partition. Indeed, when the block partition is unknown, they have to evaluate $\mathcal{O}(N^3)$ candidate blocks to determine which block is adopted. Bayesian approaches [52], [53] design nonconvex optimization problems considering the unknown block partition, and present their iterative solvers. However, the iterative algorithms require $\mathcal{O}(N^3)$ operations at every iteration due to the matrix inversion, which are not affordable in particular for large-scale problems. In addition, due to local minima of the nonconvex optimization problem, it is difficult to elude the influence of initial values given for algorithms. Thus, it is highly desired to realize a convex method which can handle the unknown blocks $\mathcal{B}_1^\star, \ldots, \mathcal{B}_{K^\star}^\star$ whose sizes are different, with low computational complexity.

In this paper, we propose a convex recovery method for

Manuscript received May 4, 2021; revised November 19, 2021; accepted February 23, 2022. Date of publication xxxxxx xx, 2022; date of current version xxxxxx xx, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Caroline Chaux. This work was supported in part by the Japan Society for the Promotion of Science under Grants-in-Aid 21K17827, 21H01592, and 19K20361. *(Corresponding author: Hiroki Kuroda.)*

The authors are with the College of Information Science and Engineering, Ritsumeikan University, Shiga 525-8577, Japan (e-mail: {kuroda, d-kita}@media.ritsumei.ac.jp).

Digital Object Identifier xx.xxxx/TSP.2022.xxxxxxx

$\boldsymbol{x}^\star$ block-sparse across the unknown blocks $\mathcal{B}_1^\star, \ldots, \mathcal{B}_{K^\star}^\star$. Our major contribution is to design a novel convex penalty function, named *latent optimally partitioned (LOP)-ℓ_2/ℓ_1 penalty*, which can automatically adapt the block partition for the signal of interest. More precisely, we first introduce a nonconvex penalty function as the minimum of the mixed ℓ_2/ℓ_1 norm over all possible block partitions. Then, to derive a tight convex relaxation of the nonconvex penalty function, using a variational representation of the ℓ_2 norm (see Lemma 1), we represent the nonconvex penalty function by the minimization of a convex function under the ℓ_0 pseudo-norm constraint on latent variables. Finally, the LOP-ℓ_2/ℓ_1 penalty is derived by replacing the ℓ_0 pseudo-norm constraint with its best convex approximation, i.e., the ℓ_1 norm constraint. To compute an optimal solution of the proposed block-sparse recovery model, we reformulate it into a convex optimization problem involving the latent variables of the LOP-ℓ_2/ℓ_1 penalty. Although the reformulated convex problem involves a discontinuous function which is difficult to handle, we show that it can be solved by combining proximal splitting techniques [54]–[62] and theory of perspective functions [63]–[65]. As a concrete example, we develop an $\mathcal{O}(N)$ iterative algorithm with guaranteed convergence to the optimal solution. Numerical experiments on both synthetic examples and real-world data show that the proposed method achieves superior estimation accuracy to existing convex methods including the LGL method.

The rest of this paper is organized as follows. In Section II, we introduce the LOP-ℓ_2/ℓ_1 penalty, and also show its extension to multi-dimensional signals. In Section III, we design the block-sparse recovery model with the LOP-ℓ_2/ℓ_1 penalty, and develop the iterative algorithm which converges to the optimal solution of the model. Section IV provides numerical examples, followed by conclusion in Section V.

A preliminary short version of this paper was presented at a conference [66].

Notations: $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{C}$ respectively denote the sets of all real numbers, all nonnegative real numbers, and all complex numbers. We denote the cardinality of a set $\mathcal{S}$ by $|\mathcal{S}|$. For matrices or vectors, we denote the simple transpose and the Hermitian transpose respectively by $(\cdot)^\top$ and $(\cdot)^{\mathrm{H}}$. For $\boldsymbol{x} = (x_1, \ldots, x_N)^\top \in \mathbb{C}^N$ and an index set $\mathcal{I} \subset \{1, \ldots, N\}$, $\boldsymbol{x}_{\mathcal{I}} := (x_n)_{n\in\mathcal{I}} \in \mathbb{C}^{|\mathcal{I}|}$ denotes the subvector of $\boldsymbol{x}$ indexed by $\mathcal{I}$. We define the support of $\boldsymbol{x} \in \mathbb{C}^N$ by $\mathrm{supp}(\boldsymbol{x}) := \{n \in \{1, \ldots, N\} \mid x_n \neq 0\}$. The ℓ_2 norm, the ℓ_1 norm, and the ℓ_0 pseudo-norm of $\boldsymbol{x} \in \mathbb{C}^N$ are respectively defined by $\|\boldsymbol{x}\|_2 := \sqrt{\boldsymbol{x}^{\mathrm{H}}\boldsymbol{x}}$, $\|\boldsymbol{x}\|_1 := \sum_{n=1}^N |x_n|$, and $\|\boldsymbol{x}\|_0 := |\mathrm{supp}(\boldsymbol{x})|$. We define the operator norm of $\boldsymbol{L} \in \mathbb{C}^{M\times N}$ by $\|\boldsymbol{L}\|_{\mathrm{op}} := \max_{\boldsymbol{x}\neq\boldsymbol{0}}[\|\boldsymbol{L}\boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2]$.

Let $(\mathcal{U}, \|\cdot\|)$ be a finite-dimensional Hilbert space. A function $f\colon \mathcal{U} \to \mathbb{R}\cup\{\infty\}$ is called proper if its effective domain $\mathrm{dom}(f) := \{\boldsymbol{u} \in \mathcal{U} \mid f(\boldsymbol{u}) < \infty\}$ is nonempty. A function $f\colon \mathcal{U} \to \mathbb{R}\cup\{\infty\}$ is lower semicontinuous if its lower level set $\{\boldsymbol{u} \in \mathcal{U} \mid f(\boldsymbol{u}) \leq a\}$ is closed for every $a \in \mathbb{R}$. A function $f\colon \mathcal{U} \to \mathbb{R}\cup\{\infty\}$ is convex if $f(\beta\boldsymbol{u} + (1-\beta)\boldsymbol{v}) \leq \beta f(\boldsymbol{u}) + (1-\beta) f(\boldsymbol{v})$ for every $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$ and $\beta \in (0,1)$. The set of all proper lower semicontinuous convex functions from $\mathcal{U}$ to $\mathbb{R}\cup\{\infty\}$ is denoted by $\Gamma_0(\mathcal{U})$. A function $f\colon \mathcal{U} \to \mathbb{R}\cup\{\infty\}$ is called coercive if $\lim_{\|\boldsymbol{u}\|\to\infty} f(\boldsymbol{u}) = \infty$. The proximity operator of $f \in \Gamma_0(\mathcal{U})$ is defined by $\mathrm{prox}_f(\boldsymbol{u}) := \arg\min_{\boldsymbol{v}\in\mathcal{U}} \left[f(\boldsymbol{v}) + \frac{1}{2}\|\boldsymbol{u} - \boldsymbol{v}\|^2\right]$.

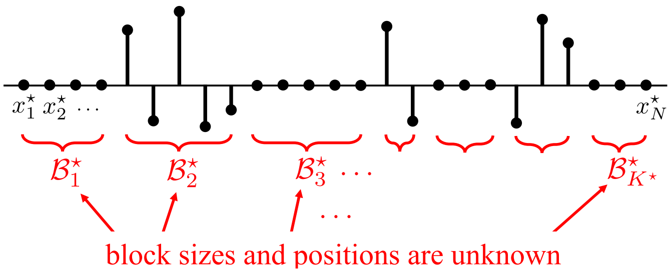

$x_1^\star$ $x_2^\star$ $\cdots$
$x_N^\star$
$\mathcal{B}_1^\star$
$\mathcal{B}_2^\star$
$\mathcal{B}_3^\star$ $\cdots$
$\cdots$
$\mathcal{B}_{K^\star}^\star$
block sizes and positions are unknown

Fig. 1. Illustration of a block-sparse signal, where we consider that the block partition is unknown a priori, i.e., the sizes and the positions of $\mathcal{B}_1^\star, \ldots, \mathcal{B}_{K^\star}^\star$ are unknown.

II. Design of Proposed Penalty Function

To evaluate the block-sparsity of $\boldsymbol{x} \in \mathbb{C}^N$ across the fixed non-overlapping blocks $\mathcal{B}_1, \ldots, \mathcal{B}_j \subset \{1, \ldots, N\}$, the mixed ℓ_2/ℓ_1 norm is widely used as a convex penalty function. The mixed ℓ_2/ℓ_1 norm is defined as the sum (i.e. the ℓ_1 norm) of the block-wise ℓ_2 norms:

$$\|\boldsymbol{x}\|_{2,1}^{(\mathcal{B}_k)_{k=1}^j} := \sum_{k=1}^j \sqrt{|\mathcal{B}_k|}\|\boldsymbol{x}_{\mathcal{B}_k}\|_2, \tag{1}$$

where we employ the weight $\sqrt{|\mathcal{B}_k|}$ so that larger blocks are penalized more heavily. Note that this weight is consistent with the existing studies assuming *known structures*, e.g, [21]–[24].[1] The mixed ℓ_2/ℓ_1 norm promotes the block-sparsity by pushing components in $\mathcal{B}_k$ toward zeros together. However, since the mixed ℓ_2/ℓ_1 norm excessively penalizes blocks composed of both zero and nonzero components, its performance degrades when $\mathcal{B}_1, \ldots, \mathcal{B}_j$ do not match with the ground-truth $\mathcal{B}_1^\star, \ldots, \mathcal{B}_{K^\star}^\star$.

To avoid the block mismatch, the block partition is adapted for $\boldsymbol{x}$ in the proposed penalty function named *latent optimally partitioned (LOP)-ℓ_2/ℓ_1 penalty*. As shown in Fig. 2, we first introduce a nonconvex penalty function $\psi_K(\boldsymbol{x})$ as the minimum of the mixed ℓ_2/ℓ_1 norm over the partition of at most K blocks. Then, as a suitable convex relaxation of $\psi_K(\boldsymbol{x})$, we derive the LOP-ℓ_2/ℓ_1 penalty $\Psi_\alpha(\boldsymbol{x})$. Concretely, we define $\psi_K(\boldsymbol{x})$ by

$$\begin{aligned}\psi_K(\boldsymbol{x}) &:= \min_{j\in\{1,\ldots,K\}} \left[\min_{(\mathcal{B}_k)_{k=1}^j \in \mathcal{P}_j} \|\boldsymbol{x}\|_{2,1}^{(\mathcal{B}_k)_{k=1}^j}\right] \\ &= \min_{j\in\{1,\ldots,K\}} \left[\min_{(\mathcal{B}_k)_{k=1}^j \in \mathcal{P}_j} \sum_{k=1}^j \sqrt{|\mathcal{B}_k|}\|\boldsymbol{x}_{\mathcal{B}_k}\|_2\right], \end{aligned} \tag{2}$$

[1] The weight is more important for the proposed framework as we consider the minimization of (1) over all possible block partitions (see Remark 1).

where $\mathcal{P}_j$ consists of all j block partitions of $\{1,\dots,N\}$, i.e.,

$$(\mathcal{B}_k)_{k=1}^{j} \in \mathcal{P}_j$$
$$\Leftrightarrow \begin{cases} \displaystyle\bigcup_{k=1}^{j} \mathcal{B}_k = \{1,\dots,N\}, \\ \mathcal{B}_k \neq \varnothing \quad (k=1,\dots,j), \\ \mathcal{B}_k \cap \mathcal{B}_{k'} = \varnothing \quad (k \neq k'), \\ \mathcal{B}_k = \{n \in \{1,\dots,N\} \mid n_k \leq n \leq m_k\} \\ \text{ for some } n_k, m_k \in \{1,\dots,N\} \quad (k=1,\dots,j). \end{cases} \tag{3}$$

Note that K can be set to an upper bound of the ground-truth $K^\star$. Note also that, since the sizes of $\mathcal{B}_1,\dots,\mathcal{B}_j$ are adapted, the non-overlapping condition $\mathcal{B}_k \cap \mathcal{B}_{k'} = \varnothing$ is not restrictive. In other words, since the block sizes are adapted, overlapping of blocks is unnecessary for the proposed penalty (see Fig. 2 for an example where zero and nonzero components are correctly separated by the non-overlapping blocks).

Remark 1 (Validity of the weight of the mixed ℓ_2/ℓ_1 norm)**.** The weight $\sqrt{|\mathcal{B}_k|}$ in (1) is needed to eliminate the trivial solution in (2). If we omit the weight in (1), the minimum of (2) is always attained when all variables are gathered into a single block, i.e., when $j=1$, because the inequality[2]

$$\|\boldsymbol{x}_{\mathcal{B}}\|_2 \leq \|\boldsymbol{x}_{\mathcal{B}'}\|_2 + \|\boldsymbol{x}_{\mathcal{B}''}\|_2$$

holds for any block $\mathcal{B} \subset \{1,\dots,N\}$ decomposed as $\mathcal{B} = \mathcal{B}' \cup \mathcal{B}''$ with $\mathcal{B}'$ and $\mathcal{B}''$ satisfying $\mathcal{B}' \neq \varnothing$, $\mathcal{B}'' \neq \varnothing$, and $\mathcal{B}' \cap \mathcal{B}'' = \varnothing$. Thus, we have to penalize larger blocks more heavily. We demonstrate the validity of the weight $\sqrt{|\mathcal{B}_k|}$ as follows. Suppose that a block $\mathcal{B}$ is decomposed into non-overlapping blocks $\mathcal{B}'$ and $\mathcal{B}''$ satisfying the above conditions. In addition, suppose that $\boldsymbol{x}_{\mathcal{B}'} \neq \boldsymbol{0}$ and $\boldsymbol{x}_{\mathcal{B}''} = \boldsymbol{0}$. Then, from $|\mathcal{B}| > |\mathcal{B}'|$, $\|\boldsymbol{x}_{\mathcal{B}}\|_2 = \|\boldsymbol{x}_{\mathcal{B}'}\|_2 \neq 0$, and $\|\boldsymbol{x}_{\mathcal{B}''}\|_2 = 0$, we have

$$\sqrt{|\mathcal{B}|}\|\boldsymbol{x}_{\mathcal{B}}\|_2 > \sqrt{|\mathcal{B}'|}\|\boldsymbol{x}_{\mathcal{B}'}\|_2 + \sqrt{|\mathcal{B}''|}\|\boldsymbol{x}_{\mathcal{B}''}\|_2.$$

This implies that the objective of (2) is decreased by decomposing $\mathcal{B}$ into $\mathcal{B}'$ and $\mathcal{B}''$ which respectively correspond to nonzero components and zero components. Namely, the trivial solution is avoided thanks to the weight $\sqrt{|\mathcal{B}_k|}$. Note that, since the number of blocks in (2) is at most K due to the constraint set, it is impossible to decompose blocks unlimitedly.

To derive the convex relaxation of $\psi_K(\boldsymbol{x})$, we exploit the following lemma which shows a variational representation of the ℓ_2 norm.

Lemma 1. *Define* $\phi\colon \mathbb{C} \times \mathbb{R} \to \mathbb{R}_+ \cup \{\infty\}$ *by*

$$\phi(x,\tau) := \begin{cases} \dfrac{|x|^2}{2\tau} + \dfrac{\tau}{2}, & \text{if } \tau > 0; \\ 0, & \text{if } x = 0 \text{ and } \tau = 0; \\ \infty, & \text{otherwise}, \end{cases} \tag{4}$$

which is a coercive lower semicontinuous convex function.

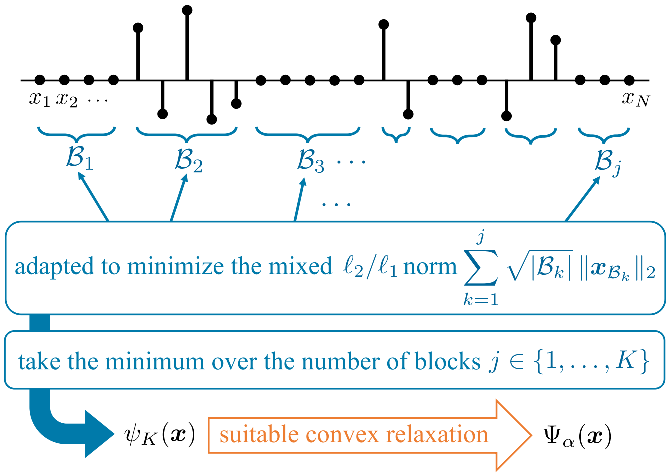

$x_1\, x_2\, \cdots$
x_N
$\mathcal{B}_1$
$\mathcal{B}_2$
$\mathcal{B}_3\ \cdots$
$\cdots$
$\mathcal{B}_j$
adapted to minimize the mixed ℓ_2/ℓ_1 norm $\sum_{k=1}^{j}\sqrt{|\mathcal{B}_k|}\,\|\boldsymbol{x}_{\mathcal{B}_k}\|_2$
take the minimum over the number of blocks $j \in \{1,\dots,K\}$
$\psi_K(\boldsymbol{x})$
suitable convex relaxation
$\Psi_\alpha(\boldsymbol{x})$

Fig. 2. Outline of design of the latent optimally partitioned (LOP)-ℓ_2/ℓ_1 penalty $\Psi_\alpha(\boldsymbol{x})$.

Then, the block-wise ℓ_2 norm is variationally represented as

$$\sqrt{|\mathcal{B}|}\|\boldsymbol{x}_{\mathcal{B}}\|_2 = \min_{\tau \in \mathbb{R}} \sum_{n \in \mathcal{B}} \phi(x_n, \tau), \tag{5}$$

for any nonempty $\mathcal{B} \subset \{1,\dots,N\}$, where the minimum is attained when $\tau = \|\boldsymbol{x}_{\mathcal{B}}\|_2/\sqrt{|\mathcal{B}|}$.

Proof. See Appendix B.

Remark 2. Lemma 1 is a slight modification of the existing result

$$\sqrt{|\mathcal{B}|}\|\boldsymbol{x}_{\mathcal{B}}\|_2 = \inf_{\tau > 0} \sum_{n \in \mathcal{B}} \left[\frac{|x_n|^2}{2\tau} + \frac{\tau}{2}\right]$$

used in, e.g., [49] for the analysis of the LGL penalty and [67] for the design of a directed-graph-sparse penalty. Since the domain of $\sum_{n\in\mathcal{B}} \left[\frac{|x_n|^2}{2\tau} + \frac{\tau}{2}\right]$ is not closed, this representation is inappropriate for the use in concrete optimization algorithms, as the applicability of the algorithm of [67] is limited [68]. Indeed, the study of [49] uses this representation only for the theoretical analysis of the LGL penalty but not for the design of the penalty and the corresponding optimization. Lemma 1 resolves this issue by using the lower semicontinuous convex function ϕ.

Applying Lemma 1 for each $\sqrt{|\mathcal{B}_k|}\|\boldsymbol{x}_{\mathcal{B}_k}\|_2$ in (2), we can rewrite $\psi_K(\boldsymbol{x})$ as

$$\psi_K(\boldsymbol{x}) = \min_{j \in \{1,\dots,K\}} \left[\min_{(\mathcal{B}_k)_{k=1}^{j} \in \mathcal{P}_j, \boldsymbol{\tau} \in \mathbb{R}^j} \sum_{k=1}^{j} \sum_{n \in \mathcal{B}_k} \phi(x_n, \tau_k)\right].$$

Note that the particular weight $\sqrt{|\mathcal{B}_k|}$ in (1) allows the application of Lemma 1. Introducing a latent vector $\boldsymbol{\sigma} = (\sigma_1,\dots,\sigma_N)^\top \in \mathbb{R}^N$ as

$$\sigma_n = \tau_k \quad (n \in \mathcal{B}_k) \qquad \text{for } k = 1,\dots,j,$$

as illustrated in Fig. 3, we see that $\boldsymbol{\sigma}$ is characterized by the condition that

$$\|\boldsymbol{D}\boldsymbol{\sigma}\|_0 \leq j-1,$$

[2] This inequality follows from the triangle inequality because $\bar{\boldsymbol{x}}^{\mathcal{B}} = \bar{\boldsymbol{x}}^{\mathcal{B}'} + \bar{\boldsymbol{x}}^{\mathcal{B}''}$ due to $\mathcal{B} = \mathcal{B}' \cup \mathcal{B}''$ and $\mathcal{B}' \cap \mathcal{B}'' = \varnothing$, where $\bar{\boldsymbol{x}}^{\mathcal{I}} \in \mathbb{R}^N$ is defined for $\mathcal{I} \subset \{1,\dots,N\}$ by $\bar{x}_n^{\mathcal{I}} = x_n$ if $n \in \mathcal{I}$ and $\bar{x}_n^{\mathcal{I}} = 0$ otherwise.

where $\boldsymbol{D}$ is the first difference operator defined by

$$\boldsymbol{D} := \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(N-1)\times N},$$

because $\mathcal{B}_k$ consists of only consecutive indices for each $k = 1,\ldots,j$. Moreover, from $\bigcup_{k=1}^{j} \mathcal{B}_k = \{1,\ldots,N\}$ and $\mathcal{B}_k \cap \mathcal{B}_{k'} = \varnothing \quad (k \neq k')$ in (3), we have

$$\sum_{k=1}^{j} \sum_{n\in\mathcal{B}_k} \phi(x_n,\tau_k) = \sum_{n=1}^{N} \phi(x_n,\sigma_n).$$

Thus, we can represent $\psi_K(\boldsymbol{x})$ as

$$\begin{aligned}\psi_K(\boldsymbol{x}) &= \min_{j\in\{1,\ldots,K\}} \left[\min_{\substack{\boldsymbol{\sigma}\in\mathbb{R}^N \\ \|\boldsymbol{D}\boldsymbol{\sigma}\|_0 \le j-1}} \sum_{n=1}^{N} \phi(x_n,\sigma_n) \right] \\ &= \min_{\substack{\boldsymbol{\sigma}\in\mathbb{R}^N \\ \|\boldsymbol{D}\boldsymbol{\sigma}\|_0 \le K-1}} \sum_{n=1}^{N} \phi(x_n,\sigma_n).\end{aligned}$$

Finally, based on the fact that the ℓ_1 norm is the tightest convex relaxation of the ℓ_0 pseudo-norm in the constraint, we derive the LOP-ℓ_2/ℓ_1 penalty as

$$\Psi_\alpha(\boldsymbol{x}) := \min_{\substack{\boldsymbol{\sigma}\in\mathbb{R}^N \\ \|\boldsymbol{D}\boldsymbol{\sigma}\|_1 \le \alpha}} \sum_{n=1}^{N} \phi(x_n,\sigma_n), \tag{6}$$

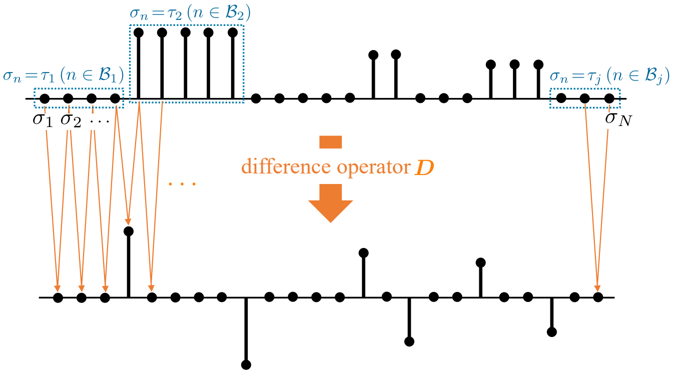
where $\alpha \in \mathbb{R}_+$ is a tuning parameter related to the number of blocks.

Theorem 1. *For any $\alpha \in \mathbb{R}_+$, Ψ_α takes finite values, i.e., $\mathrm{dom}(\Psi_\alpha) = \mathbb{C}^N$. Moreover, Ψ_α is coercive, continuous, nonnegative, and convex. Note that this implies $\Psi_\alpha \in \Gamma_0(\mathbb{C}^N)$.*

Proof. See Appendix C.

Intuitively, Ψ_α evaluates the block-sparsity of $\boldsymbol{x}$ with automatically adapted block partition as follows. Since the constraint $\|\boldsymbol{D}\boldsymbol{\sigma}\|_1 \le \alpha$ promotes the sparsity of $\boldsymbol{D}\boldsymbol{\sigma}$, $\boldsymbol{\sigma}$ is encouraged to be block-wise constant. At the same time, since $\phi(x,\tau)$ is basically $\frac{|x|^2}{2\tau} + \frac{\tau}{2}$ (see (4)), roughly speaking, minimizing $\sum_{n=1}^{N} \phi(x_n,\sigma_n)$ w.r.t. $\boldsymbol{\sigma}$ yields the weighted quadratic sum of $\boldsymbol{x}$ where larger components of $\boldsymbol{x}$ have smaller weights. These mechanisms make $\Psi_\alpha(\boldsymbol{x})$ the weighted quadratic sum of $\boldsymbol{x}$ with block-wise constant weights where blocks consisting of larger components have smaller weights. Thus, if large components of $\boldsymbol{x}$ are clustered in several blocks, the value of $\Psi_\alpha(\boldsymbol{x})$ becomes small. Note that the number of blocks is controlled by α. Thus, with a reasonable choice of α, decreasing the value of $\Psi_\alpha(\boldsymbol{x})$ can promote the block-sparsity of $\boldsymbol{x}$ with the adapted block partition. Remark that the computational difficulty is resolved because blocks are implicitly controlled by the vector $\boldsymbol{\sigma}$ of size N (see Remark 5 in Section III for the overall complexity of the proposed method).

As its special instances, the LOP-ℓ_2/ℓ_1 penalty reproduces the mixed ℓ_2/ℓ_1 norm with the coarsest partition and the finest partition. This is more precisely shown in the next theorem.



$\sigma_n = \tau_2\ (n\in\mathcal{B}_2)$
$\sigma_n = \tau_1\ (n\in\mathcal{B}_1)$
$\sigma_n = \tau_j\ (n\in\mathcal{B}_j)$
σ_1 σ_2 $\cdots$
σ_N
difference operator $\boldsymbol{D}$
$\cdots$

Fig. 3. Illustration of the latent vector $\boldsymbol{\sigma}$ and its difference $\boldsymbol{D}\boldsymbol{\sigma}$ for $j = 7$. It can be seen that $\boldsymbol{D}\boldsymbol{\sigma}$ has only $j-1 = 6$ nonzero components.

Theorem 2. *The LOP-ℓ_2/ℓ_1 penalty reproduces the mixed ℓ_2/ℓ_1 norm with the largest block $\{1,\ldots,N\}$ and the smallest blocks $(\{k\})_{k=1}^{N}$ respectively for $\alpha = 0$ and $\alpha \to \infty$, i.e.,*

$$\Psi_0(\boldsymbol{x}) = \|\boldsymbol{x}\|_{2,1}^{\{1,\ldots,N\}} = \sqrt{N}\|\boldsymbol{x}\|_2, \tag{7}$$

$$\lim_{\alpha\to\infty} \Psi_\alpha(\boldsymbol{x}) = \|\boldsymbol{x}\|_{2,1}^{(\{k\})_{k=1}^{N}} = \|\boldsymbol{x}\|_1. \tag{8}$$

Proof. See Appendix D.

This theorem also suggests that the tuning parameter α indeed controls the number of blocks in the LOP-ℓ_2/ℓ_1 penalty, though blocks do not explicitly appear in (6).

Remark 3 (How to choose α)**.** We show that a reasonable choice of α is to use

$$\alpha \ge 2\bar{\kappa}\bar{\sigma}, \tag{9}$$

where $\bar{\kappa}$ is the number of nonzero blocks and $\bar{\sigma}$ is the average of the standard deviation (i.e. the square root of the signal power) of nonzero blocks for the target signal:

$$\begin{aligned}\bar{\kappa} &:= |\mathcal{K}|, \\ \bar{\sigma} &:= \frac{1}{\bar{\kappa}} \sum_{k\in\mathcal{K}} \frac{\|\boldsymbol{x}^\star_{\mathcal{B}^\star_k}\|_2}{\sqrt{|\mathcal{B}^\star_k|}},\end{aligned}$$

where $\mathcal{K} := \{k \in \{1,\ldots,K^\star\} \mid \boldsymbol{x}^\star_{\mathcal{B}^\star_k} \neq \boldsymbol{0}\}$. We begin by computing an ideal $\boldsymbol{\sigma}^\star$ from $\boldsymbol{x}^\star$. Since $\boldsymbol{\sigma}^\star$ is desired to be constant for the block $\mathcal{B}^\star_k$, by considering the minimization of $\sum_{n\in\mathcal{B}^\star_k} \phi(x^\star_n,\sigma)$ w.r.t. σ, we have

$$\sigma^\star_n = \frac{\|\boldsymbol{x}^\star_{\mathcal{B}^\star_k}\|_2}{\sqrt{|\mathcal{B}^\star_k|}} \quad (n\in\mathcal{B}^\star_k) \qquad \text{for } k = 1,\ldots,K^\star,$$

from Lemma 1. For simplicity, suppose that $\mathcal{B}^\star_k$ and $\mathcal{B}^\star_{k+1}$ are consecutive blocks. Then, we have

$$\begin{aligned}\|\boldsymbol{D}\boldsymbol{\sigma}^\star\|_1 &= \sum_{k=1}^{K^\star-1} \left| \frac{\|\boldsymbol{x}^\star_{\mathcal{B}^\star_{k+1}}\|_2}{\sqrt{|\mathcal{B}^\star_{k+1}|}} - \frac{\|\boldsymbol{x}^\star_{\mathcal{B}^\star_k}\|_2}{\sqrt{|\mathcal{B}^\star_k|}} \right| \\ &\le 2\sum_{k=1}^{K^\star} \frac{\|\boldsymbol{x}^\star_{\mathcal{B}^\star_k}\|_2}{\sqrt{|\mathcal{B}^\star_k|}} = 2\bar{\kappa}\bar{\sigma}.\end{aligned}$$

Note that this inequality is expected to be tight because a nonzero block is usually adjacent to zero blocks. In order that $\boldsymbol{\sigma}^{\star}$ is contained in the constraint $\|\boldsymbol{D}\boldsymbol{\sigma}\|_1 \leq \alpha$ in (6), α should be chosen from the range (9). Note that, from our experiments shown in Section IV, the performance of the LOP-ℓ_2/ℓ_1 penalty seems fairly robust against the choice of α from the range slightly larger than $2\bar{\kappa}\bar{\sigma}$.

Remark 4 (Generalization to multi-dimensions)**.** For a matrix $\boldsymbol{X} \in \mathbb{C}^{N\times M}$, the LOP-$\ell_2/\ell_1$ penalty can be naturally extended by replacing the 1D difference operator with a 2D one:

$$\Psi_\alpha^{(2\mathrm{d})}(\boldsymbol{X}) := \min_{\substack{\boldsymbol{\Sigma}\in\mathbb{R}^{N\times M}\\ \|D_{2\mathrm{d}}(\boldsymbol{\Sigma})\|_1\leq\alpha}} \sum_{n=1}^{N}\sum_{m=1}^{M} \phi(X_{n,m}, \Sigma_{n,m}), \tag{10}$$

where $D_{2\mathrm{d}}\colon \mathbb{R}^{N\times M} \to \mathbb{R}^{(N-1)M+N(M-1)}$ computes differences in vertical and horizontal directions by

$$\begin{aligned} D_{2\mathrm{d}}(\boldsymbol{\Sigma}) &= \begin{bmatrix} D_{\mathrm{v}}(\boldsymbol{\Sigma}) \\ D_{\mathrm{h}}(\boldsymbol{\Sigma}) \end{bmatrix}, \\ D_{\mathrm{v}}(\boldsymbol{\Sigma}) &= \left(\Sigma_{n+1,m} - \Sigma_{n,m}\right)_{(n,m)\in\{1,\ldots,N-1\}\times\{1,\ldots,M\}}, \\ D_{\mathrm{h}}(\boldsymbol{\Sigma}) &= \left(\Sigma_{n,m+1} - \Sigma_{n,m}\right)_{(n,m)\in\{1,\ldots,N\}\times\{1,\ldots,M-1\}}. \end{aligned}$$

Note that we do not only focus on rectangular blocks, and $\Psi_\alpha^{(2\mathrm{d})}$ can manage blocks of various shapes because most components of $D_{2\mathrm{d}}(\boldsymbol{\Sigma})$ are zeros when $\boldsymbol{\Sigma}$ is constant on each block, irrespective of block shapes. The LOP-ℓ_2/ℓ_1 penalty can be naturally extended to signals in multi-dimensional spaces, e.g., pth-order tensors, by modifying the difference operator in similar ways.

III. Proposed Block-Sparse Recovery Model and Its Optimization Algorithm

We present a block-sparse recovery model using the LOP-ℓ_2/ℓ_1 penalty (6), and develop an iterative algorithm which converges to an optimal solution of the model. Specifically, we consider the following regularization model:

$$\operatorname*{minimize}_{\boldsymbol{x}\in\mathbb{C}^N} f(\boldsymbol{L}\boldsymbol{x}) + \lambda\Psi_\alpha(\boldsymbol{x}), \tag{11}$$

where $f(\boldsymbol{L}\boldsymbol{x})$ is some convex data-fidelity function with $f\colon \mathbb{C}^J \to \mathbb{R}_+$ and $\boldsymbol{L} \in \mathbb{C}^{J\times N}$, and $\lambda > 0$ is the regularization parameter. We suppose that $f \in \Gamma_0(\mathbb{C}^J)$, and its proximity operator can be computed efficiently. Such examples include the square error $\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$ and the absolute error $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1$, where $\boldsymbol{y}$ is the the known observation vector and $\boldsymbol{A}$ is the known measurement matrix. For these errors, we respectively set[3] $f(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{u}\|_2^2$ and $f(\boldsymbol{u}) = \|\boldsymbol{y}-\boldsymbol{u}\|_1$ for $\boldsymbol{u} = \boldsymbol{L}\boldsymbol{x}$ and $\boldsymbol{L} = \boldsymbol{A}$. The existence of the solution of (11) is guaranteed by the coercivity of $\Psi_\alpha \in \Gamma_0(\mathbb{C}^N)$, as shown in the next theorem.

Theorem 3. *For any $\lambda > 0$, the proposed model* (11) *has an optimal solution.*

Proof. It follows from Theorem 1 by [69, Corollary 11.16].

We show that an optimal solution of the proposed model (11) can be obtained as follows. By plugging the definition of $\Psi_\alpha(\boldsymbol{x})$ in (6) into (11), the optimization problem (11) is translated into

$$\left.\begin{aligned} \operatorname*{minimize}_{(\boldsymbol{x},\boldsymbol{\sigma})\in\mathbb{C}^N\times\mathbb{R}^N}\ & f(\boldsymbol{L}\boldsymbol{x}) + \lambda\varphi(\boldsymbol{x},\boldsymbol{\sigma}) \\ \text{subject to}\ & \|\boldsymbol{D}\boldsymbol{\sigma}\|_1 \leq \alpha \end{aligned}\right\}, \tag{12}$$

where we let

$$\varphi(\boldsymbol{x},\boldsymbol{\sigma}) := \sum_{n=1}^{N} \phi(x_n, \sigma_n).$$

Algorithm 1: Solver for the proposed model (11)

```
Input: $\gamma > 0$, $0 < \mu_1 \leq \frac{1}{\sqrt{\|\boldsymbol{L}\|_{\mathrm{op}}^2+1}}$, $0 < \mu_2 \leq \frac{1}{\sqrt{5}}$, and
       $\boldsymbol{x}^{(0)} \in \mathbb{C}^N, \boldsymbol{\sigma}^{(0)} \in \mathbb{R}^N, \boldsymbol{u}^{(0)} \in \mathbb{C}^J$,
       $\boldsymbol{\eta}^{(0)} \in \mathbb{R}^{N-1}, \boldsymbol{r}_1^{(0)} \in \mathbb{C}^J, \boldsymbol{r}_2^{(0)} \in \mathbb{R}^{N-1}$.
for $i = 0, 1, 2, \ldots$ do
    $\tilde{\boldsymbol{x}}^{(i+1)} = \boldsymbol{x}^{(i)} + \mu_1\boldsymbol{L}^{\mathrm{H}}(\boldsymbol{r}_1^{(i)} - \mu_1(\boldsymbol{L}\boldsymbol{x}^{(i)} - \boldsymbol{u}^{(i)}))$
    $\tilde{\boldsymbol{\sigma}}^{(i+1)} = \boldsymbol{\sigma}^{(i)} + \mu_2\boldsymbol{D}^{\top}(\boldsymbol{r}_2^{(i)} - \mu_2(\boldsymbol{D}\boldsymbol{\sigma}^{(i)} - \boldsymbol{\eta}^{(i)}))$
    $\tilde{\boldsymbol{u}}^{(i+1)} = \boldsymbol{u}^{(i)} - \mu_1(\boldsymbol{r}_1^{(i)} - \mu_1(\boldsymbol{L}\boldsymbol{x}^{(i)} - \boldsymbol{u}^{(i)}))$
    $\tilde{\boldsymbol{\eta}}^{(i+1)} = \boldsymbol{\eta}^{(i)} - \mu_2(\boldsymbol{r}_2^{(i)} - \mu_2(\boldsymbol{D}\boldsymbol{\sigma}^{(i)} - \boldsymbol{\eta}^{(i)}))$
    $(\boldsymbol{x}^{(i+1)}, \boldsymbol{\sigma}^{(i+1)}) = \mathrm{prox}_{\gamma\lambda\varphi}(\tilde{\boldsymbol{x}}^{(i+1)}, \tilde{\boldsymbol{\sigma}}^{(i+1)})$
                                                        // see (13) and (14)
    $\boldsymbol{u}^{(i+1)} = \mathrm{prox}_{\gamma f}(\tilde{\boldsymbol{u}}^{(i+1)})$          // see (15) or (16)
    $\boldsymbol{\eta}^{(i+1)} = P_{B_1^\alpha}(\tilde{\boldsymbol{\eta}}^{(i+1)})$          // see (17) and (18)
    $\boldsymbol{r}_1^{(i+1)} = \boldsymbol{r}_1^{(i)} - \mu_1(\boldsymbol{L}\boldsymbol{x}^{(i+1)} - \boldsymbol{u}^{(i+1)})$
    $\boldsymbol{r}_2^{(i+1)} = \boldsymbol{r}_2^{(i)} - \mu_2(\boldsymbol{D}\boldsymbol{\sigma}^{(i+1)} - \boldsymbol{\eta}^{(i+1)})$
```

The optimization problem (12) is convex but relatively difficult to solve because it involves the discontinuous function $\varphi(\boldsymbol{x},\boldsymbol{\sigma})$. This implies that a simple gradient descent strategy may fail to solve (12). We show that an iterative algorithm with guaranteed convergence to the solution of (12) can be developed by combining proximal splitting techniques [54]–[62] and theory of perspective functions [63]–[65]. As a concrete example, applying the primal-dual algorithm [54]–[59], we obtain the proposed algorithm shown in Algorithm 1 (see Appendix E for the derivation). The operators in Algorithm 1 can be computed as follows. Interpreting $\phi(x,\sigma)$ as the sum of *perspective function* [63] of $|x|^2/2$ and the linear function $\sigma/2$, based on [65, Example 2.4], we can compute the proximity operator of $\gamma\lambda\varphi$ as

$$\mathrm{prox}_{\gamma\lambda\varphi}(\boldsymbol{x},\boldsymbol{\sigma}) = \left(\mathrm{prox}_{\gamma\lambda\phi}(x_n,\sigma_n)\right)_{n=1}^{N}, \tag{13}$$

with

$$\begin{aligned} &\mathrm{prox}_{\gamma\lambda\phi}(x,\sigma) \\ &= \begin{cases} (0,0), & \text{if } 2\gamma\lambda\sigma + |x|^2 \leq \gamma^2\lambda^2; \\ (0, \sigma - \frac{\gamma\lambda}{2}), & \text{if } x = 0 \text{ and } 2\sigma > \gamma\lambda; \\ \left(x - \gamma\lambda s\frac{x}{|x|}, \sigma + \gamma\lambda\frac{s^2-1}{2}\right), & \text{otherwise,} \end{cases} \end{aligned} \tag{14}$$

[3] For the square error, we can also set as $f(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{u}\|_2^2$ and $\boldsymbol{L}$ to the identity matrix, which would be beneficial when, e.g., $\boldsymbol{A}^{\mathrm{H}}\boldsymbol{A}$ is diagonal.

where $s > 0$ is the unique positive root of

$$s^3 + \left(\frac{2}{\gamma\lambda}\sigma + 1\right)s - \frac{2}{\gamma\lambda}|x| = 0,$$

and can be explicitly given via Cardano's formula as follows. Let $p = \frac{2}{\gamma\lambda}\sigma + 1$, $q = -\frac{2}{\gamma\lambda}|x|$, and $D = -\frac{q^2}{4} - \frac{p^3}{27}$. Then,

$$s = \begin{cases} \sqrt[3]{-\frac{q}{2} + \sqrt{-D}} + \sqrt[3]{-\frac{q}{2} - \sqrt{-D}}, & \text{if } D < 0; \\ 2\sqrt[3]{-\frac{q}{2}}, & \text{if } D = 0; \\ 2\sqrt[3]{\sqrt{\frac{q^2}{4} + D}}\cos\left(\frac{\arctan(-2\sqrt{D}/q)}{3}\right), & \text{if } D > 0, \end{cases}$$

where $\sqrt[3]{\cdot}$ denotes the real cubic root. Note that the unique existence of the positive root follows from the uniqueness of the proximity operator [64, Theorem 3.1]. The proximity operator of γf depends on the employed data-fidelity function. For the square error $f(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}\|_2^2$, we have

$$\mathrm{prox}_{\frac{\gamma}{2}\|\boldsymbol{y}-\cdot\|_2^2}(\boldsymbol{u}) = \frac{\gamma\boldsymbol{y} + \boldsymbol{u}}{\gamma + 1}, \tag{15}$$

and for the absolute error $f(\boldsymbol{u}) = \|\boldsymbol{y} - \boldsymbol{u}\|_1$,

$$\mathrm{prox}_{\gamma\|\boldsymbol{y}-\cdot\|_1}(\boldsymbol{u}) = \left(u_j + \frac{\gamma}{\max\{|y_j - u_j|, \gamma\}}(y_j - u_j)\right)_{j=1}^{J}. \tag{16}$$

The ℓ_1 ball projection $P_{B_1^\alpha}$ can be computed as

$$P_{B_1^\alpha}(\boldsymbol{\eta}) = \begin{cases} \boldsymbol{\eta}, & \text{if } \|\boldsymbol{\eta}\|_1 \leq \alpha; \\ (a_n \mathrm{sign}(\eta_n))_{n=1}^{N-1}, & \text{otherwise}, \end{cases} \tag{17}$$

with

$$a_n := \max\left\{|\eta_n| - \frac{1}{T}\left(\sum_{t=1}^{T}\rho_t - \alpha\right), 0\right\}, \tag{18}$$

where $\rho_1, \ldots, \rho_{N-1}$ are obtained by sorting $|\eta_1|, \ldots, |\eta_{N-1}|$ in descending order, and

$$T := \max\left\{t \in \{1, \ldots, N-1\} \;\middle|\; \frac{1}{t}\left(\sum_{n=1}^{t}\rho_n - \alpha\right) < \rho_t\right\}.$$

The convergence of Algorithm 1 to an optimal solution of the proposed model (11) is confirmed as follows.

Theorem 4. *Choose $\gamma > 0$, μ_1, μ_2, and initial values $\boldsymbol{x}^{(0)}$, $\boldsymbol{\sigma}^{(0)}$, $\boldsymbol{u}^{(0)}$, $\boldsymbol{\eta}^{(0)}$, $\boldsymbol{r}_1^{(0)}$, $\boldsymbol{r}_2^{(0)}$ which satisfy the setting in Algorithm 1. Then, $(\boldsymbol{x}^{(i)})_{i=1}^{\infty}$ generated by Algorithm 1 converges to an optimal solution of* (11).

Proof. See Appendix F.

Since Algorithm 1 also solves (12), we obtain not only the solution for $\boldsymbol{x}$ but also the solution for $\boldsymbol{\sigma}$, say $\hat{\boldsymbol{\sigma}}$. Since $\hat{\boldsymbol{\sigma}}$ is the optimized latent vector of the LOP-ℓ_2/ℓ_1 penalty (6), we can recover the block partition from $\hat{\boldsymbol{\sigma}}$. More precisely, since $\hat{\boldsymbol{\sigma}}$ is expected to be block-wise constant, the endpoints of blocks are recovered from the positions of nonzero components of $\boldsymbol{D}\hat{\boldsymbol{\sigma}}$ (see Figs. 5 and 6 in Section IV for numerical examples).

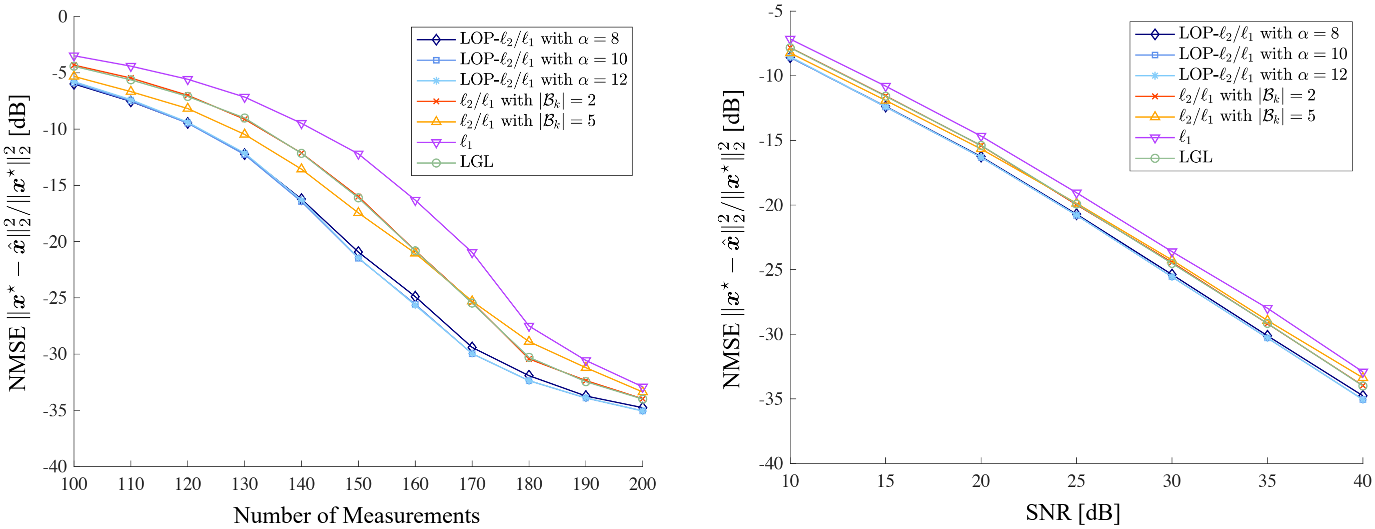
Remark 5 (Computational complexity)**.** We summarize the computational cost of Algorithm 1 per an iteration. We can compute $\tilde{\boldsymbol{x}}^{(i+1)}$, $\tilde{\boldsymbol{u}}^{(i+1)}$, and $\boldsymbol{r}_1^{(i+1)}$ with $\mathcal{O}(JN)$ operations, and $\tilde{\boldsymbol{\sigma}}^{(i+1)}$, $\tilde{\boldsymbol{\eta}}^{(i+1)}$, and $\boldsymbol{r}_2^{(i+1)}$ with $\mathcal{O}(N)$ operations since $\boldsymbol{D}$ has only $2(N-1)$ nonzero entries. The proximity operator of $\gamma\lambda\varphi$ can be computed by (13) and (14) with $\mathcal{O}(N)$ operations. For both the square error and the absolute error, $\mathrm{prox}_{\gamma f}$ can be computed with $\mathcal{O}(J)$ operations by (15) or (16). While the computation of the ℓ_1 ball projection $P_{B_1^\alpha}$ by (17) and (18) requires $\mathcal{O}(N\log N)$ operations for sorting, we can use sophisticated ℓ_1 ball projection algorithms having $\mathcal{O}(N)$ expected complexity, e.g., [70]. Summarizing, Algorithm 1 has $\mathcal{O}(N)$ complexity per an iteration. Thus, the proposed method has the same computational complexity as the ℓ_1 regularization method and the mixed ℓ_2/ℓ_1 regularization method using the non-overlapping blocks, when the primal-dual algorithm (26) in Appendix E is applied. The computational complexity of the LGL method depends on the design of overlapping blocks (see Appendix A). If all possible blocks are used, the computational complexity of the LGL method is $\mathcal{O}(N^3)$, which is significantly higher than the proposed method. Even when blocks are restricted to the overlapping blocks of a fixed size B, since the number of variables in the LGL model is $\mathcal{O}(BN)$, the computational complexity of the LGL method is about B-time as high as the proposed method. Note that the nonconvex methods [50]–[53] have $\mathcal{O}(N^3)$ computational complexity, which is significantly more expensive than the convex methods. In addition, it should be noted that the nonconvex methods have the problem of the dependence on initial values.

IV. Numerical Experiments

To show the effectiveness of the proposed block-sparse recovery model (11) using the LOP-ℓ_2/ℓ_1 penalty (6), we conduct numerical experiments on both synthetic and real-world data. We compare the LOP-ℓ_2/ℓ_1 model with existing convex penalties exploiting the block-sparsity: the mixed ℓ_2/ℓ_1 penalty (1) and the LGL penalty, shown in (23) of Appendix A, which is the state-of-the-art of structured sparse penalty. We also show the ℓ_1 penalty for reference of the non-structured sparse penalty. Note that we here do not include the nonconvex methods [50]–[53] because their computational complexity is significantly higher than the convex methods (see Remark 5) and the settings of initial values need separate discussion.

A. Synthetic Examples

We consider the estimation of a block-sparse signal $\boldsymbol{x}^\star \in \mathbb{R}^N$ from noisy compressive measurements. More precisely, we define the measurements by $\boldsymbol{y} := \boldsymbol{A}\boldsymbol{x}^\star + \boldsymbol{\varepsilon}$, where the entries of $\boldsymbol{A} \in \mathbb{R}^{d\times N}$ $(d < N)$ are drawn from i.i.d. Gaussian distribution $\mathcal{N}(0, 1)$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ is the white Gaussian noise. The block-sparse signal $\boldsymbol{x}^\star$ having 4 nonzero blocks and 80 nonzero components with $N = 250$ is randomly generated by the following scheme. The sizes of 4 nonzero blocks are randomly determined under the condition that the sum of the sizes is 80. The nonzero blocks are randomly located without



(a) NMSE against the number of measurements where SNR is fixed to 40dB.

(b) NMSE against SNR where the number of measurements is fixed to 200.

Fig. 4. Comparison of the regularization models for the recovery of block-sparse signals from noisy compressive measurements, where the results are averaged over 100 independent trials.

overlap, i.e., under the condition $|\mathrm{supp}(\boldsymbol{x}^\star)| = 80$. The zero blocks are automatically determined from the locations of the nonzero blocks. Amplitudes of nonzero components are drawn from i.i.d. $\mathcal{N}(0, 1)$. Note that we randomly generate the block partition for each trial to investigate the average performance for various block partitions.

In the proposed LOP-ℓ_2/ℓ_1 model (11), we use the square error by setting $f(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}\|_2^2$ and $\boldsymbol{L} = \boldsymbol{A}$. The existing penalties are also combined with the square error in similar ways. Note that the regularization parameter is tuned independently for each model to obtain the best accuracy. For the proposed model, the regularization parameter λ is tuned after the selection of α, though the optimal λ is almost the same for α used in our experiments. The proposed model (11) is solved by Algorithm 1, where we set $\gamma = 10^{-1}$, $\mu_1 = \frac{1}{\sqrt{\|\boldsymbol{A}\|_{\mathrm{op}}^2+1}}$, and $\mu_2 = \frac{1}{\sqrt{5}}$ which satisfy the condition for the convergence to an optimal solution. The existing models are also solved by applying the primal-dual algorithm (26) in Appendix E. We terminate the iteration when the norm of the differences between the variables of successive iterates is below the threshold 10^{-4}.

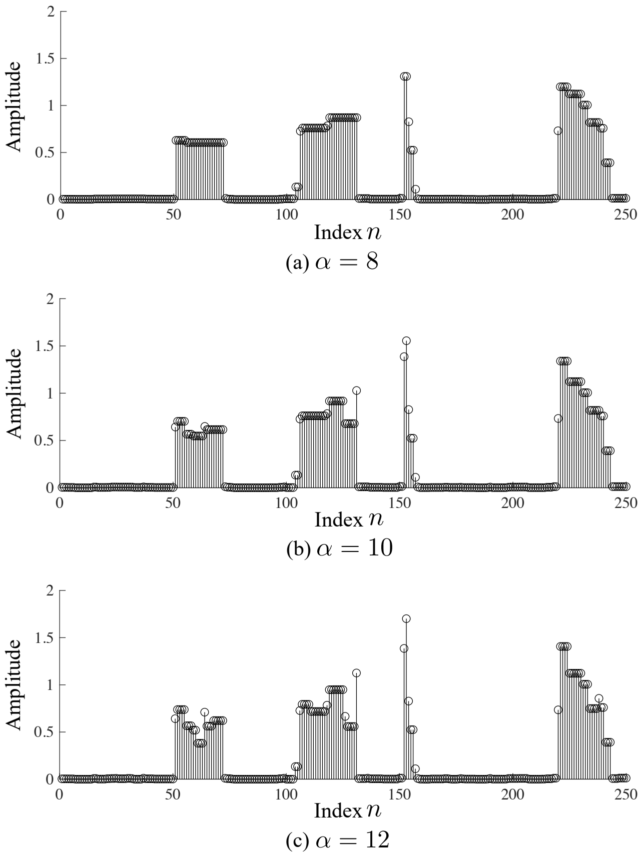
We compare the LOP-ℓ_2/ℓ_1 model and the existing models in terms of the normalized mean square error (NMSE)

$$\frac{\|\boldsymbol{x}^\star - \hat{\boldsymbol{x}}\|_2^2}{\|\boldsymbol{x}^\star\|_2^2},$$

where $\hat{\boldsymbol{x}}$ is the solution of each model. In Figs. 4(a) and (b), we show the NMSE respectively against the number d of measurements and the SNR $E[\|\boldsymbol{A}\boldsymbol{x}^\star\|_2^2]/E[\|\boldsymbol{\varepsilon}\|_2^2]$, where the results are averaged over 100 independent trials. Note that the variance of each component of $\boldsymbol{\varepsilon}$ is determined after the generation of $\boldsymbol{A}$ and $\boldsymbol{x}^\star$ according to the specified SNR. The mixed ℓ_2/ℓ_1 model is tested for block sizes 2 and 5, and the LGL model uses overlapping blocks of size 2 in (23). Different block sizes are given to the mixed ℓ_2/ℓ_1 model to see the performance dependency on the block size. More precisely, the block sizes 2 and 5 are chosen for the best performances respectively for the cases where the numbers of measurements are large and small. From Figs. 4(a) and (b), we see that the LOP-ℓ_2/ℓ_1 model outperforms the existing models. Since $2\bar{\kappa}\bar{\sigma}$ in (9) of Remark 3 is approximately 8 in this experiment,[4] we see that the performance of the LOP-ℓ_2/ℓ_1 model is fairly robust against the choice of the tuning parameter α, if α is chosen from the range slightly larger than $2\bar{\kappa}\bar{\sigma}$.

For the proposed method, we also examine the performance of the block partition estimation. Since the proposed LOP-ℓ_2/ℓ_1 model (11) is solved as (12), we also obtain $\hat{\boldsymbol{\sigma}}$ the solution of (12) for the variable $\boldsymbol{\sigma}$. Since $\hat{\boldsymbol{\sigma}}$ is expected to take constant values for the same block, the blocks can be recovered by checking the values of $\hat{\boldsymbol{\sigma}}$. Fig. 5 shows $\hat{\boldsymbol{\sigma}}$ obtained for an example of trials where the number of measurements is 200 and the SNR is 40dB. In Fig. 6, we show the block partition computed from the nonzero positions of $\boldsymbol{D}\boldsymbol{\sigma}$, where we discard the components of $\boldsymbol{D}\boldsymbol{\sigma}$ whose magnitudes are smaller than the threshold 0.3 for the visibility of the result. For reference, we also show the ground-truth signal $\boldsymbol{x}^\star$ in Fig. 6. The proposed method correctly estimates the zero blocks, but tends to divide the nonzero blocks into smaller blocks. Since nonzero components are randomly generated in our experiments, nonzero components generated for the same block might have a slight deviation in the magnitudes (see, e.g., the last nonzero block from the left in Fig. 6(a)). Thus, it can be said that the blocks estimated by the proposed method adequately reflect the property of the generated ground-truth signal. We also see that, increasing α, the nonzero blocks are divided into smaller blocks, while the zero blocks are kept almost the same. Since nonzero components in the same block have a slight deviation, the division of nonzero blocks

[4] This is because the number $\bar{\kappa}$ of nonzero blocks is set to 4, and the sample standard deviation $\bar{\sigma}$ is expected to be close to 1 since the nonzero components are drawn from $\mathcal{N}(0, 1)$.

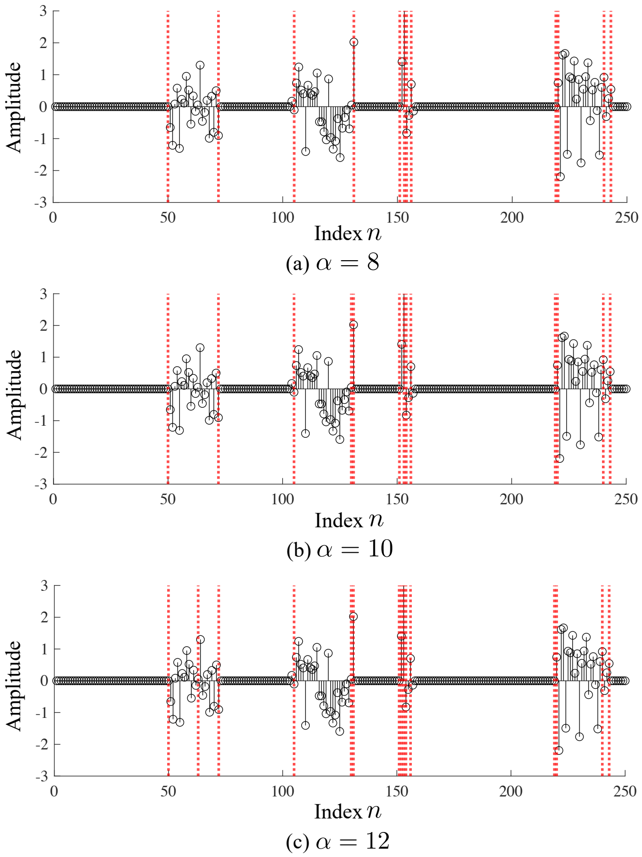


(a) $\alpha = 8$

(b) $\alpha = 10$

(c) $\alpha = 12$

Fig. 5. The latent vector $\hat{\boldsymbol{\sigma}}$ optimized in the LOP-ℓ_2/ℓ_1 penalty $\Psi_\alpha(\boldsymbol{x})$ for a trial of experiments where SNR and the number of measurements are respectively 40dB and 200.



(a) $\alpha = 8$

(b) $\alpha = 10$

(c) $\alpha = 12$

Fig. 6. Block partition computed from $\hat{\boldsymbol{\sigma}}$ shown in Fig. 5, where the endpoints of blocks are shown in red dot lines. The ground-truth signal $\boldsymbol{x}^\star$ is shown for reference.

is harmless for the estimation of $\boldsymbol{x}^\star$. This is hypothesized to be a reason of the robustness of the proposed method against the choice of α in terms of the estimation accuracy of $\boldsymbol{x}^\star$.

Lastly, we investigate how close the proposed LOP-ℓ_2/ℓ_1 model is to the mixed ℓ_2/ℓ_1 model using the *ground-truth* blocks $\mathcal{B}_1^\star, \ldots, \mathcal{B}_{K^\star}^\star$, though which are not available in practice. For simplicity, we here fix the ground-truth blocks as $\mathcal{B}_1^\star = \{1, \ldots, 50\}$, $\mathcal{B}_2^\star = \{51, \ldots, 50 + B^\star\}$, $\mathcal{B}_3^\star = \{51 + B^\star, \ldots, 150\}$, $\mathcal{B}_4^\star = \{151, \ldots, 150 + B^\star\}$, $\mathcal{B}_5^\star = \{151 + B^\star, \ldots, 250\}$, where only $\mathcal{B}_2^\star$ and $\mathcal{B}_4^\star$ contain nonzero components, and $B^\star$ is the size of these nonzero blocks. Note that $2\bar{\kappa}\bar{\sigma}$ in (9) of Remark 3 approximately equals to 4 in this setting. Fig. 7 shows the NMSE of the proposed model with several choices of α and the mixed ℓ_2/ℓ_1 model using the ground-truth blocks, where the results are averaged over 100 independent trials. In Fig. 7, we also show the ℓ_1 model and the ℓ_2 model, which respectively correspond to the LOP-ℓ_2/ℓ_1 model with $\alpha \to \infty$ and $\alpha = 0$ (see Theorem 2). Since the mixed ℓ_2/ℓ_1 model uses the ground-truth blocks in this experiment, it shows the best performance. Although the proposed model does not use the information on the ground-truth blocks, the proposed model with α slightly larger than $2\bar{\kappa}\bar{\sigma}$ shows comparable performance to the mixed ℓ_2/ℓ_1 model using the ground-truth blocks. This validates the effectiveness of the optimization of the block partition in the proposed method.

B. Application to Phased Array Weather Radar

A major goal of the phased array weather radar (PAWR) [40], [41] is to estimate the backscattered signal $\boldsymbol{X}^\star \in \mathbb{C}^{M \times N}$ from noisy observations $\boldsymbol{Y} := \boldsymbol{A}\boldsymbol{X}^\star + \boldsymbol{\mathcal{E}} \in \mathbb{C}^{d \times N}$ obtained by transmitting fan beams, where $\boldsymbol{A} \in \mathbb{C}^{d \times M}$ is a known array manifold matrix, $\boldsymbol{\mathcal{E}} \in \mathbb{C}^{d \times N}$ is the white Gaussian noise, and M, N, and d respectively are the numbers of elevation angles, pulses, and array elements. This estimation problem is called *beamforming* in the PAWR literature [37]–[39]. As shown in [38], [39], the backscattered signal $\boldsymbol{X}^\star$ exhibits block-sparsity in the Fourier domain for each elevation angle. Since the block partition depends on the unknown mean and standard deviation of the Doppler frequency at each elevation angle, it is also unknown and different for each elevation angle. Thus, the proposed approach is suitable for the PAWR beamforming. Specifically, we design a LOP-ℓ_2/ℓ_1 model for the PAWR beamforming as

$$\underset{\boldsymbol{X} \in \mathbb{C}^{M \times N}}{\text{minimize}} \ \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\|_{\text{fro}}^2 + \lambda \tilde{\Psi}_\alpha^{(\text{2d})}(\boldsymbol{F}\boldsymbol{X}^\top), \tag{19}$$

where $\|\cdot\|_{\text{fro}}$ denotes the Frobenius norm, and $\boldsymbol{F} \in \mathbb{C}^{N \times N}$ is the normalized discrete Fourier transform matrix. To exploit the column-wise block-sparsity of $\boldsymbol{F}\boldsymbol{X}^\top$, we use $\tilde{\Psi}_\alpha^{(\text{2d})}$ slightly modified from (10) by replacing D_{2d} with the vertical difference operator D_{v}. To apply Algorithm 1, we rewrite (19) in the form of (12) as follows. Since $\boldsymbol{F}$ is a unitary matrix, we have

$$\begin{aligned}\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{X}\|_{\mathrm{fro}}^2 &= \|\boldsymbol{Y}^\top-\boldsymbol{X}^\top\boldsymbol{A}^\top\|_{\mathrm{fro}}^2\\ &= \|\boldsymbol{F}\boldsymbol{Y}^\top-\boldsymbol{F}\boldsymbol{X}^\top\boldsymbol{A}^\top\|_{\mathrm{fro}}^2.\end{aligned}$$

Thus, by letting $\boldsymbol{Z}=\boldsymbol{F}\boldsymbol{X}^\top$ and substituting the definition of $\tilde{\Psi}_\alpha^{(2\mathrm{d})}$, the optimization problem (19) can be translated into

$$\left.\begin{aligned}\underset{(\boldsymbol{Z},\boldsymbol{\Sigma})\in\mathbb{C}^{N\times M}\times\mathbb{R}^{N\times M}}{\text{minimize}}\ \frac{1}{2}\|\boldsymbol{F}\boldsymbol{Y}^\top-\boldsymbol{Z}\boldsymbol{A}^\top\|_{\mathrm{fro}}^2+\lambda\widetilde{\varphi}(\boldsymbol{Z},\boldsymbol{\Sigma})\\ \text{subject to } \|D_{\mathrm{v}}(\boldsymbol{\Sigma})\|_1\leq\alpha\end{aligned}\right\}, \tag{20}$$

where

$$\widetilde{\varphi}(\boldsymbol{Z},\boldsymbol{\Sigma}):=\sum_{n=1}^{N}\sum_{m=1}^{M}\phi(Z_{n,m},\Sigma_{n,m}).$$

Further, by defining $\bar{\boldsymbol{A}}\in\mathbb{C}^{dN\times NM}$ and $\bar{\boldsymbol{D}}_{\mathrm{v}}\in\mathbb{R}^{(N-1)M\times NM}$ satisfying[5]

$$\begin{aligned}(\forall\boldsymbol{Z}\in\mathbb{C}^{N\times M})\quad &\mathrm{vec}(\boldsymbol{Z}\boldsymbol{A}^\top)=\bar{\boldsymbol{A}}\mathrm{vec}(\boldsymbol{Z}),\\ (\forall\boldsymbol{\Sigma}\in\mathbb{R}^{N\times M})\quad &D_{\mathrm{v}}(\boldsymbol{\Sigma})=\bar{\boldsymbol{D}}_{\mathrm{v}}\mathrm{vec}(\boldsymbol{\Sigma}),\end{aligned}$$

and $\bar{\boldsymbol{y}}:=\mathrm{vec}(\boldsymbol{F}\boldsymbol{Y}^\top)\in\mathbb{C}^{dN}$, we can rewrite (20) as

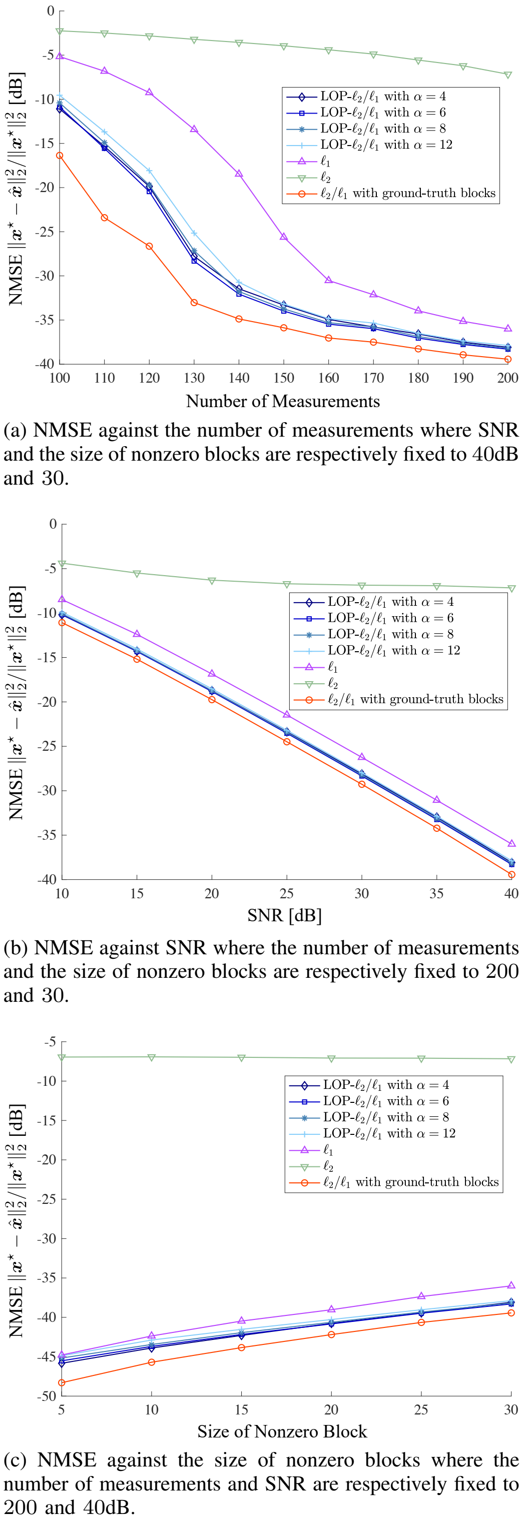
$$\left.\begin{aligned}\underset{(\bar{\boldsymbol{z}},\bar{\boldsymbol{\sigma}})\in\mathbb{C}^{NM}\times\mathbb{R}^{NM}}{\text{minimize}}\ \frac{1}{2}\|\bar{\boldsymbol{y}}-\bar{\boldsymbol{A}}\bar{\boldsymbol{z}}\|_2^2+\lambda\varphi(\bar{\boldsymbol{z}},\bar{\boldsymbol{\sigma}})\\ \text{subject to } \|\bar{\boldsymbol{D}}_{\mathrm{v}}\bar{\boldsymbol{\sigma}}\|_1\leq\alpha\end{aligned}\right\}. \tag{21}$$

Since (21) is in the form of (12), we can solve (21) by Algorithm 1 with setting[6] $f(\bar{\boldsymbol{z}})=\frac{1}{2}\|\bar{\boldsymbol{y}}-\bar{\boldsymbol{A}}\bar{\boldsymbol{z}}\|_2^2$, $\boldsymbol{L}=\boldsymbol{I}$, and $\boldsymbol{D}=\bar{\boldsymbol{D}}_{\mathrm{v}}$. From $\hat{\boldsymbol{z}}$ the solution of (21), we obtain the solution of (19) by $\hat{\boldsymbol{X}}=(\boldsymbol{F}^{\mathrm{H}}\mathrm{vec}^{-1}(\hat{\boldsymbol{z}}))^\top$, where vec^{-1} is the inverse of the vectorization operator. Existing penalties are also combined with the square error and solved in similar ways, where the regularization parameters are tuned independently to obtain the best results.

We conduct a numerical simulation in a setting similar to [37]–[39]. The backscattered signal $\boldsymbol{X}^\star\in\mathbb{C}^{M\times N}$ is generated based on the real reflection intensity measured by the PAWR at Osaka University,[7] where the elevation angles are chosen uniformly from $-15°$ to $30°$ degrees with $M=110$, and the mean and the standard deviation of the Doppler frequency are respectively drawn from uniform distributions of $[-1.25,1.25]$ and $[0.0314,0.1257]$ in kHz for each elevation angle. We set $N=50$ and $d=128$, and $\boldsymbol{\mathcal{E}}$ as the white Gaussian noise of the standard deviation $\sqrt{5}$. Table I shows the NMSE

$$\frac{\|\boldsymbol{X}^\star-\hat{\boldsymbol{X}}\|_{\mathrm{fro}}^2}{\|\boldsymbol{X}^\star\|_{\mathrm{fro}}^2},$$

averaged over 50 independent trails, where the LGL model uses overlapping blocks of size 2, and ℓ_2/ℓ_1 (a) and (b) respectively refer to the mixed ℓ_2/ℓ_1 models using block sizes 2 and 4. The result shows that the LOP-ℓ_2/ℓ_1 model achieves better estimation accuracy than the existing models, and is quite robust against the choice of the tuning parameter α. In Fig. 8, we show examples of the power spectra of the ground-truth and the estimates, i.e., squared magnitudes of entries of $\boldsymbol{F}\boldsymbol{X}^{\star\top}$ and $\boldsymbol{F}\hat{\boldsymbol{X}}^\top$. It can be seen that the LOP-ℓ_2/ℓ_1 model



(a) NMSE against the number of measurements where SNR and the size of nonzero blocks are respectively fixed to 40dB and 30.

(b) NMSE against SNR where the number of measurements and the size of nonzero blocks are respectively fixed to 200 and 30.

(c) NMSE against the size of nonzero blocks where the number of measurements and SNR are respectively fixed to 200 and 40dB.

Fig. 7. Comparison of the LOP-ℓ_2/ℓ_1 model against the mixed ℓ_2/ℓ_1 model using the ground-truth blocks, where the ℓ_1 model and the ℓ_2 model are also shown for reference of the LOP-ℓ_2/ℓ_1 model with $\alpha\to\infty$ and $\alpha=0$.

[5] We define the vectorization operator by $\mathrm{vec}(\boldsymbol{V}):=(\boldsymbol{v}_1^\top,\ldots,\boldsymbol{v}_q^\top)^\top\in\mathbb{C}^{pq}$ for $\boldsymbol{V}:=(\boldsymbol{v}_1,\ldots,\boldsymbol{v}_q)\in\mathbb{C}^{p\times q}$.

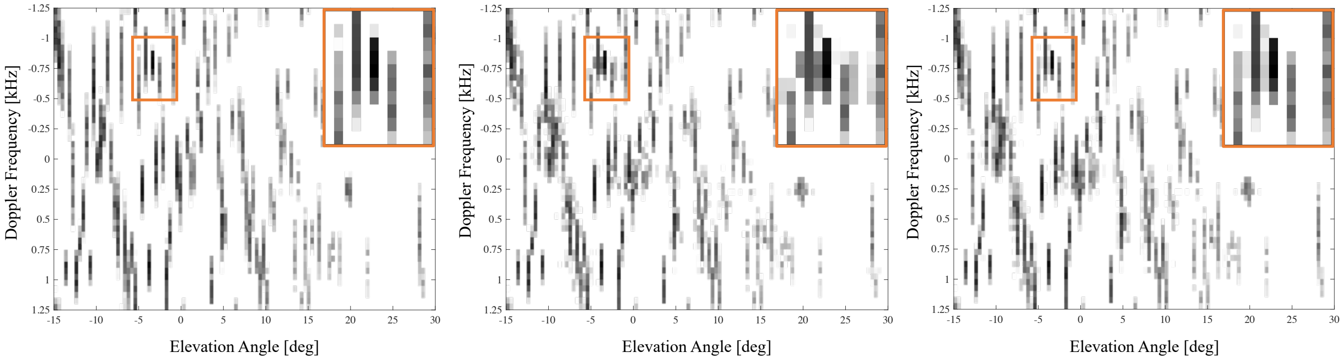
[6] We use this setting because $\bar{\boldsymbol{A}}$ is a block-diagonal matrix. The proximity operator of $f(\bar{\boldsymbol{z}})=\frac{1}{2}\|\bar{\boldsymbol{y}}-\bar{\boldsymbol{A}}\bar{\boldsymbol{z}}\|_2^2$ can be computed as $\mathrm{prox}_{\frac{\gamma}{2}\|\bar{\boldsymbol{y}}-\bar{\boldsymbol{A}}\cdot\|_2^2}(\bar{\boldsymbol{z}})=(\gamma\bar{\boldsymbol{A}}^{\mathrm{H}}\bar{\boldsymbol{A}}+\boldsymbol{I})^{-1}(\gamma\bar{\boldsymbol{A}}^{\mathrm{H}}\bar{\boldsymbol{y}}+\bar{\boldsymbol{z}})$.

[7] The PAWR data is courtesy of Prof. Tomoo Ushio, Dr. Hiroshi Kikuchi, and Dr. Eiichi Yoshikawa.

TABLE I

COMPARISON OF THE REGULARIZATION MODELS FOR SIMULATION OF PHASED ARRAY WEATHER RADAR IN TERMS OF NMSE IN dB AVERAGED OVER 50 INDEPENDENT TRIALS.

LOP-ℓ_2/ℓ_1 $(\alpha = 120M)$	LOP-ℓ_2/ℓ_1 $(\alpha = 130M)$	LOP-ℓ_2/ℓ_1 $(\alpha = 140M)$	ℓ_2/ℓ_1 (a)	ℓ_2/ℓ_1 (b)	ℓ_1	LGL
-15.62	$\mathbf{-15.72}$	-15.71	-14.78	-14.03	-13.39	-14.83



(a) Ground-truth

(b) Estimate by LGL model

(c) Estimate by LOP-ℓ_2/ℓ_1 model

Fig. 8. Examples of power spectra of ground-truth and estimates of the backscattered signal, shown in dB, for phased array weather radar simulation.

reduces the artifacts caused in the LGL model, e.g., in the zoomed areas of Fig. 8.

C. Application of Proposed 2D Penalty to Speech Denoising

To show the effectiveness of the 2D extension (10) of the LOP-ℓ_2/ℓ_1 penalty, we conduct experiments on the speech denoising exploiting block-sparsity of the spectrogram. We generate noisy speech as $\boldsymbol{y} := \boldsymbol{s}^\star + \boldsymbol{\varepsilon} \in \mathbb{R}^d$, where $\boldsymbol{s}^\star$ is a 2-second clip of speech taken from [71] with 16kHz sampling rate (i.e. $d = 32000$), and $\boldsymbol{\varepsilon}$ is the white Gaussian noise. The spectrogram $\boldsymbol{X}^\star := \mathcal{F}(\boldsymbol{s}^\star)$ of the clean speech, where $\mathcal{F}$ denotes the short-time Fourier transform (STFT), is expected to be block-sparse because entries of $\boldsymbol{X}^\star$ in adjacent frequency bins and frames tend to be zeros or nonzeros together. The block partition depends on the characteristics of the speech to be recovered, and thus is unknown a priori. Hence, the proposed approach is suitable for exploiting the block-sparsity of the spectrogram. Specifically, we apply the 2D LOP-ℓ_2/ℓ_1 penalty in (10) to the speech denoising as

$$\hat{\boldsymbol{X}} \in \arg\min_{\boldsymbol{X}\in\mathbb{C}^{N\times M}} \frac{1}{2}\|\boldsymbol{y} - \mathcal{F}^{-1}(\boldsymbol{X})\|_2^2 + \lambda\Psi_\alpha^{(2\mathrm{d})}(\boldsymbol{X}), \quad (22)$$

where N and M are respectively the numbers of frequency bins and frames, and $\mathcal{F}^{-1}$ denotes the inverse STFT. We implement the STFT as a *semi-orthogonal transformation* by using the hann window of 32ms (i.e. 512 samples) with 75% overlap and appropriate zero padding. Note that this setting implies $N = 512$ and $M = 253$. Since $D_{2\mathrm{d}}$ and $\mathcal{F}^{-1}$ are linear operators, we can solve (22) by Algorithm 1 through reformulation with the vectorization of $\boldsymbol{X}$, similarly to the reformulation of (20) into (21), where we should choose $\mu_2 \leq 1/3$ according to the operator norm of $D_{2\mathrm{d}}$. Note that $\mathcal{F}^{-1}$ can be computed as $\mathcal{F}^*$, the adjoint of $\mathcal{F}$, thanks to the semi-orthogonality of $\mathcal{F}$. The existing penalties are used in manners similar to (22), where the regularization parameters are tuned independently to obtain the best NMSE.

Experiments are conducted for male and female speech with input SNRs of 5dB and 10dB. In Table II, we show the NMSE

$$\frac{\|\boldsymbol{s}^\star - \mathcal{F}^{-1}(\hat{\boldsymbol{X}})\|_2^2}{\|\boldsymbol{s}^\star\|_2^2},$$

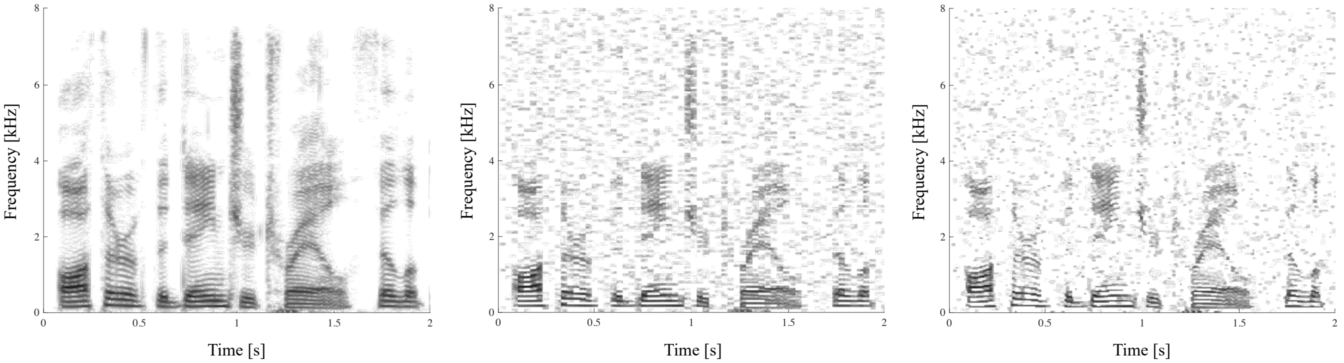
averaged over 20 independent trials, where ℓ_2/ℓ_1 (a) and (b) respectively denote the mixed ℓ_2/ℓ_1 models using block sizes 2×4 and 2×2, which yield best results among block sizes $\{1, 2, 4, 8\} \times \{1, 2, 4, 8\}$. We here do not include the LGL model since the use of 2D overlapping blocks is computationally expensive. From Table II, we see that the 2D LOP-ℓ_2/ℓ_1 model achieves better estimation accuracy than the existing models for every experimental condition. In addition, the performance of the 2D LOP-ℓ_2/ℓ_1 model is fairly robust against the choice of α. Fig. 9 shows the magnitude spectrograms of the clean speech and the estimates, i.e., magnitudes of entries of $\boldsymbol{X}^\star$ and $\hat{\boldsymbol{X}}$, for the denoising of male speech with 10dB input SNR. Rectangular artifacts are observed in Fig. 9(b) which shows the magnitude spectrogram obtained by the mixed ℓ_2/ℓ_1 model with the block size 2×4. These artifacts are considered to be caused by the rectangular blocks used in the mixed ℓ_2/ℓ_1 norm. Note that, since the blocks are used explicitly, it is difficult to employ blocks of various shapes in the mixed ℓ_2/ℓ_1 norm due to the computational complexity. On the other hand, the 2D LOP-ℓ_2/ℓ_1 penalty can handle blocks of various shapes because the blocks are implicitly controlled with the 2D difference operator $D_{2\mathrm{d}}$ (see Remark 4). Indeed, from Fig. 9(c), we see that the 2D LOP-ℓ_2/ℓ_1 model considerably reduces such artifacts, and more accurately recovers detailed shapes of the spectrogram.

V. CONCLUSION

We presented a convex recovery method for block-sparse signals whose block partitions are not available a priori. We first introduce a nonconvex penalty function ψ_K in (2), where

TABLE II
COMPARISON OF THE REGULARIZATION MODELS FOR SPEECH DENOISING IN TERMS OF NMSE IN dB AVERAGED OVER 20 INDEPENDENT TRIALS.

	LOP-ℓ_2/ℓ_1 ($\alpha = 0.006NM$)	LOP-ℓ_2/ℓ_1 ($\alpha = 0.007NM$)	LOP-ℓ_2/ℓ_1 ($\alpha = 0.008NM$)	ℓ_2/ℓ_1 (a)	ℓ_2/ℓ_1 (b)	ℓ_1
Male, 5dB	$\mathbf{-12.36}$	-12.28	-12.23	-12.16	-11.98	-11.85
Male, 10dB	$\mathbf{-15.78}$	-15.73	-15.68	-15.56	-15.41	-15.28
Female, 5dB	-14.05	$\mathbf{-14.07}$	-14.06	-13.69	-13.49	-13.47
Female, 10dB	-17.37	$\mathbf{-17.43}$	-17.41	-17.04	-16.90	-16.85



(a) Clean speech (b) Estimate by mixed ℓ_2/ℓ_1 model (c) Estimate by LOP- ℓ_2/ℓ_1 model

Fig. 9. Examples of magnitude spectrograms of clean speech and estimates, shown in dB, for the denoising of male speech with the input SNR of 10dB.

the block partition is optimized for the signal of interest by minimizing the mixed ℓ_2/ℓ_1 norm. Then, as the tight convex relaxation of ψ_K, we derive the proposed latent optimally partitioned (LOP)-ℓ_2/ℓ_1 penalty Ψ_α in (6). For the block-sparse recovery model (11) using the LOP-ℓ_2/ℓ_1 penalty, we develop the iterative solver as Algorithm 1 which is guaranteed to converge to an optimal solution. Numerical experiments on both synthetic and real-world data illustrate the effectiveness of the proposed method.

APPENDIX A
BRIEF REVIEW OF CONVEX BLOCK-SPARSE PENALTIES USING OVERLAPPING BLOCKS

A simple modification of the mixed ℓ_2/ℓ_1 norm is to use, in (1), overlapping blocks $\bar{\mathcal{B}}_1, \ldots, \bar{\mathcal{B}}_{\bar{K}} \subset \{1, \ldots, N\}$ such that

$$\begin{cases} \bigcup_{k=1}^{\bar{K}} \bar{\mathcal{B}}_k = \{1, \ldots, N\}, \\ \bar{\mathcal{B}}_k \neq \varnothing \quad (k = 1, \ldots, \bar{K}), \\ \bar{\mathcal{B}}_k = \{n \in \{1, \ldots, N\} \mid \bar{n}_k \leq n \leq \bar{m}_k\} \\ \quad \text{for some } \bar{n}_k, \bar{m}_k \in \{1, \ldots, N\} \quad (k = 1, \ldots, \bar{K}), \end{cases}$$

where the non-overlapping condition $\bar{\mathcal{B}}_k \cap \bar{\mathcal{B}}_{k'} = \varnothing \quad (k \neq k')$ is omitted, e.g., as in [42]–[47]. However, this modification cannot remove the blocks containing both zero and nonzero components. To alleviate this drawback, the latent group lasso (LGL) method [48], [49] is presented to select relevant blocks from the pre-defined overlapping blocks $\bar{\mathcal{B}}_1, \ldots, \bar{\mathcal{B}}_{\bar{K}}$. Specifically, the LGL penalty is defined by[8]

$$\mathrm{LGL}(\boldsymbol{x}) := \min_{\substack{(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{\bar{K}}) \in \mathbb{C}^{N \times \bar{K}} \\ \boldsymbol{x} = \sum_{k=1}^{\bar{K}} \boldsymbol{w}_k, \\ \mathrm{supp}(\boldsymbol{w}_k) \subset \bar{\mathcal{B}}_k \quad (k=1,\ldots,\bar{K})}} \sum_{k=1}^{\bar{K}} \sqrt{|\bar{\mathcal{B}}_k|} \|\boldsymbol{w}_k\|_2. \tag{23}$$

The LGL penalty can also be written as

$$\mathrm{LGL}(\boldsymbol{x}) = \min_{\substack{\tilde{\boldsymbol{w}} \in \mathbb{C}^{\tilde{N}} \\ \boldsymbol{x} = \boldsymbol{G}\tilde{\boldsymbol{w}}}} \sum_{k=1}^{\bar{K}} \sqrt{|\bar{\mathcal{B}}_k|} \|\tilde{\boldsymbol{w}}_{\mathcal{I}_k}\|_2,$$

where $\tilde{N} = \sum_{k=1}^{\bar{K}} |\bar{\mathcal{B}}_k|$,

$$\mathcal{I}_k := \left\{ n \in \{1, \ldots, \tilde{N}\} \;\middle|\; 1 + \sum_{m=1}^{k-1} |\bar{\mathcal{B}}_m| \leq n \leq \sum_{m=1}^{k} |\bar{\mathcal{B}}_m| \right\},$$

$$\boldsymbol{G} := \begin{bmatrix} \boldsymbol{G}_1 & \cdots & \boldsymbol{G}_{\bar{K}} \end{bmatrix} \in \mathbb{R}^{N \times \tilde{N}},$$

$$\boldsymbol{G}_k := \begin{bmatrix} \boldsymbol{O}_{\bar{n}_k - 1, |\bar{\mathcal{B}}_k|} \\ \boldsymbol{I}_{|\bar{\mathcal{B}}_k|} \\ \boldsymbol{O}_{N - \bar{m}_k, |\bar{\mathcal{B}}_k|} \end{bmatrix} \in \mathbb{R}^{N \times |\bar{\mathcal{B}}_k|},$$

and $\boldsymbol{I}_p \in \mathbb{R}^{p \times p}$ and $\boldsymbol{O}_{p,q} \in \mathbb{R}^{p \times q}$ respectively denote the identity matrix of order p and the zero matrix of size $p \times q$. Thus, when the LGL penalty is combined with some convex data-fidelity function, the resulting regularization model can be solved as a convex optimization problem of $\tilde{\boldsymbol{w}} \in \mathbb{C}^{\tilde{N}}$. However, it is intractable to use all possible blocks as $\bar{\mathcal{B}}_1, \ldots, \bar{\mathcal{B}}_{\bar{K}}$ because the problem size becomes too large, i.e., $\tilde{N} = N(N+1)(N+2)/6$, in this case. Due to this issue, $\bar{\mathcal{B}}_1, \ldots, \bar{\mathcal{B}}_{\bar{K}}$ are typically restricted to blocks of a fixed size.

[8] By setting $\bar{\mathcal{B}}_k$ to other than blocks, the LGL method is applicable to other structured sparsity, while this paper focuses on the block-sparsity.

On the other hand, the proposed approach is free from such issue because the block partition is implicitly optimized in the proposed penalty (6) with the latent vector $\boldsymbol{\sigma}$ of size N.

Appendix B
Proof of Lemma 1

First, we prove the properties of ϕ defined by (4). Notice that ϕ can be written as

$$\phi(x,\tau)=\frac{h(x,\tau)}{2}+\frac{\tau}{2},$$

where

$$h(x,\tau):=\begin{cases}\frac{|x|^2}{\tau}, & \text{if } \tau>0;\\ 0, & \text{if } x=0 \text{ and } \tau=0;\\ \infty, & \text{otherwise.}\end{cases}$$

Since h is *perspective function* of $|x|^2$, it is lower semicontinuous convex [63, Proposition 2.3 and Example 2.6], and thus ϕ is lower semicontinuous convex. The coercivity of ϕ is proven by showing $\phi(x,\tau)\to\infty \quad (|x|+|\tau|\to\infty)$ as follows.

a) (Case of $|\tau|\to\infty$) Clear from $\phi(x,\tau)\geq\frac{\tau}{2}$.

b) (Case of $|\tau|$ is bounded) In this case, we have $|x|\to\infty$ and $\exists R>0$ such that $|\tau|\leq R$. Moreover, since $\phi(x,\tau)=\infty$ when $\tau\leq 0$ and $x\neq 0$, we have

$$(\forall x\neq 0)\quad \phi(x,\tau)\geq\frac{|x|^2}{2R},$$

which implies $\phi(x,\tau)\to\infty$ due to $|x|\to\infty$.

Next, we prove the equation (5) as follows.

a) (Case of $\boldsymbol{x}_{\mathcal{B}}\neq\mathbf{0}$) For $\tau>0$, we have

$$\sum_{n\in\mathcal{B}}\phi(x_n,\tau)=\sum_{n\in\mathcal{B}}\left(\frac{|x_n|^2}{2\tau}+\frac{\tau}{2}\right)=\frac{\|\boldsymbol{x}_{\mathcal{B}}\|_2^2}{2\tau}+\frac{\tau}{2}|\mathcal{B}|.$$

Since $\boldsymbol{x}_{\mathcal{B}}\neq\mathbf{0}$ in this case, $\frac{\|\boldsymbol{x}_{\mathcal{B}}\|_2^2}{2\tau}+\frac{|\mathcal{B}|}{2}\tau$ is a strictly convex function of $\tau>0$, and the minimum is attained when $\tau=\|\boldsymbol{x}_{\mathcal{B}}\|_2/\sqrt{|\mathcal{B}|}$. Thus, we have

$$\min_{\tau>0}\sum_{n\in\mathcal{B}}\phi(x_n,\tau)=\sqrt{|\mathcal{B}|}\|\boldsymbol{x}_{\mathcal{B}}\|_2.$$

For $\tau=0$, since there exists $n\in\mathcal{B}$ such that $x_n\neq 0$ in this case, we have $\sum_{n\in\mathcal{B}}\phi(x_n,0)=\infty$. For $\tau<0$, the definition of ϕ in (4) readily implies $\sum_{n\in\mathcal{B}}\phi(x_n,\tau)=\infty$. Summarizing, we have the equation (5) where the minimum is attained when $\tau=\|\boldsymbol{x}_{\mathcal{B}}\|_2/\sqrt{|\mathcal{B}|}$.

b) (Case of $\boldsymbol{x}_{\mathcal{B}}=\mathbf{0}$) For $\tau>0$, since $x_n=0$ for every $n\in\mathcal{B}$ in this case, we have

$$\sum_{n\in\mathcal{B}}\phi(x_n,\tau)=\sum_{n\in\mathcal{B}}\left(\frac{|x_n|^2}{2\tau}+\frac{\tau}{2}\right)=\frac{\tau}{2}|\mathcal{B}|>0.$$

For $\tau=0$, we have

$$\sum_{n\in\mathcal{B}}\underbrace{\phi(x_n,0)}_{=0}=0.$$

For $\tau<0$, we have $\sum_{n\in\mathcal{B}}\phi(x_n,\tau)=\infty$. Thus, the minimum is attained when $\tau=0$, and

$$\min_{\tau\in\mathbb{R}}\sum_{n\in\mathcal{B}}\phi(x_n,\tau)=0.$$

Since $\boldsymbol{x}_{\mathcal{B}}=\mathbf{0}$ in this case, this implies (5) where the minimum is attained when $\tau=0=\|\boldsymbol{x}_{\mathcal{B}}\|_2/\sqrt{|\mathcal{B}|}$.

Appendix C
Proof of Theorem 1

Since Ψ_α is defined by the partial minimization of the coercive lower semicontinuous convex function $\sum_{n=1}^N\phi(x_n,\sigma_n)$ under the closed convex constraint $\|\boldsymbol{D\sigma}\|_1\leq\alpha$, and

$$\left\{\boldsymbol{\sigma}\in\mathbb{R}^N\ \middle|\ \sum_{n=1}^N\phi(x_n,\sigma_n)<\infty \text{ and } \|\boldsymbol{D\sigma}\|_1\leq\alpha\right\}\neq\varnothing \tag{24}$$

for any $\alpha\in\mathbb{R}_+$, Ψ_α takes finite values for every $\boldsymbol{x}\in\mathbb{C}^N$, i.e., $\operatorname{dom}(\Psi_\alpha)=\mathbb{C}^N$, by [69, Proposition 11.15]. Note that (24) holds for $\alpha\in\mathbb{R}_+$ since $\boldsymbol{D\sigma}=\mathbf{0}$ for any $\boldsymbol{\sigma}$ satisfying $\sigma_1=\sigma_2=\cdots=\sigma_N$. The convexity of Ψ_α follows from the convexity of $\sum_{n=1}^N\phi(x_n,\sigma_n)$ and $\|\boldsymbol{D\sigma}\|_1\leq\alpha$ by [69, Proposition 8.35]. Since Ψ_α is convex and $\operatorname{dom}(\Psi_\alpha)=\mathbb{C}^N$, Ψ_α is continuous by [69, Corollary 8.40]. From $\sum_{n=1}^N\phi(x_n,\sigma_n)\geq 0$, Ψ_α is nonnegative. Lastly, the coercivity of Ψ_α follows from

$$\begin{aligned}\Psi_\alpha(\boldsymbol{x})&=\min_{\substack{\boldsymbol{\sigma}\in\mathbb{R}^N\\ \|\boldsymbol{D\sigma}\|_1\leq\alpha}}\sum_{n=1}^N\phi(x_n,\sigma_n)\\ &\geq\min_{\boldsymbol{\sigma}\in\mathbb{R}^N}\sum_{n=1}^N\phi(x_n,\sigma_n)\\ &=\sum_{n=1}^N\left[\min_{\sigma_n\in\mathbb{R}}\phi(x_n,\sigma_n)\right]=\|\boldsymbol{x}\|_1,\end{aligned}$$

where the last equality is obtained by applying Lemma 1 for each of $\min_{\sigma_n\in\mathbb{R}}\phi(x_n,\sigma_n)\quad(n=1,\ldots,N)$.

Appendix D
Proof of Theorem 2

The equation (7) is proven as

$$\begin{aligned}\Psi_0(\boldsymbol{x})&=\min_{\substack{\boldsymbol{\sigma}\in\mathbb{R}^N\\ \|\boldsymbol{D\sigma}\|_1\leq 0}}\sum_{n=1}^N\phi(x_n,\sigma_n)\\ &=\min_{\sigma\in\mathbb{R}}\sum_{n=1}^N\phi(x_n,\sigma)=\sqrt{N}\|\boldsymbol{x}\|_2,\end{aligned}$$

where the second equality holds due to

$$\|\boldsymbol{D\sigma}\|_1\leq 0\Leftrightarrow\boldsymbol{D\sigma}=\mathbf{0}\Leftrightarrow\sigma_1=\sigma_2=\cdots=\sigma_N,$$

and the last equality follows from Lemma 1. Next, (8) is proven as

$$\lim_{\alpha\to\infty}\Psi_\alpha(\boldsymbol{x})=\min_{\boldsymbol{\sigma}\in\mathbb{R}^N}\sum_{n=1}^N\phi(x_n,\sigma_n)=\|\boldsymbol{x}\|_1,$$

where the last equality holds by Lemma 1.

Appendix E
Primal-Dual Algorithm and Reformulation of (12)

The Chambolle-Pock primal-dual algorithm [54], [55] and the Loris-Verhoeven algorithm [56], [57] are widely used for solving convex optimization problems. As explained in [62], these algorithms are essentially the same for the following problem

$$\underset{\boldsymbol{v}\in\mathcal{V}}{\operatorname{minimize}}\, G(\boldsymbol{v}) \text{ subject to } \boldsymbol{H}\boldsymbol{v} = \boldsymbol{0}, \tag{25}$$

and the iteration can be written as

$$\left\lfloor\begin{aligned} \tilde{\boldsymbol{v}}^{(i+1)} &= \boldsymbol{v}^{(i)} + \boldsymbol{H}^*(\boldsymbol{r}^{(i)} - \boldsymbol{H}\boldsymbol{v}^{(i)}) \\ \boldsymbol{v}^{(i+1)} &= \operatorname{prox}_{\gamma G}(\tilde{\boldsymbol{v}}^{(i+1)}) \\ \boldsymbol{r}^{(i+1)} &= \boldsymbol{r}^{(i)} - \boldsymbol{H}\boldsymbol{v}^{(i+1)}, \end{aligned}\right. \tag{26}$$

where we suppose that $G \in \Gamma_0(\mathcal{V})$, $\boldsymbol{H}\colon \mathcal{V} \to \mathcal{W}$ is a linear operator, $\boldsymbol{H}^*\colon \mathcal{W} \to \mathcal{V}$ is the adjoint of $\boldsymbol{H}$, $\mathcal{V}$ and $\mathcal{W}$ are finite-dimensional Hilbert spaces, and $\gamma > 0$. This algorithm is also known as the linearized augmented Lagrangian algorithm [58], [59]. The convergence of (26) to an optimal solution of (25) is shown in the following lemma from, e.g., [62, Theorem 5.3] and [58, Corollary 1].

Lemma 2. *Suppose that an optimal solution of* (25) *exists, the so-called qualification condition*[9] $\boldsymbol{0} \in \operatorname{ri}(\boldsymbol{H}(\operatorname{dom}(G)))$ *holds,* $\|\boldsymbol{H}\|_{\mathrm{op}} \leq 1$*, and* $\gamma > 0$*. Let* $\boldsymbol{v}^{(0)} \in \mathcal{V}$ *and* $\boldsymbol{r}^{(0)} \in \mathcal{W}$*. Then,* $(\boldsymbol{v}^{(i)})_{i=1}^{\infty}$ *generated by* (26) *converges to an optimal solution of* (25).

We rewrite (12) into the form of (25) by introducing auxiliary variables $\boldsymbol{u} = \boldsymbol{L}\boldsymbol{x}$ and $\boldsymbol{\eta} = \boldsymbol{D}\boldsymbol{\sigma}$:

$$\left.\begin{aligned} &\underset{(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\eta})\in\mathbb{C}^N\times\mathbb{R}^N\times\mathbb{C}^J\times\mathbb{R}^{N-1}}{\operatorname{minimize}} \underbrace{f(\boldsymbol{u}) + \lambda\varphi(\boldsymbol{x},\boldsymbol{\sigma}) + \iota_{B_1^\alpha}(\boldsymbol{\eta})}_{=:G(\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\eta})} \\ &\text{subject to } \underbrace{\begin{bmatrix} \mu_1\boldsymbol{L} & \boldsymbol{O} & -\mu_1\boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \mu_2\boldsymbol{D} & \boldsymbol{O} & -\mu_2\boldsymbol{I} \end{bmatrix}}_{=:\boldsymbol{H}} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\sigma} \\ \boldsymbol{u} \\ \boldsymbol{\eta} \end{bmatrix} = \boldsymbol{0} \end{aligned}\right\}, \tag{27}$$

where $\iota_{B_1^\alpha}(\boldsymbol{\eta})$ is the indicator function of the ℓ_1 ball and defined by

$$\iota_{B_1^\alpha}(\boldsymbol{\eta}) := \begin{cases} 0, & \text{if } \|\boldsymbol{\eta}\|_1 \leq \alpha; \\ \infty, & \text{otherwise,} \end{cases}$$

$\boldsymbol{I}$ and $\boldsymbol{O}$ respectively denote the identity and zero matrices of appropriate sizes, and μ_1 and μ_2 satisfying

$$\left.\begin{aligned} 0 &< \mu_1 \leq \frac{1}{\sqrt{\|\boldsymbol{L}\|_{\mathrm{op}}^2 + 1}} \\ 0 &< \mu_2 \leq \frac{1}{\sqrt{5}} \left(< \frac{1}{\sqrt{\|\boldsymbol{D}\|_{\mathrm{op}}^2 + 1}} \right) \end{aligned}\right\} \tag{28}$$

are introduced so that $\|\boldsymbol{H}\|_{\mathrm{op}} \leq 1$. Applying the primal-dual algorithm (26) to the problem (27), we obtain the proposed algorithm shown as Algorithm 1 in Section III, where the updates are written in terms of $(\boldsymbol{x}, \boldsymbol{\sigma}, \boldsymbol{u}, \boldsymbol{\eta})$. Note that the proximity operator of γG in (26) reduces to the proximity operators of $\gamma\lambda\varphi$, γf, and $\gamma\iota_{B_1^\alpha}$, since γG is the separable sum of these functions. Note also that the proximity operator of $\gamma\iota_{B_1^\alpha}$ reduces to the ℓ_1 ball projection $P_{B_1^\alpha}$.

Appendix F
Proof of Theorem 4

First, the existence of the solution of (27) is guaranteed from Theorem 3 and the equivalence between (11), (12), and (27). The qualification condition for (27) is confirmed as follows. The effective domain of G is given by

$$\begin{aligned} \operatorname{dom}(G) &= \operatorname{dom}(\varphi) \times \operatorname{dom}(f) \times \operatorname{dom}(\iota_{B_1^\alpha}) \\ &= (\mathbb{C} \times \mathbb{R}_{++} \cup \{(0,0)\})^N \times \mathbb{C}^J \times B_1^\alpha, \end{aligned}$$

where $\mathbb{R}_{++}$ denotes the set of all positive real numbers, and $B_1^\alpha = \{\boldsymbol{\eta} \in \mathbb{R}^{N-1} \mid \|\boldsymbol{\eta}\|_1 \leq \alpha\}$. Thus, we have

$$\begin{aligned} \boldsymbol{H}(\operatorname{dom}(G)) &= \left\{ \begin{bmatrix} \mu_1\boldsymbol{L}\boldsymbol{x} - \mu_1\boldsymbol{u} \\ \mu_2\boldsymbol{D}\boldsymbol{\sigma} - \mu_2\boldsymbol{\eta} \end{bmatrix} \middle| (\boldsymbol{x},\boldsymbol{\sigma},\boldsymbol{u},\boldsymbol{\eta}) \in \operatorname{dom}(G) \right\} \\ &= (\boldsymbol{L}(\mathbb{C}^N) - \mathbb{C}^J) \times (\boldsymbol{D}(\mathbb{R}^N) - \mu_2 B_1^\alpha), \end{aligned} \tag{29}$$

where we use the relation $\boldsymbol{D}(\mathbb{R}_{++}^N) = \boldsymbol{D}(\mathbb{R}^N)$. This relation holds because, for any $\boldsymbol{\sigma} \in \mathbb{R}^N$, $\boldsymbol{D}(\boldsymbol{\sigma} + c\boldsymbol{1}) = \boldsymbol{D}\boldsymbol{\sigma}$ and $\boldsymbol{\sigma} + c\boldsymbol{1} \in \mathbb{R}_{++}^N$ can be satisfied with some $c \in \mathbb{R}_{++}$, where $\boldsymbol{1} := (1, 1, \ldots, 1)^\top \in \mathbb{R}_{++}^N$. Since the expression (29) implies that $\boldsymbol{H}(\operatorname{dom}(G))$ is a nonempty convex set satisfying the symmetry $\boldsymbol{H}(\operatorname{dom}(G)) = -\boldsymbol{H}(\operatorname{dom}(G))$, the qualification condition $\boldsymbol{0} \in \operatorname{ri}(\boldsymbol{H}(\operatorname{dom}(G)))$ holds by [69, Example 6.10]. The condition $\|\boldsymbol{H}\|_{\mathrm{op}} \leq 1$ is checked as follows. Since

$$\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}} = \begin{bmatrix} \mu_1^2(\boldsymbol{L}\boldsymbol{L}^{\mathrm{H}} + \boldsymbol{I}) & \boldsymbol{O} \\ \boldsymbol{O} & \mu_2^2(\boldsymbol{D}\boldsymbol{D}^\top + \boldsymbol{I}) \end{bmatrix}$$

is block-diagonal, with $\rho(\cdot)$ denoting the largest eigenvalue of the argument, we have

$$\begin{aligned} \|\boldsymbol{H}\|_{\mathrm{op}} &= \sqrt{\rho(\boldsymbol{H}\boldsymbol{H}^{\mathrm{H}})} \\ &= \max\left\{ \sqrt{\rho(\mu_1^2(\boldsymbol{L}\boldsymbol{L}^{\mathrm{H}} + \boldsymbol{I}))}, \sqrt{\rho(\mu_2^2(\boldsymbol{D}\boldsymbol{D}^\top + \boldsymbol{I}))} \right\} \\ &= \max\left\{ \mu_1\sqrt{\rho(\boldsymbol{L}\boldsymbol{L}^{\mathrm{H}}) + 1}, \mu_2\sqrt{\rho(\boldsymbol{D}\boldsymbol{D}^\top) + 1} \right\} \\ &= \max\left\{ \mu_1\sqrt{\|\boldsymbol{L}\|_{\mathrm{op}}^2 + 1}, \mu_2\sqrt{\|\boldsymbol{D}\|_{\mathrm{op}}^2 + 1} \right\} \leq 1, \end{aligned}$$

where the last inequality holds since μ_1 and μ_2 are chosen to satisfy (28). Thus, by Lemma 2, $(\boldsymbol{x}^{(i)}, \boldsymbol{\sigma}^{(i)}, \boldsymbol{u}^{(i)}, \boldsymbol{\eta}^{(i)})_{i=1}^{\infty}$ converges to an optimal solution of (27). The convergence of $(\boldsymbol{x}^{(i)})_{i=1}^{\infty}$ to the solution of (11) follows from the relation between (11), (12), and (27).

Acknowledgment

We would like to acknowledge helpful discussions with Dr. Akira Hirabayashi. We would also like to thank Prof. Tomoo Ushio, Dr. Hiroshi Kikuchi, and Dr. Eiichi Yoshikawa for providing us with the PAWR data. Finally, we would like to thank the anonymous reviewers for their valuable comments on the original version of the manuscript.

[9] For a set $S \subset \mathcal{V}$, $\boldsymbol{H}(S) := \{\boldsymbol{H}\boldsymbol{v} \in \mathcal{W} \mid \boldsymbol{v} \in S\}$. For a set $C \subset \mathcal{W}$, $\operatorname{ri}(C)$ denotes the relative interior of C (see, e.g., [69, Definition 6.9]).

References

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Review*, vol. 43, no. 1, pp. 129–159, 2001.

[3] R. G. Baraniuk, "Compressive sensing [lecture notes]," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 118–121, Jul. 2007.

[4] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[5] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[6] J. J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3601–3608, Oct. 2005.

[7] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.

[8] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[9] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.

[10] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, 2008.

[11] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Ann. Statist.*, vol. 38, no. 2, pp. 894–942, 2010.

[12] D. Ge, X. Jiang, and Y. Ye, "A note on the complexity of L_p minimization," *Math. Programm.*, vol. 129, no. 2, pp. 285–299, 2011.

[13] J. Yu, A. Eriksson, T.-J. Chin, and D. Suter, "An adversarial optimization approach to efficient outlier removal," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2011, pp. 399–406.

[14] A. Argyriou, R. Fygel, and N. Srebro, "Sparse prediction with the k-support norm," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1466–1474.

[15] P. Yin, Y. Lou, Q. He, and J. Xin, "Minimization of ℓ_{1-2} for compressed sensing," *SIAM J. Sci. Comput.*, vol. 37, no. 1, pp. 536–563, 2015.

[16] J. Gotoh, A. Takeda, and K. Tono, "DC formulations and algorithms for sparse optimization problems," *Math. Programm.*, vol. 169, no. 1, pp. 141–176, 2018.

[17] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.

[18] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.

[19] X. Lv, G. Bi, and C. Wan, "The group lasso for stable recovery of block-sparse signal representations," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1371–1382, Apr. 2011.

[20] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4094–4107, Aug. 2012.

[21] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.

[22] Y. Nardi and A. Rinaldo, "On the asymptotic properties of the group lasso estimator for linear models," *Electron. J. Statist.*, vol. 2, pp. 605–633, 2008.

[23] F. Wei and J. Huang, "Consistent group selection in high-dimensional linear regression," *Bernoulli*, vol. 16, no. 4, pp. 1369–1384, 2010.

[24] J. Huang, P. Breheny, and S. Ma, "A selective review of group selection in high-dimensional models," *Statist. Sci.*, vol. 27, no. 4, pp. 481–499, Nov. 2012.

[25] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *J. Mach. Learn. Res.*, vol. 9, pp. 1179–1225, Jun. 2008.

[26] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Roy. Statist. Soc. Ser. B*, vol. 70, no. 1, pp. 53–71, 2008.

[27] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1830–1839, May 2008.

[28] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 101–111, Jan. 2003.

[29] L. Yu, H. Sun, J. Barbot, and G. Zheng, "Bayesian compressive sensing for cluster structured sparse signals," *Signal Process.*, vol. 92, no. 1, pp. 259–269, 2012.

[30] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa, "Compressive sensing for background subtraction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 155–168.

[31] Z. Gao, L. Cheong, and Y. Wang, "Block-sparse RPCA for salient motion detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1975–1987, Oct. 2014.

[32] J. Fang, L. Zhang, and H. Li, "Two-dimensional pattern-coupled sparse Bayesian learning via generalized approximate message passing," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2920–2930, Jun. 2016.

[33] F. Gustafsson, *Adaptive filtering and change detection*. Wiley, 2000.

[34] H. Kuroda, M. Yamagishi, and I. Yamada, "Exploiting sparsity in tight-dimensional spaces for piecewise continuous signal recovery," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6363–6376, Dec. 2018.

[35] ——, "Segmentation of piecewise ARX processes by exploiting sparsity in tight-dimensional spaces," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Jul. 2019, 5 pages.

[36] L. Wang, L. Zhao, G. Bi, C. Wan, and L. Yang, "Enhanced ISAR imaging by exploiting the continuity of the target scene," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5736–5750, Sep. 2014.

[37] E. Yoshikawa, T. Ushio, Z. Kawasaki, S. Yoshida, T. Morimoto, F. Mizutani, and M. Wada, "MMSE beam forming on fast-scanning phased array weather radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 3077–3088, May 2013.

[38] D. Kitahara, M. Nakahara, A. Hirabayashi, E. Yoshikawa, H. Kikuchi, and T. Ushio, "Nonlinear beamforming via convex optimization for phased array weather radar," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 1831–1835.

[39] D. Kitahara, H. Kuroda, A. Hirabayashi, E. Yoshikawa, H. Kikuchi, and T. Ushio, "Nonlinear beamforming based on group-sparsities of periodograms for phased array weather radar," *IEEE Trans. Geosci. Remote Sens.*, 2022, doi:10.1109/TGRS.2022.3154118. (early access)

[40] F. Mizutani, T. Ushio, E. Yoshikawa, S. Shimamura, H. Kikuchi, M. Wada, S. Satoh, and T. Iguchi, "Fast-scanning phased-array weather radar with angular imaging technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2664–2673, May 2018.

[41] B. Isom, R. Palmer, R. Kelley, J. Meier, D. Bodine, M. Yeary, B.-L. Cheong, Y. Zhang, T.-Y. Yu, and M. I. Biggerstaff, "The atmospheric imaging radar: Simultaneous volumetric observations using a phased array weather radar," *J. Atmos. Ocean. Technol.*, vol. 30, no. 4, pp. 655–675, Apr. 2013.

[42] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, 2011.

[43] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing, "Smoothing proximal gradient method for general structured sparse regression," *Ann. Appl. Stat.*, vol. 6, no. 2, pp. 719–752, 2012.

[44] İ. Bayram and Ö. D. Akyıldız, "Primal-dual algorithms for audio decomposition using mixed norms," *Signal, Image Video Process.*, vol. 8, no. 1, pp. 95–110, 2014.

[45] P.-Y. Chen and I. W. Selesnick, "Translation-invariant shrinkage/thresholding of group sparse signals," *Signal Process.*, vol. 94, pp. 476–489, 2014.

[46] G. Peyré and J. Fadili, "Group sparsity with overlapping partition functions," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2011, pp. 303–307.

[47] L. Yuan, J. Liu, and J. Ye, "Efficient methods for overlapping group lasso," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2011, pp. 352–360.

[48] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2009, pp. 433–440.

[49] G. Obozinski, L. Jacob, and J.-P. Vert, "Group lasso with overlaps: The latent group lasso approach," preprint arXiv:1110.0413, 2011, 60 pages.

[50] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *J. Mach. Learn. Res.*, vol. 12, no. 11, 2011.

[51] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.

[52] J. Fang, Y. Shen, H. Li, and P. Wang, "Pattern-coupled sparse Bayesian learning for recovery of block-sparse signals," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 360–372, Jan. 2015.

[53] L. Wang, L. Zhao, S. Rahardja, and G. Bi, “Alternative to extended block sparse Bayesian learning and its relation to pattern-coupled sparse Bayesian learning,” *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2759–2771, May 2018.

[54] A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging,” *J. Math. Imaging Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[55] E. Esser, X. Zhang, and T. F. Chan, “A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science,” *SIAM J. Imaging Sci.*, vol. 3, no. 4, pp. 1015–1046, 2010.

[56] I. Loris and C. Verhoeven, “On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty,” *Inverse Problems*, vol. 27, no. 12, Nov. 2011, 15 pages.

[57] P. Chen, J. Huang, and X. Zhang, “A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration,” *Inverse Problems*, vol. 29, no. 2, Jan. 2013, 33 pages.

[58] M. Yamagishi and I. Yamada, “Nonexpansiveness of a linearized augmented Lagrangian operator for hierarchical convex optimization,” *Inverse Problems*, vol. 33, no. 4, Mar. 2017, 35 pages.

[59] J. Yang and X. Yuan, “Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization,” *Math. Comput.*, vol. 82, no. 281, pp. 301–329, 2013.

[60] D. Gabay and B. Mercier, “A dual algorithm for the solution of nonlinear variational problems via finite element approximation,” *Comput. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.

[61] J. Eckstein and D. P. Bertsekas, “On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators,” *Math. Programm.*, vol. 55, no. 1, pp. 293–318, Apr. 1992.

[62] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi, “Proximal splitting algorithms for convex optimization: A tour of recent advances, with new twists,” preprint arXiv:1912.00137v7, 2021, 65 pages.

[63] P. L. Combettes, “Perspective functions: Properties, constructions, and examples,” *Set-Valued and Variational Analysis*, vol. 26, no. 2, pp. 247–264, 2018.

[64] P. L. Combettes and C. L. Müller, “Perspective functions: Proximal calculus and applications in high-dimensional statistics,” *J. Math. Anal. Appl.*, vol. 457, no. 2, pp. 1283–1306, 2018.

[65] ——, “Perspective maximum likelihood-type estimation via proximal decomposition,” *Electron. J. Statist.*, vol. 14, no. 1, pp. 207–238, 2020.

[66] H. Kuroda, D. Kitahara, and A. Hirabayashi, “A convex penalty for block-sparse signals with unknown structures,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2021, pp. 5430–5434.

[67] C. A. Micchelli, J. M. Morales, and M. Pontil, “Regularizers for structured sparsity,” *Adv. Comput. Math.*, vol. 38, no. 3, pp. 455–489, 2013.

[68] F. Lara, “A note on “Reguralizers for structured sparsity”,” *Adv. Comput. Math.*, vol. 44, no. 4, pp. 1321–1323, 2018.

[69] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, 2nd ed. Springer, 2011.

[70] L. Condat, “Fast projection onto the simplex and the ℓ_1 ball,” *Math. Programm.*, vol. 158, no. 1, pp. 575–585, 2016.

[71] J. Kominek and A. W. Black, “The CMU Arctic speech databases,” in *Proc. ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

Hiroki Kuroda (M’19) received the B.E. degree in computer science, and the M.E. and Ph.D. degrees in information and communications engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 2013, 2015 and 2019, respectively. From April 2017 to March 2019, he was a recipient of the Research Fellowship of the Japan Society for the Promotion of Science (JSPS). From April 2019 to March 2020, he was a postdoctoral researcher with the National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. Since April 2020, he has been an Assistant Professor with the College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan. His current research interests are in signal processing and its applications, sparse modeling, and optimization theory.

He received the Young Researcher Award from the IEICE Technical Group on Signal Processing in 2018 and the Seiichi Tejima Doctoral Dissertation Award from the Tokyo Institute of Technology in 2020.

Daichi Kitahara (M’18) received the B.E. degree in computer science in 2012, and the M.E. and Ph.D. degrees in communications and computer engineering in 2014 and 2017 from the Tokyo Institute of Technology, Tokyo, Japan, respectively. From 2014 to 2017, he was a recipient of the Research Fellowship of the Japan Society for the Promotion of Science (JSPS). Since 2017, he has been an Assistant Professor with the College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan. His current research interests are in signal processing and its applications, inverse problem, optimization theory, and multivariate spline theory.

He received the Young Researcher Awards from the IEICE Technical Group on Signal Processing, the SICE Technical Committee on Remote Sensing, and the IEEE Computational Intelligence Society Japan Chapter in 2013, 2015, and 2016, respectively. He also received the IEEE Signal Processing Society (SPS) Japan Student Journal Paper Award from the IEEE SPS Tokyo Joint Chapter in 2016 and the Yasujiro Niwa Outstanding Paper Award from the Tokyo Denki University in 2018.