第36回信号処理シンポジウム
2021年11月9日〜12日（リモート開催）

Exploiting Graph-Structured Sparsity via Convex Optimization

Hiroki Kuroda† Daichi Kitahara†
†College of Information Science and Engineering, Ritsumeikan University
Email: {kuroda, d-kita}@media.ritsumei.ac.jp

Abstract In this paper, we present a novel convex recovery model for graph-structured sparse signals. While various structured sparsities can be treated as the graph-structured sparsity, recovery of graph-structured sparse signals is known to be a NP-hard problem and thus existing frameworks rely on approximation algorithms, for which global optimality cannot be ensured in general. To resolve this difficulty, we propose a convex penalty function derived as a tight convex relaxation of a penalty function designed as the minimum of the mixed ℓ_2/ℓ_1 norm over the groups defined by connected components of the candidate subgraphs. We can obtain a global optimal solution of the proposed graph-structured recovery model with use of proximal splitting techniques. Numerical experiments show the effectiveness of the proposed recovery model.

1 Introduction

Structural information of sparsity patterns is the key for the further enhancement of sparsity-aware signal processing and machine learning. Various structured sparsities such as block-sparsity and tree-structured sparsity can be encoded as the graph-structured sparsity [1–3] with appropriately designed underlying graphs. Besides its generality, graph-structured sparsity has its own numerous applications, which include clustering of biological network [4] and anomaly detection [5–8] of, e.g., water distribution pipe breakage [9], traffic congestion [10], and disease outbreak [11].

While the graph-structured sparsity can model various structural information of nonzero patterns of the target signals, it is challenging to effectively exploit the graph-structured sparsity for signal processing and machine learning. Since graph-structured sparse signals are originally modeled by a nonconvex set, and it is known that the projection onto this set is a NP-hard problem [1]. Due to this difficulty, the existing methods, e.g., [1–3], rely on approximation algorithms, where the computation of a globally optimal solution is challenging. On the other hand, existing convex regularization frameworks [12–20] requires predefined subgraphs in prior to the estimation, and thus cannot flexibly incorporate the graph-structured sparsity.

In this paper, for recovery of graph-structured sparse signals, we present a novel convex framework where the relevant subgraph is automatically identified. We first design a nonconvex penalty function as the minimum of the mixed ℓ_2/ℓ_1 norm over the set of groups induced by subgraphs of the underlying graph. Then, the proposed convex penalty function is derived as its tight convex relaxation by utilizing the variational representation of the ℓ_2 norm [21, Lemma 1] and the introduction of latent variables characterized by a difference operator on the underlying graph. We design a graph-structured recovery model with the proposed penalty, whose globally optimal solution can be obtained by applying proximal splitting techniques. Numerical experiments for recovery of graph-structured sparse signals from linear measurements demonstrate the effectiveness of the proposed method.

2 Preliminaries

2.1 Mathematical Notions

$\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ respectively denote the sets of all real numbers, all nonnegative real numbers, and all positive real numbers. For matrices or vectors, we denote the transpose by $(\cdot)^\top$. For $\boldsymbol{x} = (x_1, \dots, x_N)^\top \in \mathbb{R}^N$ and an index set $\mathcal{I} \subset \{1, \dots, N\}$, $\boldsymbol{x}_\mathcal{I} := (x_n)_{n\in\mathcal{I}}$ denotes the subvector of $\boldsymbol{x}$ indexed by $\mathcal{I}$. We define the support of $\boldsymbol{x} \in \mathbb{R}^N$ by $\mathrm{supp}(\boldsymbol{x}) := \{n \in \{1, \dots, N\} \,|\, x_n \neq 0\}$. For a set $\mathcal{A}$, $|\mathcal{A}|$ denotes the cardinality of $\mathcal{A}$. The ℓ_2 norm and the ℓ_1 norm of $\boldsymbol{x} \in \mathbb{R}^N$ are respectively defined by $\|\boldsymbol{x}\|_2 := \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ and $\|\boldsymbol{x}\|_1 := \sum_{n=1}^N |x_n|$. The weighted ℓ_0 pseudo-norm with the weight vector $\boldsymbol{w} \in \mathbb{R}^N_{++}$ is defined by $\|\boldsymbol{x}\|_{0,\boldsymbol{w}} := \sum_{n=1}^N w_n \mathbb{1}(x_n)$, where $\mathbb{1}(x) = 0$ if $x = 0$ and $\mathbb{1}(x) = 1$ otherwise.

We denote an undirected graph by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of vertices and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is a set of edges. Any nonempty graph $\mathcal{G}$ can be uniquely decomposed into *connected components*, that is, *maximal connected subgraphs* [22] which form a *partition* of $\mathcal{V}$. We denote the connected components of $\mathcal{G}$ by

$$c_k(\mathcal{G}) \subset \mathcal{V} \quad \text{for } k = 1, \dots, \kappa(\mathcal{G}),$$

where $\kappa(\mathcal{G})$ is the number of connected components.

2.2 Problem Formulation

We consider the estimation of $\boldsymbol{x}^\star \in \mathbb{R}^N$ supposed to be graph-structured sparse with an underlying graph $\mathcal{G}_0 = (\mathcal{V}_0, \mathcal{E}_0)$ where $\mathcal{V}_0 := \{1, \dots, N\}$. Namely, nonzero entries of $\boldsymbol{x}^\star$ are clustered in several connected subgraphs of $\mathcal{G}_0$. In other words, $\boldsymbol{x}^\star$ is group-sparse by setting the groups to the connected components of $\mathcal{G}^\star = (\mathcal{V}_0, \mathcal{E}^\star)$,

where $\mathcal{E}^\star \subset \mathcal{E}_0$ is defined by eliminating irrelevant edges from $\mathcal{E}_0$. More precisely, we suppose that

$$\boldsymbol{x}^\star_{c_k(\mathcal{G}^\star)} \approx \boldsymbol{0} \text{ for many } k \in \{1, \dots, \kappa(\mathcal{G}^\star)\}.$$

To evaluate the cost of edges eliminated from $\mathcal{E}_0$, we also consider weights of edges, which are represented by a function $w_0 \colon \mathcal{E} \to \mathbb{R}_{++}$. Note that, if such weights are not available, we can simply set $w_0(e) = 1$ for every $e \in \mathcal{E}$.

3 Proposed Graph-Structured Sparse Recovery

3.1 Design of Penalty Function

We focus on the observation that the graph-structured sparsity can be evaluated as the group-sparsity by setting the groups to the connected components of an appropriate subgraph of the underlying graph $\mathcal{G}_0$. The group-sparsity can be effectively promoted by the mixed ℓ_2/ℓ_1 norm, i.e., the sum of the ℓ_2 norm of subvectors of the groups. However, it is not suitable to directly apply the mixed ℓ_2/ℓ_1 norm to graph-structured sparse signals because the appropriate subgraph is unknown a priori. Thus, we newly introduce a penalty function as the minimum of the mixed ℓ_2/ℓ_1 norm over candidate subgraphs of $\mathcal{G}_0$. Although the introduced penalty function is nonconvex, we derive its tight convex relaxation by using the variational representation of the ℓ_2 norm [21, Lemma 1] and introducing latent variables.

Concretely, we begin by defining the mixed ℓ_2/ℓ_1 norm induced by a graph $\mathcal{G}$:

$$\|\boldsymbol{x}\|^{\mathcal{G}}_{2,1} := \sum_{k=1}^{\kappa(\mathcal{G})} \sqrt{|c_k(\mathcal{G})|}\, \|\boldsymbol{x}_{c_k(\mathcal{G})}\|_2, \tag{1}$$

where $\boldsymbol{x}_{c_k(\mathcal{G})}$ is the subvector indexed by the connected component $c_k(\mathcal{G})$, and $\sqrt{|c_k(\mathcal{G})|}$ is the weight introduced based on the suggestions in, e.g., [14–17], for group-sparse recovery. To automatically identify the appropriate subgraph, we design the penalty function $\psi^{\mathcal{G}_0}_d(\boldsymbol{x})$ by

$$\psi^{\mathcal{G}_0}_d(\boldsymbol{x}) := \min_{\mathcal{G} \in \mathcal{S}_d} \|\boldsymbol{x}\|^{\mathcal{G}}_{2,1}, \tag{2}$$

where $\mathcal{S}_d$ consists of all candidate subgraphs of $\mathcal{G}_0$ defined by eliminating edges whose total weights are less than or equal to $d \in \mathbb{R}_+$, i.e.,

$$\mathcal{S}_d := \left\{ \mathcal{G} = (\mathcal{V}_0, \mathcal{E}) \;\middle|\; \mathcal{E} \subset \mathcal{E}_0 \text{ and } \sum_{e \in \mathcal{E}_0 \setminus \mathcal{E}} w_0(e) \le d \right\}. \tag{3}$$

To derive the convex relaxation of $\psi^{\mathcal{G}_0}_d(\boldsymbol{x})$, we exploit the following lemma, from [21], which shows a variational representation of the ℓ_2 norm.

Lemma 1. *Define a coercive lower semicontinuous convex function $\phi \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+ \cup \{\infty\}$ by*

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

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

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

for any index set $\mathcal{I} \subset \{1, \dots, N\}$.

By applying Lemma 1 for each $\sqrt{|c_k(\mathcal{G})|}\, \|\boldsymbol{x}_{c_k(\mathcal{G})}\|_2$ in the definition of $\|\boldsymbol{x}\|^{\mathcal{G}}_{2,1}$ in (1), we can rewrite $\psi^{\mathcal{G}_0}_d(\boldsymbol{x})$ in (2) as

$$\psi^{\mathcal{G}_0}_d(\boldsymbol{x}) = \min_{\mathcal{G} \in \mathcal{S}_d} \left[\min_{\boldsymbol{\tau} \in \mathbb{R}^{\kappa(\mathcal{G})}} \sum_{k=1}^{\kappa(\mathcal{G})} \sum_{n \in c_k(\mathcal{G})} \phi(x_n, \tau_k) \right].$$

Let us introduce a latent vector $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_N)^\top \in \mathbb{R}^N$ by

$$\sigma_n = \tau_k \;\; (n \in c_k(\mathcal{G})) \quad \text{for } k = 1, \dots, \kappa(\mathcal{G}).$$

Notice that $c_k(\mathcal{G})$ $(k = 1, \dots, \kappa(\mathcal{G}))$ are the connected components of $\mathcal{G}$, and the sum of the weights of edges eliminated from $\mathcal{G}_0$ is less than or equal to d as shown in (3). Thus, $\boldsymbol{\sigma}$ is characterized by the condition that

$$\|\boldsymbol{D}\boldsymbol{\sigma}\|_{0,\boldsymbol{w}} \le d,$$

where $\boldsymbol{D} \in \mathbb{R}^{|\mathcal{E}_0| \times N}$ is the difference operator on the graph $\mathcal{G}_0$, i.e.,

$$\boldsymbol{D}\boldsymbol{\sigma} = (\sigma_n - \sigma_{n'})_{(n,n') \in \mathcal{E}_0},$$

and $\boldsymbol{w} = (w(e))_{e \in \mathcal{E}_0} \in \mathbb{R}^{|\mathcal{E}_0|}$. Moreover, since the connected components form a partition of $\mathcal{V}_0$, i.e., $\bigcup_{k=1}^{\kappa(\mathcal{G})} c_k(\mathcal{G}) = \mathcal{V}_0 = \{1, \dots, N\}$ and $c_k(\mathcal{G}) \cap c_{k'}(\mathcal{G}) = \varnothing$ $(k \ne k')$, we have

$$\sum_{k=1}^{\kappa(\mathcal{G})} \sum_{n \in c_k(\mathcal{G})} \phi(x_n, \tau_k) = \sum_{n=1}^{N} \phi(x_n, \sigma_n).$$

Thus, we finally have

$$\psi^{\mathcal{G}_0}_d(\boldsymbol{x}) = \min_{\substack{\boldsymbol{\sigma} \in \mathbb{R}^N \\ \|\boldsymbol{D}\boldsymbol{\sigma}\|_{0,\boldsymbol{w}} \le d}} \sum_{n=1}^{N} \phi(x_n, \sigma_n).$$

By replacing the weighted ℓ_0 pseudo-norm in the constraint with its convex envelope, i.e., the weighted ℓ_1 norm, we derive the proposed convex penalty as

$$\Psi^{\mathcal{G}_0}_\alpha(\boldsymbol{x}) := \min_{\substack{\boldsymbol{\sigma} \in \mathbb{R}^N \\ \|\boldsymbol{W}\boldsymbol{D}\boldsymbol{\sigma}\|_1 \le \alpha}} \sum_{n=1}^{N} \phi(x_n, \sigma_n), \tag{6}$$

where $\alpha \in \mathbb{R}_+$ is a tuning parameter related to the the sum of the weights of edges eliminated from $\mathcal{G}_0$, and $\boldsymbol{W} = \operatorname{diag}(\boldsymbol{w})$.

3.2 Optimization Algorithm

We present a graph-structured sparse recovery model using the proposed penalty (6). Specifically, we consider the following regularization model:

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}}\, f(\boldsymbol{L}\boldsymbol{x}) + \lambda\Psi_\alpha^{\mathcal{G}_0}(\boldsymbol{x}), \tag{7}$$

where $f(\boldsymbol{L}\boldsymbol{x})$ is some convex data-fidelity function with $f\colon \mathbb{R}^J \to \mathbb{R}_+$ and $\boldsymbol{L} \in \mathbb{R}^{J\times N}$, and $\lambda \in \mathbb{R}_{++}$ is the regularization parameter. We suppose that $f \in \Gamma_0(\mathbb{R}^J)$, and its proximity operator can be computed efficiently. Such examples include the square error $f(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{u}\|_2^2$ and the absolute error $f(\boldsymbol{u}) = \|\boldsymbol{y}-\boldsymbol{u}\|_1$, where $\boldsymbol{y}$ is the known observation vector and $\boldsymbol{L}$ is set to the known measurement matrix. We can also apply (7) to classification problems by using, e.g., the logistic loss $f(\boldsymbol{u}) = \sum_{j=1}^J \log(1+\exp(y_j u_j))$ and the hinge loss $f(\boldsymbol{u}) = \sum_{j=1}^J \max\{1-y_j u_j, 0\}$, where $y_j \in \{-1,+1\}$ is a label for the j-th training data $\boldsymbol{a}_j \in \mathbb{R}^N$, and $\boldsymbol{L} = (\boldsymbol{a}_1,\dots,\boldsymbol{a}_J)^\top$.

Plugging the definition of $\Psi_\alpha(\boldsymbol{x})$ in (6) into (7), we can solve the optimization problem (7) as

$$\left.\begin{aligned}\underset{(\boldsymbol{x},\boldsymbol{\sigma})\in\mathbb{R}^N\times\mathbb{R}^N}{\text{minimize}}\; & f(\boldsymbol{L}\boldsymbol{x}) + \lambda\sum_{n=1}^N \phi(x_n,\sigma_n)\\ \text{subject to}\; & \|\boldsymbol{W}\boldsymbol{D}\boldsymbol{\sigma}\|_1 \le \alpha\end{aligned}\right\}, \tag{8}$$

We obtain Algorithm 1 by applying the primal-dual algorithm [23–28] to the reformulated problem (8) with further slight reformulations. Based on [29, Example 2.4], we can compute the proximity operator of $\gamma\lambda\varphi$ in Algorithm 1 by

$$\operatorname{prox}_{\gamma\lambda\varphi}(\boldsymbol{x},\boldsymbol{\sigma}) = \left(\operatorname{prox}_{\gamma\lambda\phi}(x_n,\sigma_n)\right)_{n=1}^N, \tag{9}$$

with

$$\operatorname{prox}_{\gamma\lambda\phi}(x,\sigma) = \begin{cases}(0,0), & \text{if } 2\gamma\lambda\sigma + |x|^2 \le \gamma^2\lambda^2;\\ (0,\sigma-\frac{\gamma\lambda}{2}), & \text{if } x=0 \text{ and } 2\sigma > \gamma\lambda;\\ \left(x-\gamma\lambda s\frac{x}{|x|}, \sigma+\gamma\lambda\frac{s^2-1}{2}\right), & \text{otherwise,}\end{cases} \tag{10}$$

where $s \in \mathbb{R}_{++}$ is the unique positive root of

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

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. The proximity operator of γf depends on the employed data-fidelity function. The proximity operator of $\gamma\iota_{B_1^\alpha}$ reduces to the ℓ_1 ball projection $P_{B_1^\alpha}$, which can be computed as

$$P_{B_1^\alpha}(\boldsymbol{\eta}) = \begin{cases}\boldsymbol{\eta}, & \text{if } \|\boldsymbol{\eta}\|_1 \le \alpha;\\ (a_n\operatorname{sign}(\eta_n))_{n=1}^{N-1}, & \text{otherwise,}\end{cases} \tag{11}$$

with

$$a_n := \max\left\{|\eta_n| - \sum_{t=1}^T \frac{\rho_t-\alpha}{T}, 0\right\}, \tag{12}$$

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

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

Algorithm 1: Solver for the proposed model (7)

Input: $\gamma>0$, $\mu_1 \in \left(0, \frac{1}{\sqrt{\|\boldsymbol{L}\|_{\mathrm{op}}^2+1}}\right]$, $\mu_2 \in \left(0, \frac{1}{\sqrt{\|\boldsymbol{W}\boldsymbol{D}\|_{\mathrm{op}}^2+1}}\right]$

for $i = 0,1,2,\dots$ **do**

$\tilde{\boldsymbol{x}}^{(i+1)} = \boldsymbol{x}^{(i)} + \mu_1\boldsymbol{L}^\top(\boldsymbol{r}_1^{(i)} - \mu_1(\boldsymbol{L}\boldsymbol{x}^{(i)} - \boldsymbol{u}^{(i)}))$

$\tilde{\boldsymbol{\sigma}}^{(i+1)} = \boldsymbol{\sigma}^{(i)} + \mu_2\boldsymbol{D}^\top\boldsymbol{W}(\boldsymbol{r}_2^{(i)} - \mu_2(\boldsymbol{W}\boldsymbol{D}\boldsymbol{\sigma}^{(i)} - \boldsymbol{\eta}^{(i)}))$

$\tilde{\boldsymbol{u}}^{(i+1)} = \boldsymbol{u}^{(i)} - \mu_1(\boldsymbol{r}_1^{(i)} - \mu_1(\boldsymbol{L}\boldsymbol{x}^{(i)} - \boldsymbol{u}^{(i)}))$

$\tilde{\boldsymbol{\eta}}^{(i+1)} = \boldsymbol{\eta}^{(i)} - \mu_2(\boldsymbol{r}_2^{(i)} - \mu_2(\boldsymbol{W}\boldsymbol{D}\boldsymbol{\sigma}^{(i)} - \boldsymbol{\eta}^{(i)}))$

$(\boldsymbol{x}^{(i+1)}, \boldsymbol{\sigma}^{(i+1)}) = \operatorname{prox}_{\gamma\lambda\varphi}(\tilde{\boldsymbol{x}}^{(i+1)}, \tilde{\boldsymbol{\sigma}}^{(i+1)})$ // see (9) and (10)

$\boldsymbol{u}^{(i+1)} = \operatorname{prox}_{\gamma f}(\tilde{\boldsymbol{u}}^{(i+1)})$

$\boldsymbol{\eta}^{(i+1)} = P_{B_1^\alpha}(\tilde{\boldsymbol{\eta}}^{(i+1)})$ // see (11) and (12)

$\boldsymbol{r}_1^{(i+1)} = \boldsymbol{r}_1^{(i)} - \mu_1(\boldsymbol{L}\boldsymbol{x}^{(i+1)} - \boldsymbol{u}^{(i+1)})$

$\boldsymbol{r}_2^{(i+1)} = \boldsymbol{r}_2^{(i)} - \mu_2(\boldsymbol{W}\boldsymbol{D}\boldsymbol{\sigma}^{(i+1)} - \boldsymbol{\eta}^{(i+1)})$

4 Numerical Experiments

We consider the estimation of a graph-structured sparse signal $\boldsymbol{x}^\star \in \mathbb{R}^N$ from noisy compressive measurements. The underlying graph $\mathcal{G}_0$ is randomly generated under the condition that each node has 3 edges. Then, 2 subgraphs of $\mathcal{G}_0$ are randomly chosen for 150 nonzero entries of $\boldsymbol{x}^\star$, where N is set to 500. Amplitudes of the nonzero entries are drawn from i.i.d. Gaussian distribution $\mathcal{N}(0,1)$. The selection scheme of subgraphs are as follows. First, the numbers of nodes of subgraphs are randomly chosen from $\{1,2,\dots,149\}$ under the constraint that their sum equals to 150. Each subgraph is initialized with a node from $\mathcal{G}_0$. Then, for a node randomly chosen from the subgraph, adjacent nodes are added into the subgraph until the number of nodes becomes the prescribed number. If some nodes are shared by the subgraphs chosen by this process, we reject these subgraphs, and repeat the process until the prescribed condition $\operatorname{supp}(\boldsymbol{x}^\star) = 150$ is met. The measurements are generated 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. $\mathcal{N}(0,1)$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ is the white Gaussian noise.

Table 1: Normalized mean square error between $\boldsymbol{x}^\star$ and $\hat{\boldsymbol{x}}$ averaged over 100 trials.

Number of measurements	Proposed	ℓ_1
250	$\mathbf{-8.27}$	-6.89
300	$\mathbf{-17.24}$	-14.75
350	$\mathbf{-30.11}$	-29.16

We compare the proposed regularization model (7) against the ℓ_1 norm regularization model. In the proposed regularization model (7), 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 ℓ_1 norm is also combined with the square error in similar ways. Note that the regularization parameter is tuned independently for each model to obtain the best results. We terminate the iteration when the norm of the differences between the variables of successive iterates is below the threshold 10^{-4}. In Table 1, we show the normalized mean square error (NMSE) $\|\boldsymbol{x}^\star - \hat{\boldsymbol{x}}\|_2^2/\|\boldsymbol{x}^\star\|_2^2$ against the number of measurements, where $\hat{\boldsymbol{x}}$ is the solution of each model and the results are averaged over 100 independent trials. The results show that the proposed penalty function yields better estimation accuracy than the ℓ_1 regularization model.

References

[1] C. Hegde, P. Indyk, and L. Schmidt, "A nearly-linear time framework for graph-structured sparsity," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2015, pp. 928–937.

[2] F. Chen and B. Zhou, "A generalized matching pursuit approach for graph-structured sparsity," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 1389–1395.

[3] B. Zhou, F. Chen, and Y. Ying, "Stochastic iterative hard thresholding for graph-structured sparsity optimization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, vol. 97, pp. 7563–7573.

[4] J. Peng, L. Zhu, Y. Wang, and J. Chen, "Mining relationships among multiple entities in biological networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 3, pp. 769–776, 2020.

[5] J. Cadena, F. Chen, and A. Vullikanti, "Graph anomaly detection based on steiner connectivity and density," *Proc. IEEE*, vol. 106, no. 5, pp. 829–845, 2018.

[6] B. Zhou and F. Chen, "Graph-structured sparse optimization for connected subgraph detection," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2016, pp. 709–718.

[7] F. Chen, B. Zhou, A. Alim, and L. Zhao, "A generic framework for interesting subspace cluster detection in multi-attributed networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2017, pp. 41–50.

[8] F. Jie, C. Wang, F. Chen, L. Li, and X. Wu, "Block-structured optimization for anomalous pattern detection in interdependent networks," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2019, pp. 1138–1143.

[9] D. De Oliveira, D. Neill, J. Garrett, and L. Soibelman, "Detection of patterns in water distribution pipe breakage using spatial scan statistics for point events in a physical network," *J. Comput. Civ. Eng.*, vol. 25, no. 1, pp. 21–30, Jan. 2011.

[10] B. Anbaroğlu, T. Cheng, and B. Heydecker, "Non-recurrent traffic congestion detection on heterogeneous urban road networks," *Transportmetrica A: Transport Science*, vol. 11, no. 9, pp. 754–771, 2015.

[11] S. Speakman, E. McFowland, and D. B. Neill, "Scalable detection of anomalous patterns with connectivity constraints," *Journal of Computational and Graphical Statistics*, vol. 24, no. 4, pp. 1014–1033, 2015.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

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

[21] H. Kuroda and D. Kitahara, "Block-sparse recovery with optimal block partition," preprint TechRxiv:14538552, May 2021.

[22] R. Diestel, *Graph Theory*, Springer, 5th edition, 2017.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] P. L. Combettes and C. L. Müller, "Perspective maximum likelihood-type estimation via proximal decomposition," *Electron. J. Statist.*, vol. 14, no. 1, pp. 207–238, 2020.