A Generalized Formulation for Group Selection via ADMM

Chengyu Ke*, Sunyoung Shin†, Yifei Lou‡, Miju Ahn*

Abstract

This paper studies a statistical learning model where the model coefficients have a pre-determined non-overlapping group sparsity structure. We consider a combination of a loss function and a regularizer to recover the desired group sparsity patterns, which can embrace many existing works. We analyze the stationary solution of the proposed formulation, obtaining a sufficient condition for the stationary solution to achieve optimality and establishing a bound of the distance from the solution to a reference point that is related to the ground-truth from a probabilistic interpretation. We develop an efficient algorithm that adopts an alternating direction method of multiplier (ADMM), showing that the iterates converge to a stationary solution under certain conditions. In the numerical experiment, we implement the algorithm for generalized linear models with convex and nonconvex group regularizers to evaluate the model performance on various data types, noise levels, and sparsity settings.

Keywords: ADMM, Group sparsity, Group variable selection, Nonconvex programming

1 Introduction

Statistical learning models have been extensively used for data interpretations and predictions in many fields collecting a large amount of data, such as computer vision and computational biology. Given a dataset, a statistical learning model is utilized to estimate unknown model coefficients by minimizing a certain loss function such as mean squared error and cross-entropy [17]. With a wide range of features, only a small subset of them actually corresponds to nonzero model coefficients; as a result, fitting a model by blindly incorporating all the available features may cause a misleading interpretation of contributing features and an inaccurate prediction. In an attempt to select meaningful features automatically, feature selection techniques [4, 16, 20] have been developed to produce a sparsely fitted model with many zero coefficients. For example, regularization methods, appending a penalty function to the loss function, discard insignificant features by setting small model coefficients to zero. A well-known regularization term is the ℓ_1-norm of the model coefficients, known as least absolute shrinkage and selection operator (LASSO) [51]. LASSO has been widely used in diverse settings to tackle practical problems such as forecasting corporate bankruptcy and assessing drug efficacy [50, 52]. However, the convex ℓ_1 penalty may mistakenly suppress large coefficients while shrinking small coefficients to zero, which is referred to as biasedness [14]. To alleviate this issue, nonconvex regularization functions have been proposed, including smoothly clipped absolute deviation (SCAD) [14], minimax concave penalty (MCP) [64], transformed ℓ_1 (TL1) [36, 66], $\ell_{1/2}$ penalty [25, 61], scale-invariant ℓ_1 [44, 54], and logarithm penalty [9]. These nonconvex regularizers try to retain the large model coefficients while applying a similar shrinkage of smaller coefficients as the convex ℓ_1-norm.

Group selection is a variant of feature selection when features have a group structure; specifically, each feature belongs to one or more groups. A group structure is non-overlapping if each feature is assigned to exactly one group. In the non-overlapping group selection, the features in the same group are required to have the corresponding coefficients altogether nonzero or zero, which is called group sparsity. The group structure of the features is reasonable in many applications. For example, a categorical variable should be converted

*Department of Operations Research and Engineering Management, Southern Methodist University, Dallas, TX 75205, USA. (cke@smu.edu, mijua@smu.edu)

†Department of Mathematics, Pohang University of Science and Technology, Pohang, Gyeongbuk, South Korea, 37673. (sunyoungshin@postech.ac.kr)

‡Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080, USA. (yifei.lou@utdallas.edu)

to a group of several binary dummy variables prior to model fitting. Group sparsity can be enforced into a model by extending regularization techniques to a group setting. Group LASSO [63, 34] achieves group sparsity with the convex $\ell_{2,1}$ norm, i.e., the sum of the ℓ_2 norms of the coefficients from the same group. Exploiting the convexity, many algorithms were applied and developed to solve for the group LASSO, including the gradient descent [68], the coordinate descent [24, 45], the second-order cone program [27], the semismooth Newton method [67], the subspace acceleration method [11], and the alternating direction method of multipliers (ADMM) [6, 12]. However, group LASSO may inherently have bias in the same way as LASSO [20]. Consequently, nonconvex regularizers such as $\ell_{0,2}$ penalty [21], group capped-ℓ_1 [37, 43], group MCP [20], group SCAD [55], and group LOG [22], were introduced. A variety of numerical algorithms are considered to solve for the nonconvex optimization, including primal dual active set algorithm [21], smoothing penalty algorithm [37], difference-of-convex algorithm [43], coordinate descent algorithms [58, 41], group descent algorithm [8], coordinate majorization descent algorithm [57], and iterative reweighted algorithm [22].

We propose a generalized formulation for fitting a statistical learning model with a non-overlapping group structure. For example, we establish the connection of loss functions under our setting to generalized linear models (GLMs) [33]. Many existing regularizations can be regarded as special cases of our generalized framework; we consider ℓ_1 penalty, SCAD, MCP, TL1, and $\ell_{1/2}$ as case studies. Our optimization problem is nonconvex if we use nonconvex regularization terms. Instead of a global optimum that might be hard to obtain, our analysis is based on the stationary solution, which can be provably achieved by using our algorithm. Stationary solutions were studied for nonconvex programming analyses, especially for the difference of convex (DC) problems [38, 53, 32]. In the DC literature, it has been shown that a certain kind of stationarity is a necessary condition for local optimality [38]. Furthermore, the stationary solution can achieve local and even global optimality [3] under suitable conditions. In this paper, we identify such conditions based on a restricted strong convexity (RSC) assumption [18]. RSC was originally analyzed for the convex LASSO problem where the global minimizer is compared to the ground-truth. We relax the requirement of having the ground-truth vector by a reference point that can be tied to the ground-truth in a probabilistic interpretation. We further provide a bound for the distance from the stationary solution obtained by our approach to this reference point.

To solve our proposed model, we design an iterative scheme based on the ADMM framework, which is an efficient method for solving large-scale optimization problems in statistics, machine learning, and related fields [10, 15, 47, 59, 69]. Although ADMM is originally designed for convex optimization, it can be extended to nonconvex problems, and its global convergence can be proved under certain conditions [19, 39, 56]. In our problem formulation, ADMM provides an iterative scheme that involves two subproblems that can be minimized sequentially. One subproblem involving the loss function is assumed to be convex, and hence it can be solved efficiently by closed-form solution or iterative methods. The other subproblem is related to the nonconvex regularization terms. We rely on proximal operators [40] to produce a stationary solution and characterize conditions for the global optimality of stationary solutions. Theoretically, we prove the subsequence convergence of the ADMM framework. In other words, the sequence generated by the ADMM has a subsequence convergence to a stationary point of our proposed model.

We conduct in-depth experiments on various datasets under different learning settings. We use synthetic datasets for linear regression and Poisson regression, and one real dataset for logistic regression. With hyperparameters tuned by cross-validation, our framework with various group regularization terms is applied to those datasets. An overall evaluation indicates that nonconvex penalty functions consistently outperform convex penalty functions.

We summarize our major contributions as follows,

(1) We introduce a generalized formulation for a non-overlapping group sparsity problem that can be reduced to many existing works;
(2) We investigate properties of the stationary solutions of the problem to achieve optimality as well as the bound on the distance from the solution to a reference point which is closely related to the optima of the loss function;
(3) We prove the subsequence convergence of ADMM iterates to a stationary point of our proposed model;
(4) Our numerical experiments on various models under different group structure settings evaluate the performance of existing group sparsity functions under the proposed formulation, showing the advantages of nonconvex penalties over convex ones.

The rest of the paper is organized as follows. In Section 2, we introduce a generalized formulation for non-overlapping group selection. Section 3 analyzes the conditions for global optimality of stationary solutions to our optimization problem and their properties under the RSC assumption on the loss function. In Section 4, we present the ADMM framework for minimizing the proposed model with convergence analysis. In Section 5, we present in-depth numerical experiments on both synthetic and real datasets for linear, Poisson, and logistic regressions. Section 6 concludes the paper.

2 The Proposed Framework

We consider a statistical learning problem defined by a coefficient vector $\mathbf{x} \in \mathbb{R}^d$ with a pre-determined group structure. We restrict our attention to a non-overlapping group selection by assuming that components of the variable $\mathbf{x}$ from m non-overlapping groups, denoted by $\mathcal{G}_k$ for $k \in \{1, \ldots, m\}$. Specifically $\mathcal{G}_k$ is a set of indices of $\mathbf{x}$ that belong to the kth group. The setting of non-overlapping groups implies that all $\mathcal{G}_k$'s are mutually exclusive. The notation $|\mathcal{G}_k|$ represents the cardinality of the set $\mathcal{G}_k$, and $\mathbf{x}_{\mathcal{G}_k} \in \mathbb{R}^{|\mathcal{G}_k|}$ is defined as a subvector of $\mathbf{x}$ that only consists of the coefficients in the group $\mathcal{G}_k$. For the rest of the paper, we denote $\mathcal{G}^{\max} \triangleq \max\limits_{1 \le k \le m} \sqrt{|\mathcal{G}_k|}$.

We aim to penalize the complexity of the group structure while minimizing the loss of the model simultaneously. To this end, we propose a general framework

$$\min_{\mathbf{x}} \; F_\lambda(\mathbf{x}) \triangleq L(\mathbf{x}) + \lambda \sum_{k=1}^{m} \sqrt{|\mathcal{G}_k|} \; \underbrace{p(\|\mathbf{x}_{\mathcal{G}_k}\|_2)}_{\triangleq P_k(\mathbf{x}_{\mathcal{G}_k})}, \tag{1}$$

where the loss function $L(\cdot)$ measures the fit of the model to the observed data, the regularizer $P_k(\cdot)$ determines the group complexity for the kth group of the model coefficients, and a hyperparameter $\lambda > 0$ balances between the model fitting and the group complexity. Note that $P_k : \mathbb{R}^{|\mathcal{G}_k|} \to \mathbb{R}$ is a composite function consisting of a univariate sparsity function $p : \mathbb{R} \to \mathbb{R}$ and the norm of the coefficients corresponding to $\mathcal{G}_k$. The following set of assumptions is considered on $p(\cdot)$:

(P1). p is symmetric about zero on $\mathbb{R}$, i.e., $p(t) = p(-t)$, concave and non-decreasing on $[0, \infty)$ with $p(0) = 0$.
(P2). The derivative of p is well-defined except at 0, finite with $p'(t) \ge 0$, and $u \triangleq \sup\limits_t p'(t)$ for any $t > 0$.
(P3). There exists a constant $\mathrm{Lip_p} > 0$ such that

$$|p'(t_1) - p'(t_2)| \le \mathrm{Lip_p}|\mathrm{t_1} - \mathrm{t_2}| \quad \forall\ \mathrm{t_1}, \mathrm{t_2} > 0.$$

It is straightforward to verify that many popular sparsity-promoting functions, including SCAD [14], MCP [64], transformed ℓ_1 [36, 66], and logarithm penalty [9], satisfy all the assumptions. We assume (P1) throughout this paper and impose (P2) or (P3) wherever needed.

The assumptions on the loss function are summarized as follows.

(A1). $L(\cdot)$ is lower-bounded.
(A2). There exists a constant $\mathrm{Lip_L} > 0$ such that

$$\|\nabla L(\mathbf{x}_1) - \nabla L(\mathbf{x}_2)\|_2 \le \mathrm{Lip_L}\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \quad \forall \mathbf{x}_1,\, \mathbf{x}_2.$$

(A3). $L(\cdot)$ is convex with modulus σ, i.e., there exists a constant $\sigma \ge 0$ such that

$$L(\mathbf{x}_1) - L(\mathbf{x}_2) - \langle \nabla L(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2 \rangle \ge \frac{\sigma}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \quad \forall \mathbf{x}_1,\, \mathbf{x}_2.$$

If σ is strictly positive, $L(\cdot)$ is strongly convex.

Some sparsity functions in the literature have been recast as DC functions [3, 62, 31, 66]. Using a unified DC representation [3] of p with a convex loss, the problem (1) can be written as a DC formulation, e.g., $F_\lambda(\mathbf{x}) = g(\mathbf{x}) - h(\mathbf{x})$ where both g and h are convex functions. Subsequently, existing algorithms

[23, 28, 42, 26] can be applied to compute a critical point, e.g., $\bar{\mathbf{x}}$ with $\mathbf{0} \in \partial F_\lambda(\bar{\mathbf{x}})$. Our approach bypasses the use of a DC representation of (1) and introduces a numerical method that computes a stationary point.

The loss function used in our optimization problem (1) can be specified by generalized linear models (GLMs) that are widely used for supervised learning, where input features explain a response [33]. GLMs are an extension of ordinary linear regression beyond Gaussian response variables, where the responses b_i, $i = 1, \ldots, n$, follow any exponential family distribution with a parameter θ_i. Specifically, the probability density function of b_i takes the following canonical form:

$$f(b_i; \theta_i) = \phi(b_i) \exp\Big\{ b_i \theta_i - \psi(\theta_i) \Big\}, \tag{2}$$

where ψ is a cumulant function and ϕ is a function that is independent of θ_i [33, 60]. Both ψ and ϕ are given functions by the exponential family distribution in consideration, and the first-order derivative of the cumulant function $\psi'(\theta_i)$ is the expected value of b_i [29, 33]. Taking a Gaussian distribution with mean θ_i and variance 1, for example, one sets $\psi(\theta_i) = \frac{\theta_i^2}{2}$ and $\phi(b_i) = \frac{\exp(-b_i^2/2)}{\sqrt{2\pi}}$.

Denote the ith observation of the input features by a row vector $A_i \in \mathbb{R}^d$ of the matrix $A \in \mathbb{R}^{n \times d}$. A GLM associates $\psi'(\theta_i)$ with a linear function of the features $A_i \mathbf{x}$ such that $\theta_i = A_i \mathbf{x}$. Omitting $\phi(b_i)$ that is free of θ_i, the loss function L for the GLM is written by summing the negative exponent in (2) for all the observations (A_i, b_i), $i = 1, \ldots, n$, i.e.,

$$\frac{1}{n} \sum_{i=1}^{n} \{\psi(\theta_i) - b_i \theta_i\} = \frac{1}{n} \sum_{i=1}^{n} \{\psi(A_i \mathbf{x}) - b_i A_i \mathbf{x}\}.$$

For example, the logistic regression models binary responses $b_i \in \{0, 1\}$ by

$$\psi(\theta_i) = \log(1 + \exp(\theta_i)) = \log\{1 + \exp(A_i \mathbf{x})\}.$$

The Poisson regression for count data $b_i \in \{0, 1, 2, \ldots\}$ has $\psi(\theta_i) = \exp(\theta_i) = \exp(A_i \mathbf{x})$. Under the assumptions that the second-order derivative ψ'' is continuous, positive, and bounded above by a constant, we establish the Lipschitz continuity of the gradient of the loss function in (A2) and the convexity in (A3). The boundedness assumption on ψ'' holds for loss functions of most GLMs such as ordinary linear regression, logistic regression, and multinomial regression [29], but not Poisson regression.

3 Theoretical analysis

We present two theoretical results of the proposed model (1). Specifically, Section 3.1 establishes the global optimality of a stationary solution of (1) when the loss function is strongly convex. In Section 3.2, we define restricted strong convexity (RSC), under which the loss function is strongly convex over a subset of the feasible space, followed by providing an upper bound of the distance from the stationary solution to a reference point.

3.1 Optimality of stationary solutions

Notations: For a function $f : \mathbb{R} \to \mathbb{R}$, its derivative at a point t is denoted as $f'(t)$. The notation $f'(t^+)$ represents the right-side derivative at t, i.e.,

$$f'(t^+) \triangleq \lim_{h \to 0^+} \frac{f(t+h) - f(t)}{h}.$$

The notation $F'(\mathbf{x}; \mathbf{d})$, for $F : \mathbb{R}^d \to \mathbb{R}$, is the directional derivative of F at a point $\mathbf{x}$ along the direction $\mathbf{d}$, which is formally defined in Definition 1.

Definition 1. *Given a function $F : \mathbb{R}^d \to \mathbb{R}$, the directional derivative of F at point $\mathbf{x} \in \mathbb{R}^d$ along direction $\mathbf{d} \in \mathbb{R}^d$ is denoted as $F'(\mathbf{x}; \mathbf{d})$ and defined by*

$$F'(\mathbf{x}; \mathbf{d}) \triangleq \lim_{h \to 0^+} \frac{F(\mathbf{x} + h\mathbf{d}) - F(\mathbf{x})}{h}. \tag{3}$$

Next, we provide the definition of a stationary point used in this paper.

Definition 2. *Let $\mathbf{x}^*$ be a stationary point of an unconstrained optimization problem with an objective function $F : \mathbb{R}^d \to \mathbb{R}$ if $F'(\mathbf{x}^*; \mathbf{x} - \mathbf{x}^*) \geq 0, \ \forall \mathbf{x} \in \mathbb{R}^d$.*

Our analysis is built on the stationarity. It has been shown that stationarity is a necessary condition for a point to be a local minimum for certain DC programs [38]. We establish in Theorem 2 that a stationary solution of (1) is a global minimizer under certain conditions. The proof of Theorem 2 requires Lemma 1.

Lemma 1. *If p satisfies Assumption (P3), we have*

$$p(\|\mathbf{y}\|_2) - p(\|\mathbf{x}\|_2) \geq P'(\mathbf{x}; \mathbf{y} - \mathbf{x}) - \frac{\mathrm{Lip_p}}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \quad \forall \mathbf{x}, \ \mathbf{y} \in \mathbb{R}^n, \tag{4}$$

where $P(\mathbf{x}) \triangleq p(\|\mathbf{x}\|_2)$ and n is an ambient dimension of $\mathbf{x}$ and $\mathbf{y}$.

Proof. Let $G(\mathbf{x}) \triangleq p(\|\mathbf{x}\|_2) + \frac{\mathrm{Lip_p}}{2}\|\mathbf{x}\|_2^2$. Observe that $G(\mathbf{x}) = G_1 \circ G_2(\mathbf{x})$ where $G_1(t) \triangleq p(t) + \frac{\mathrm{Lip_p}}{2}t^2$ and $G_2(\mathbf{x}) \triangleq \|\mathbf{x}\|_2$. Then for any $0 < t_1 \leq t_2$, we have

$$\begin{aligned} G_1'(t_1) - G_1'(t_2) &= p'(t_1) - p'(t_2) + \mathrm{Lip_p}(t_1 - t_2) \\ &\leq |p'(t_1) - p'(t_2)| - \mathrm{Lip_p}|t_2 - t_1| \quad \text{since } t_1 \leq t_2 \\ &\leq \mathrm{Lip_p}|t_1 - t_2| - \mathrm{Lip_p}|t_1 - t_2| = 0 \quad \text{by (P3)}, \end{aligned}$$

which implies that $G_1'(t)$ is monotonically non-decreasing on $(0, \infty)$. Therefore, $G_1(t)$ is a convex function on the interval $[0, \infty)$. Since G_1, G_2 are both convex and G_1 is non-decreasing on $[0, \infty)$ by (P1), we conclude that their composite function $G(\mathbf{x})$ is also a convex function. Using the first order condition, i.e., $G(\mathbf{y}) \geq G(\mathbf{x}) + G'(\mathbf{x}; \mathbf{y} - \mathbf{x})$, we have

$$p(\|\mathbf{y}\|_2) + \frac{\mathrm{Lip_p}}{2}\|\mathbf{y}\|_2^2 \geq p(\|\mathbf{x}\|_2) + \frac{\mathrm{Lip_p}}{2}\|\mathbf{x}\|_2^2 + \mathrm{Lip_p}\langle \mathbf{x}, \mathbf{y} - \mathbf{x}\rangle + \mathrm{P}'(\mathbf{x}; \mathbf{y} - \mathbf{x}).$$

After simple manipulations, we deduce the desired inequality (4). □

Theorem 2. *Let Assumptions (P3) and (A3) hold with $\sigma > 0$. If $\sigma \geq \lambda \mathrm{Lip_p} \mathcal{G}^{\max}$, then any stationary solution of (1) is a global minimizer.*

Proof. Denote $\mathbf{x}^*$ as a stationary solution of (1). By Assumption (A3) and applying Lemma 1, we have

$$\begin{aligned} & F_\lambda(\mathbf{x}) - F_\lambda(\mathbf{x}^*) \\ &= L(\mathbf{x}) + \lambda \sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, p(\|\mathbf{x}_{\mathcal{G}_k}\|_2) - L(\mathbf{x}^*) - \lambda \sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, p(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2) \quad \forall \mathbf{x} \in \mathbb{R}^d \\ &\geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2 + \langle \nabla L(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle + \lambda \sum_{k=1}^m \sqrt{|\mathcal{G}_k|} \left\{ p(\|\mathbf{x}_{\mathcal{G}_k}\|_2) - p(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2) \right\} \\ &\geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2 + \langle \nabla L(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle \\ &\qquad + \lambda \sum_{k=1}^m \sqrt{|\mathcal{G}_k|} \left\{ P_k'(\mathbf{x}^*_{\mathcal{G}_k}; \mathbf{x}_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}) - \frac{\mathrm{Lip_p}}{2}\|\mathbf{x}_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\|_2^2 \right\}. \end{aligned} \tag{5}$$

Due to the stationarity (Definition 2), $\mathbf{x}^*$ satisfies

$$\langle \nabla L(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle + \lambda \sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, P_k'(\mathbf{x}^*_{\mathcal{G}_k}; \mathbf{x}_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}) \geq 0. \tag{6}$$

Consequently, the inequality (5) can be simplified as

$$F_\lambda(\mathbf{x}) - F_\lambda(\mathbf{x}^*) \geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2 - \frac{\lambda\, \mathrm{Lip_p}}{2} \sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, \|\mathbf{x}_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\|_2^2$$

$$\geq \frac{\sigma}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2 - \frac{\lambda\,\mathrm{Lip_p}}{2}\mathcal{G}^{\max}\|\mathbf{x}-\mathbf{x}^*\|_2^2 = \frac{\zeta}{2}\|\mathbf{x}-\mathbf{x}^*\|_2^2,$$

where we define $\zeta \triangleq \sigma - \lambda \mathrm{Lip_p}\mathcal{G}^{\max}$. If $\sigma - \lambda \mathrm{Lip_p}\mathcal{G}^{\max} \geq 0$, then $\zeta \geq 0$ and hence $F_\lambda(\mathbf{x}) \geq F_\lambda(\mathbf{x}^*)\ \forall \mathbf{x}$, which implies that any stationary solution $\mathbf{x}^*$ is a global minimizer. □

Our result identifies the condition on the hyperparamter λ which guarantees the global optimality of any stationary solutions. This is a generalization of [22, Theorem 2.2] which is for a special case of the group-LOG regularizer.

3.2 Stationarity under restricted strong convexity

The assumption of strong convexity may not hold for some loss functions, e.g., ordinary linear regression with an under-determined system of linear equations. As a remedy, we consider a setting where strong convexity only holds over a smaller region rather than the entire domain of the loss function. Since the scope of the strong convexity is limited, such an assumption is referred to as the restricted strong convexity (RSC) [18]. In statistics, the RSC condition for a subset of possible vector differences between the ground-truth and an estimator is imposed [35].

Let us define a reference point, denoted by $\mathbf{x}^\epsilon$, as a vector that, for $\epsilon > 0$, satisfies

$$\|\nabla L(\mathbf{x})\|_2 \leq \epsilon. \tag{7}$$

One motivation for defining such a point is that if the ground-truth were to exist, then it would satisfy (7) with high probability; such a connection between the ground-truth and an empirical solution of the sample average approximations is explained in [2] and [46, Chapter 7].

We first derive a region for RSC to hold, then provide a bound of the distance from a stationary solution to a reference point. Let S be the group-wise support set of the reference point $\mathbf{x}^\epsilon$, i.e., $S \triangleq \{k \in \{1,\dots,m\} \mid \|\mathbf{x}^\epsilon_{\mathcal{G}_k}\|_2 \neq 0\}$. Given $\delta > 0$, define the set $\mathcal{V}_\delta(S)$ as

$$\mathcal{V}_\delta(S) \triangleq \left\{ \boldsymbol{\nu} \;\middle|\; \sum_{k\notin S}\|\boldsymbol{\nu}_{\mathcal{G}_k}\|_2 \leq \delta \sum_{k\in S}\|\boldsymbol{\nu}_{\mathcal{G}_k}\|_2 \right\}.$$

It is not difficult to verify that $\mathcal{V}_\delta(S)$ is a cone; i.e., if $\mathbf{x} \in \mathcal{V}_\delta(S)$ then $\alpha\mathbf{x} \in \mathcal{V}_\delta(S)$ for any $\alpha \geq 0$. Furthermore, the set is a nonconvex cone. For example, let $\delta = 0.5, \mathcal{G}_1 = \{1,2\}, \mathcal{G}_2 = \{3,4\}$ and $S = \{2\}$. For two points $\nu_1 = (1,1,1,3)^T$ and $\nu_2 = (1,2,4,2)^T$, we have $\nu_1, \nu_2 \in \mathcal{V}_\delta(S)$ but $0.5\nu_1 + 0.5\nu_2 \notin \mathcal{V}_\delta(S)$.

Given a reference point $\mathbf{x}^\epsilon$, there exists a region that includes the vector differences between the stationary solutions of (1) and $\mathbf{x}^\epsilon$ under certain conditions; refer to the following lemma.

Lemma 3. *Let Assumptions (A3) and (P2) hold. Given a positive scalar ϵ, let $\mathbf{x}^\epsilon$ be a vector that satisfies the inequality (7). If $\mathbf{x}^*$ be a stationary solution of (1) with*

$$\epsilon < \min_{k\notin S}\sqrt{|\mathcal{G}_k|}\,\lambda\, p'(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2), \tag{8}$$

then there exists $\delta^* > 0$ such that $(\mathbf{x}^\epsilon - \mathbf{x}^*) \in \mathcal{V}_{\delta^*}(S)$.

Proof. For any stationary solution $\mathbf{x}^*$, one has

$$\langle \nabla L(\mathbf{x}^*), \mathbf{x}^\epsilon - \mathbf{x}^*\rangle + \lambda \sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, P'_k(\mathbf{x}^*_{\mathcal{G}_k}; \mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}) \geq 0. \tag{9}$$

For each summand in the second term of the left-hand side of (9), we split into two cases. If $\mathbf{x}^*_{\mathcal{G}_k} \neq \mathbf{0}$, we use (P2) to obtain

$$\begin{aligned} &P'_k(\mathbf{x}^*_{\mathcal{G}_k}; \mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}) \\ =&\left\langle p'(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2)\frac{\mathbf{x}^*_{\mathcal{G}_k}}{\|\mathbf{x}^*_{\mathcal{G}_k}\|_2}, \mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\right\rangle \begin{cases} \leq p'(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2)\|\mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\|_2, & \text{if } k \in S \\ = -p'(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2)\|\mathbf{x}^*_{\mathcal{G}_k}\|_2, & \text{if } k \notin S. \end{cases} \end{aligned} \tag{10}$$

If $\mathbf{x}^*_{\mathcal{G}_k} = \mathbf{0}$, it follows from the definition of the directional derivative that

$$P'_k(\mathbf{0}; \mathbf{x}^\epsilon_{\mathcal{G}_k}) = p'(0^+)\|\mathbf{x}^\epsilon_{\mathcal{G}_k}\|_2. \tag{11}$$

By (7) and the convexity of L, we obtain

$$\begin{aligned}\langle \nabla L(\mathbf{x}^*), \mathbf{x}^\epsilon - \mathbf{x}^*\rangle &\le \langle \nabla L(\mathbf{x}^\epsilon), \mathbf{x}^\epsilon - \mathbf{x}^*\rangle \\ &\le \|\nabla L(\mathbf{x}^\epsilon)\|_2 \|\mathbf{x}^\epsilon - \mathbf{x}^*\|_2 \le \epsilon \sum_{k=1}^m \|\mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\|_2. \end{aligned} \tag{12}$$

Substituting (10)-(12) into (9) yields

$$\begin{aligned}&\sum_{k\notin S} \sqrt{|\mathcal{G}_k|} p'(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2)\|\mathbf{x}^*_{\mathcal{G}_k}\|_2 - \frac{\epsilon}{\lambda}\sum_{k=1}^m \|\mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\|_2 \\ &\quad\le \sum_{\substack{k\in S,\\ \mathbf{x}^*_{\mathcal{G}_k}\neq \mathbf{0}}} \sqrt{|\mathcal{G}_k|} p'(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2)\|\mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\|_2 + \sum_{\substack{k\in S,\\ \mathbf{x}^*_{\mathcal{G}_k}= \mathbf{0}}} \sqrt{|\mathcal{G}_k|} p'(0^+)\|\mathbf{x}^\epsilon_{\mathcal{G}_k}\|_2.\end{aligned}$$

By the assumption (P2) where u is defined, we deduce

$$\sum_{k\notin S}\left(\sqrt{|\mathcal{G}_k|}\, p'(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2) - \frac{\epsilon}{\lambda}\right)\|\mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\|_2 \le \sum_{k\in S}\left(\frac{\epsilon}{\lambda} + u\sqrt{|\mathcal{G}_k|}\right)\|\mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\|_2.$$

The condition (8) guarantees the nonnegativity of the left-hand side of the inequality, which validates the existence of δ^*. □

Lemma 3 can be interpreted as, if there is a stationary solution that is sufficiently close to $\mathbf{x}^\epsilon$, then there exists a nonconvex cone which includes the direction $\mathbf{x}^\epsilon - \mathbf{x}^*$. This is because a stationary solution $\mathbf{x}^*_{\mathcal{G}_k}$ is more likely to meet the condition (8) if $\|\mathbf{x}^*_{\mathcal{G}_k}\|_2$ is near the origin whenever the corresponding subvector of $\mathbf{x}^\epsilon_{\mathcal{G}_k}$ is zero. Provided that the problem (1) has finitely many such stationary solutions, there exists $\mathcal{V}_\delta(S)$ that includes all directions from the stationary solutions to the reference point.

We define the RSC assumption of L over the set $\mathcal{V}_\delta(S)$:

(A4). There exists $\sigma > 0$ such that $L(\mathbf{x}_1) - L(\mathbf{x}_2) - \langle \nabla L(\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2\rangle \ge \frac{\sigma}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \quad \forall (\mathbf{x}_1 - \mathbf{x}_2) \in \mathcal{V}_\delta(S)$.

Under Assumption (A4), we provide a bound of the distance between a stationary solution of (1) and the reference point $\mathbf{x}^\epsilon$.

Theorem 4. *Let Assumptions (A4) and (P2) hold. Given a positive scalar ϵ, let $\mathbf{x}^\epsilon$ be a vector that satisfies the inequality (7). Let $\mathbf{x}^*$ be a stationary solution of (1) with*

$$\epsilon < \min_{k\notin S} \sqrt{|\mathcal{G}_k|}\, \lambda\, p'(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2). \tag{13}$$

Suppose (A4) holds over the set $\mathcal{V}_{\delta^*}(S) \ni (\mathbf{x}^\epsilon - \mathbf{x}^*)$. We have a bound

$$\|\mathbf{x}^\epsilon - \mathbf{x}^*\|_2 \le \frac{4\lambda u}{\sigma} \max_{k\in S}\sqrt{|\mathcal{G}_k|}\sqrt{\sum_{k\in S}|\mathcal{G}_k|}.$$

Proof. By the RSC assumption (A4), there exists $\sigma > 0$ such that

$$\begin{aligned}&\frac{\sigma}{2}\|\mathbf{x}^\epsilon - \mathbf{x}^*\|_2^2 \\ &\le L(\mathbf{x}^\epsilon) - L(\mathbf{x}^*) - \langle \nabla L(\mathbf{x}^*), \mathbf{x}^\epsilon - \mathbf{x}^*\rangle \\ &\le L(\mathbf{x}^\epsilon) - L(\mathbf{x}^*) + \lambda\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, P'_k(\mathbf{x}^*_{\mathcal{G}_k}; \mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}) \text{ by (9)}\end{aligned}$$

$$\leq \langle \nabla L(\mathbf{x}^\epsilon), \mathbf{x}^\epsilon - \mathbf{x}^* \rangle + \lambda \sum_{k=1}^{m} \sqrt{|\mathcal{G}_k|}\, P_k'(\mathbf{x}^*_{\mathcal{G}_k}; \mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}) \text{ by (A3)}$$
$$\leq \sum_{k \notin S} \Big(\epsilon - \lambda \sqrt{|\mathcal{G}_k|} p'(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2) \Big) \|\mathbf{x}^*_{\mathcal{G}_k}\|_2 + \sum_{k \in S} (\epsilon + \lambda u \sqrt{|\mathcal{G}_k|}) \|\mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\|_2.$$

The last inequality is obtained by applying (10)-(12). Due to (13), the first term of the right-hand side of the inequality is negative and $\epsilon < \lambda u \sqrt{|\mathcal{G}_k|}$, leading to

$$\begin{aligned}
\frac{\sigma}{2} \|\mathbf{x}^\epsilon - \mathbf{x}^*\|_2^2 &\leq 2\, \lambda u \max_{k \in S} \sqrt{|\mathcal{G}_k|} \sum_{k \in S} \|\mathbf{x}^\epsilon_{\mathcal{G}_k} - \mathbf{x}^*_{\mathcal{G}_k}\|_2 \\
&\leq 2\, \lambda u \max_{k \in S} \sqrt{|\mathcal{G}_k|} \|\mathbf{x}^\epsilon_{\mathcal{G}_S} - \mathbf{x}^*_{\mathcal{G}_S}\|_1 \text{ where } \mathcal{G}_S \triangleq \bigcup_{k \in S} \mathcal{G}_k \\
&\leq 2\, \lambda u \max_{k \in S} \sqrt{|\mathcal{G}_k|} \sqrt{|\mathcal{G}_S|} \|\mathbf{x}^\epsilon_{\mathcal{G}_S} - \mathbf{x}^*_{\mathcal{G}_S}\|_2 \\
&\leq 2\, \lambda u \max_{k \in S} \sqrt{|\mathcal{G}_k|} \sqrt{|\mathcal{G}_S|} \|\mathbf{x}^\epsilon - \mathbf{x}^*\|_2.
\end{aligned}$$

By dividing $\|\mathbf{x}^\epsilon - \mathbf{x}^*\|_2$ on both sides and substituting $|\mathcal{G}_S| = \sum_{k \in S} |\mathcal{G}_k|$, we complete the proof. □

Theorem 3 is a generalization of existing bounds shown for individual sparsity problems (without group structure) [2, 18], for which every group is a singleton such that $\mathcal{G}_1 = \{1\}, \ldots, \mathcal{G}_d = \{d\}$. For example, if we let $p(t) = t$, the problem (1) becomes LASSO regularization, and Theorem 3 exactly recovers the bound on the distance between the optimal solution of LASSO and the ground-truth shown in [18, Theorem 11.1]. The result also extends the bound derived in [2, Theorem 1] for nonconvex sparsity functions such as SCAD, MCP, and transformed ℓ_1.

4 Our algorithm

We adopt the alternating direction method of multipliers (ADMM) [6] to minimize the problem (1). Specifically, we introduce an auxiliary variable $\mathbf{z}$ and rewrite (1) equivalently as

$$\min_{\mathbf{x},\mathbf{z}} \; L(\mathbf{z}) + \lambda \sum_{k=1}^{m} \sqrt{|\mathcal{G}_k|}\, p(\|\mathbf{x}_{\mathcal{G}_k}\|_2) \qquad \text{s.t.} \qquad \mathbf{x} = \mathbf{z}. \tag{14}$$

The corresponding augmented Lagrangian function is

$$\mathcal{L}(\mathbf{x}, \mathbf{z}; \mathbf{v}) \triangleq L(\mathbf{z}) + \lambda \sum_{k=1}^{m} \sqrt{|\mathcal{G}_k|}\, p(\|\mathbf{x}_{\mathcal{G}_k}\|_2) + \rho \langle \mathbf{v}, \mathbf{x} - \mathbf{z} \rangle + \frac{\rho}{2} \|\mathbf{x} - \mathbf{z}\|_2^2, \tag{15}$$

where $\mathbf{v}$ is a Lagrangian multiplier (or dual variable) and ρ is a positive parameter. We consider a scaled form [6] in (15) by multiplying ρ in front of $\langle \mathbf{v}, \mathbf{x} - \mathbf{z} \rangle$. Consequently, ADMM iterations proceed as follows:

$$\begin{cases} \mathbf{x}^{\tau+1} \in \operatorname{argmin}_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{z}^\tau; \mathbf{v}^\tau) \\ \mathbf{z}^{\tau+1} \in \operatorname{argmin}_{\mathbf{z}} \mathcal{L}(\mathbf{x}^{\tau+1}, \mathbf{z}; \mathbf{v}^\tau) \\ \mathbf{v}^{\tau+1} = \mathbf{v}^\tau + \mathbf{x}^{\tau+1} - \mathbf{z}^{\tau+1}, \end{cases} \tag{16}$$

where τ indexes the iteration number. The $\mathbf{z}$-subproblem is written as

$$\mathbf{z}^{\tau+1} \in \operatorname*{argmin}_{\mathbf{z}} L(\mathbf{z}) + \frac{\rho}{2} \|\mathbf{x}^{\tau+1} - \mathbf{z} + \mathbf{v}^\tau\|_2^2,$$

which is convex under Assumption (A3) and hence can be solved efficiently by existing convex programming algorithms. For example, a closed-form solution can be derived if the loss function is the least squares for linear regression. In Appendix A, we provide details on how the $\mathbf{z}$-subproblem is solved for various GLM loss functions that are considered in the numerical study in Section 5. In Section 4.1, we elaborate on how to solve $\mathbf{x}$-subproblem in (16), and the convergence analysis of the ADMM scheme (16) is conducted in Section 4.2.

4.1 x-subproblem

The $\mathbf{x}$-subproblem can be decomposed into groups such that for each $k \in \{1, \cdots, m\}$,

$$\mathbf{x}_{\mathcal{G}_k}^{\tau+1} \in \underset{\mathbf{x}\in\mathbb{R}^{|\mathcal{G}_k|}}{\operatorname{argmin}} \lambda\sqrt{|\mathcal{G}_k|}p(\|\mathbf{x}\|_2) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}_{\mathcal{G}_k}^{\tau} + \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2^2 \triangleq H_{\lambda,\rho}(\mathbf{x}). \tag{17}$$

It is nonconvex due to Assumption (P1), by which a (global) optimal solution may be difficult to obtain. Corollary 5 characterizes conditions under which a stationary solution of (17) achieves the global optimality.

Corollary 5. *Let Assumption (P3) hold. If $\rho \geq \lambda \mathrm{Lip}_{\mathrm{p}}\mathcal{G}^{\max}$, then any stationary solution of (17) is a global minimizer.*

Proof. Since ρ is the strong convexity modulus of the second term in (17), the statement follows from Theorem 2. □

To solve (17), we introduce a general update scheme based on a proximal operator [5, Chapter 6], then discuss the stationarity of the obtained solution. The proximal operator of a function f is defined by

$$\mathbf{prox}_f(y;\mu) \in \underset{x}{\operatorname{argmin}} \left(\mu f(x) + \frac{1}{2}(x-y)^2\right), \tag{18}$$

where f is a univariate function and μ is a positive parameter. The notation *argmin* we use in (17) and (18) denotes a set of stationary solutions instead of the (global) optimal solutions. Specifically assuming f is a differentiable function except at 0, we define a nonzero stationary solution $\bar{x}$ if $\mu f'(\bar{x}) + (\bar{x} - y) = 0$. If $\bar{x} = 0$ is a stationary solution to (18), then there exists $\bar{v} \in \partial f(\bar{x}) = \{v | f(z) \geq f(\bar{x}) + v(z - \bar{x}),\ \forall z\ \}$ such that $\mu\bar{v} + (\bar{x} - y) = 0$. It has been shown that (18) yields a global solution for certain sparsity functions such as transformed ℓ_1 function [65, Theorem 3.1]. Multiple stationary solutions are possible for ℓ_1-ℓ_2 function [30].

We aim to find a closed-form solution to (17), denoted by $\mathbf{x}^*$. If $\mathbf{x}^* \neq \mathbf{0}$, then it satisfies the stationary condition:

$$\lambda\sqrt{|\mathcal{G}_k|}p'(\|\mathbf{x}^*\|_2)\frac{\mathbf{x}^*}{\|\mathbf{x}^*\|_2} + \rho(\mathbf{x}^* - \mathbf{z}_{\mathcal{G}_k}^{\tau} + \mathbf{v}_{\mathcal{G}_k}^{\tau}) = \mathbf{0}. \tag{19}$$

With some algebra, we deduce from (19) that $\mathbf{x}^* = c(\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau})$ for a scalar $c > 0$. The problem (17) can be written as a univariate minimization over c,

$$\min_c \lambda\sqrt{|\mathcal{G}_k|}p(c\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2) + \frac{\rho(c-1)^2}{2}\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2^2. \tag{20}$$

It can be solved by a proximal operator, i.e.,

$$c^* = \frac{1}{\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2}\mathbf{prox}_p\left(\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2; \frac{\lambda\sqrt{|\mathcal{G}_k|}}{\rho}\right). \tag{21}$$

If $c^* = 0$ is a stationary point to (20), we present in Proposition 6 that $\mathbf{x}^* = \mathbf{0}$ is a stationary point to (17). Consequently, the $\mathbf{x}$-update can be expressed as follows,

$$\mathbf{x}_{\mathcal{G}_k}^{\tau+1} = \frac{\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}}{\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2}\mathbf{prox}_p\left(\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2; \frac{\lambda\sqrt{|\mathcal{G}_k|}}{\rho}\right). \tag{22}$$

Proposition 6. *If the proximal operator in (21) returns a stationary solution, then $\mathbf{x}_{\mathcal{G}_k}^{\tau+1}$ in (22) is a stationary solution of the $\mathbf{x}$-subproblem (17).*

Proof. Suppose c^* defined in (21) is a stationary solution to (20). If $c^* \neq 0$, then it satisfies

$$\lambda\sqrt{|\mathcal{G}_k|}p'(c^*\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2) = \rho(1 - c^*)\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2,$$

Algorithm 1 Group ADMM Framework

Set hyperparameter $\lambda > 0$ and ADMM penalty parameter $\rho > 0$;
Initialize $\mathbf{x}^0, \mathbf{z}^0, \mathbf{v}^0$, $\tau = 0$.
repeat
  for $k = 1, 2, \ldots, m$ **do**
    update $\mathbf{x}_{\mathcal{G}_k}^{\tau+1}$ by (22)
  end
  $\mathbf{z}^{\tau+1} \in \underset{\mathbf{z}}{\operatorname{argmin}}\, L(\mathbf{z}) + \dfrac{\rho}{2}\|\mathbf{x}^{\tau+1} - \mathbf{z} + \mathbf{v}^{\tau}\|_2^2$
  $\mathbf{v}^{\tau+1} = \mathbf{v}^{\tau} + \mathbf{x}^{\tau+1} - \mathbf{z}^{\tau+1}$
  $\tau = \tau + 1$
until *convergence*;

which implies that $0 < c^* < 1$. Hence the corresponding $\mathbf{x}^* = c^*(\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau})$ satisfies the stationary condition given in (19).

To discuss the case of $c^* = 0$, we consider the directional derivative of the objective function for $\mathbf{x}$-update in (17) at $\mathbf{x}^*$ in the direction $\mathbf{x} - \mathbf{x}^*$ for any $\mathbf{x}$, i.e.,

$$\begin{aligned} & H'_{\lambda,\rho}(\mathbf{x}^*; \mathbf{x} - \mathbf{x}^*) \\ =& \lambda\sqrt{|\mathcal{G}_k|}P'_k(\mathbf{x}^*; \mathbf{x} - \mathbf{x}^*) + \rho\langle \mathbf{x}^* - \mathbf{z}_{\mathcal{G}_k}^{\tau} + \mathbf{v}_{\mathcal{G}_k}^{\tau}, \mathbf{x} - \mathbf{x}^*\rangle \\ \geq& \lambda\sqrt{|\mathcal{G}_k|}\lim_{h\to 0^+}\frac{p(\|\mathbf{x}^* + h(\mathbf{x} - \mathbf{x}^*)\|_2) - p(\|\mathbf{x}^*\|_2)}{h} - \rho\|\mathbf{x}^* - \mathbf{z}_{\mathcal{G}_k}^{\tau} + \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2\,\|\mathbf{x} - \mathbf{x}^*\|_2. \end{aligned}$$

Letting $\mathbf{x}^* = \mathbf{0}$ with $p(0) = 0$, we have

$$\begin{aligned} H'_{\lambda,\rho}(\mathbf{0}; \mathbf{x}) &\geq \lambda\sqrt{|\mathcal{G}_k|}\lim_{h\to 0^+}\frac{p(h\|\mathbf{x}\|_2)}{h} - \rho\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2\,\|\mathbf{x}\|_2 \\ &= \lambda\sqrt{|\mathcal{G}_k|}p'(0^+)\|\mathbf{x}\|_2 - \rho\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2\,\|\mathbf{x}\|_2 \quad \text{by L'hôpital's rule.} \end{aligned} \tag{23}$$

If $c^* = 0$ is a stationary solution to (20), there exists $\bar{u} \in \partial p(0)$ such that $\lambda\sqrt{|\mathcal{G}_k|}\bar{u} - \rho\|\mathbf{z}_{\mathcal{G}_k}^{\tau} - \mathbf{v}_{\mathcal{G}_k}^{\tau}\|_2 = 0$. Combining this with (23) provides

$$H'_{\lambda,\rho}(\mathbf{0}; \mathbf{x}) \geq \lambda\sqrt{|\mathcal{G}_k|}\|\mathbf{x}\|_2\Big(p'(0^+) - \bar{u}\Big).$$

It remains to show that $p'(0^+) \geq u$ for all $u \in \partial p(0)$. Assume there exists $\hat{u} \in \partial p(0)$ such that $p'(0^+) < \hat{u}$. From the property of the subgradient, we have $p(t) \geq p(0) + \hat{u}(t - 0)$ for all $t \in \mathbb{R}$. If we choose a strictly positive $\hat{t}$, then we must have $p(\hat{t}) \geq p(0) + \hat{u}(\hat{t} - 0) > p(0) + p'(0^+)(\hat{t} - 0)$. This contradicts the concavity of p on the domain $[0, \infty)$. Hence we conclude that if $c^* = 0$ is a stationary solution to (20), then $H'_{\lambda,\rho}(\mathbf{0}; \mathbf{x}) \geq 0$ for any $\mathbf{x}$. □

Proposition 6 and Corollary 5 indicate that if Assumption (P3) holds with $\rho \geq \lambda \mathrm{Lip_p}\mathcal{G}^{\max}$, then $\mathbf{x}_{\mathcal{G}_k}^{\tau+1}$ is a global minimizer to the problem (17).

4.2 Convergence analysis

The ADMM framework for minimizing (14) that involve both $\mathbf{x}$- and $\mathbf{z}$-subproblems is described in Algorithm 1. In this section, we present its convergence analysis. We first show that each $\mathbf{x}$- and $\mathbf{z}$-update decreases its objective value, followed by monotontic decreasing of $\{\mathcal{L}(\mathbf{x}^{\tau}, \mathbf{z}^{\tau}; \mathbf{v}^{\tau})\}$; refer to Lemmas 7, 8, and 10, respectively.

Lemma 7. *Let Assumption (P3) hold. If $\rho > \lambda \mathrm{Lip_p}\mathcal{G}^{\max}$, then for any $\mathbf{x}^{\tau+1}$ given by (16), there exists $\bar{c}_1 > 0$ such that*

$$\mathcal{L}(\mathbf{x}^{\tau+1}, \mathbf{z}^{\tau}; \mathbf{v}^{\tau}) - \mathcal{L}(\mathbf{x}^{\tau}, \mathbf{z}^{\tau}; \mathbf{v}^{\tau}) \leq -\frac{\bar{c}_1}{2}\|\mathbf{x}^{\tau+1} - \mathbf{x}^{\tau}\|_2^2. \tag{24}$$

Proof. By Proposition 6, $\mathbf{x}^{\tau+1}$ is a stationary solution to (16) such that

$$\lambda\sum_{k=1}^{m}\sqrt{|\mathcal{G}_k|}\,P'(\mathbf{x}^{\tau+1}_{\mathcal{G}_k};\mathbf{x}^{\tau}_{\mathcal{G}_k}-\mathbf{x}^{\tau+1}_{\mathcal{G}_k})+\rho\langle\mathbf{x}^{\tau+1}-\mathbf{z}^{\tau}+\mathbf{v}^{\tau},\mathbf{x}^{\tau}-\mathbf{x}^{\tau+1}\rangle\geq 0. \tag{25}$$

It follows from Lemma 1 that

$$p(\|\mathbf{x}^{\tau+1}_{\mathcal{G}_k}\|_2)-p(\|\mathbf{x}^{\tau}_{\mathcal{G}_k}\|_2)\leq\frac{\mathrm{Lip_p}}{2}\|\mathbf{x}^{\tau+1}_{\mathcal{G}_k}-\mathbf{x}^{\tau}_{\mathcal{G}_k}\|_2^2-P'(\mathbf{x}^{\tau+1}_{\mathcal{G}_k};\mathbf{x}^{\tau}_{\mathcal{G}_k}-\mathbf{x}^{\tau+1}_{\mathcal{G}_k}). \tag{26}$$

Simple calculations lead to

$$\begin{aligned}
&\mathcal{L}(\mathbf{x}^{\tau+1},\mathbf{z}^{\tau};\mathbf{v}^{\tau})-\mathcal{L}(\mathbf{x}^{\tau},\mathbf{z}^{\tau};\mathbf{v}^{\tau})\\
=&\lambda\sum_{k=1}^{m}\sqrt{|\mathcal{G}_k|}\left(p(\|\mathbf{x}^{\tau+1}_{\mathcal{G}_k}\|_2)-p(\|\mathbf{x}^{\tau}_{\mathcal{G}_k}\|_2)\right)\\
&\qquad+\frac{\rho}{2}\left(\|\mathbf{x}^{\tau+1}-\mathbf{z}^{\tau}+\mathbf{v}^{\tau}\|_2^2-\|\mathbf{x}^{\tau}-\mathbf{z}^{\tau}+\mathbf{v}^{\tau}\|_2^2\right)\\
\leq&\frac{\lambda\mathrm{Lip_p}}{2}\sum_{k=1}^{m}\sqrt{|\mathcal{G}_k|}\|\mathbf{x}^{\tau+1}_{\mathcal{G}_k}-\mathbf{x}^{\tau}_{\mathcal{G}_k}\|_2^2-\lambda\sum_{k=1}^{m}\sqrt{|\mathcal{G}_k|}\,P'(\mathbf{x}^{\tau+1}_{\mathcal{G}_k};\mathbf{x}^{\tau}_{\mathcal{G}_k}-\mathbf{x}^{\tau+1}_{\mathcal{G}_k})\\
&\qquad+\frac{\rho}{2}\langle\mathbf{x}^{\tau+1}+\mathbf{x}^{\tau}-2\mathbf{z}^{\tau}+2\mathbf{v}^{\tau},\mathbf{x}^{\tau+1}-\mathbf{x}^{\tau}\rangle\text{ by (26)}\\
\leq&\frac{\lambda\mathrm{Lip_p}}{2}\mathcal{G}^{\max}\|\mathbf{x}^{\tau+1}-\mathbf{x}^{\tau}\|_2^2+\frac{\rho}{2}\langle\mathbf{x}^{\tau}-\mathbf{x}^{\tau+1},\mathbf{x}^{\tau+1}-\mathbf{x}^{\tau}\rangle\text{ by (25)}\\
=&-\frac{\rho-\lambda\mathrm{Lip_p}\mathcal{G}^{\max}}{2}\|\mathbf{x}^{\tau+1}-\mathbf{x}^{\tau}\|_2^2.
\end{aligned}$$

If $\rho>\lambda\mathrm{Lip_p}\mathcal{G}^{\max}$, we choose $\bar{c}_1=\frac{1}{2}(\rho-\lambda\mathrm{Lip_p}\mathcal{G}^{\max})>0$ such that (24) holds. □

Lemma 8. *Let Assumption (A3) hold. There exists $\bar{c}_2>0$ such that*

$$\mathcal{L}(\mathbf{x}^{\tau+1},\mathbf{z}^{\tau+1};\mathbf{v}^{\tau})-\mathcal{L}(\mathbf{x}^{\tau+1},\mathbf{z}^{\tau};\mathbf{v}^{\tau})\leq-\frac{\bar{c}_2}{2}\|\mathbf{z}^{\tau+1}-\mathbf{z}^{\tau}\|_2^2.$$

Proof. The optimality condition of $\mathbf{z}^{\tau+1}$ is

$$\nabla L(\mathbf{z}^{\tau+1})-\rho(\mathbf{x}^{\tau+1}-\mathbf{z}^{\tau+1}+\mathbf{v}^{\tau})=\mathbf{0}. \tag{27}$$

It follows from Assumption (A3) that

$$\begin{aligned}
&\mathcal{L}(\mathbf{x}^{\tau+1},\mathbf{z}^{\tau+1};\mathbf{v}^{\tau})-\mathcal{L}(\mathbf{x}^{\tau+1},\mathbf{z}^{\tau};\mathbf{v}^{\tau})\\
=&L(\mathbf{z}^{\tau+1})-L(\mathbf{z}^{\tau})+\frac{\rho}{2}\left(\|\mathbf{x}^{\tau+1}-\mathbf{z}^{\tau+1}+\mathbf{v}^{\tau}\|_2^2-\|\mathbf{x}^{\tau+1}-\mathbf{z}^{\tau}+\mathbf{v}^{\tau}\|_2^2\right)\\
\leq&-\langle\nabla L(\mathbf{z}^{\tau+1}),\mathbf{z}^{\tau}-\mathbf{z}^{\tau+1}\rangle-\frac{\sigma}{2}\|\mathbf{z}^{\tau+1}-\mathbf{z}^{\tau}\|_2^2\\
&\qquad+\frac{\rho}{2}\langle 2\mathbf{x}^{\tau+1}-\mathbf{z}^{\tau+1}-\mathbf{z}^{\tau}+2\mathbf{v}^{\tau},\mathbf{z}^{\tau}-\mathbf{z}^{\tau+1}\rangle\text{ by (A3)}\\
=&-\frac{\sigma}{2}\|\mathbf{z}^{\tau+1}-\mathbf{z}^{\tau}\|_2^2+\frac{\rho}{2}\langle\mathbf{z}^{\tau+1}-\mathbf{z}^{\tau},\mathbf{z}^{\tau}-\mathbf{z}^{\tau+1}\rangle\text{ by (27)}\\
=&-\frac{\rho+\sigma}{2}\|\mathbf{z}^{\tau+1}-\mathbf{z}^{\tau}\|_2^2.
\end{aligned}$$

As $\rho>0$ and $\sigma\geq 0$, we choose $\bar{c}_2=\rho+\sigma>0$ that completes the proof. □

Lemma 9. *Let Assumption (A2) hold. We have*

$$\|\mathbf{v}^{\tau+1}-\mathbf{v}^{\tau}\|_2^2\leq\frac{{\mathrm{Lip_L}}^2}{\rho^2}\|\mathbf{z}^{\tau+1}-\mathbf{z}^{\tau}\|_2^2.$$

Proof. Based on (27) and $\mathbf{v}$-update formula $\mathbf{v}^{\tau+1} = \mathbf{v}^{\tau} + \mathbf{x}^{\tau+1} - \mathbf{z}^{\tau+1}$, we obtain

$$\nabla L(\mathbf{z}^{\tau+1}) - \rho \mathbf{v}^{\tau+1} = 0, \tag{28}$$

or equivalently $\mathbf{v}^{\tau+1} = \frac{\nabla L(\mathbf{z}^{\tau+1})}{\rho}$. Similarly, we have $\mathbf{v}^{\tau} = \frac{\nabla L(\mathbf{z}^{\tau})}{\rho}$. It follows from Assumption (A2) that

$$\|\mathbf{v}^{\tau+1} - \mathbf{v}^{\tau}\|_2 = \frac{1}{\rho}\|\nabla L(\mathbf{z}^{\tau+1}) - \nabla L(\mathbf{z}^{\tau})\|_2 \leq \frac{\mathrm{Lip_L}}{\rho}\|\mathbf{z}^{\tau+1} - \mathbf{z}^{\tau}\|_2. \tag{29}$$

□

Combining Lemmas 7-9, we shown in Lemma 10 that every triplet $(\mathbf{x}^{\tau+1}, \mathbf{z}^{\tau+1}; \mathbf{v}^{\tau+1})$ produced by (16) sufficiently decreases the objective value of (15).

Lemma 10. *(sufficient descent) Let Assumptions (P3), (A2) and (A3) hold. If ρ in (15) satisfies $\rho > \max\left\{\lambda \mathrm{Lip}_p \mathcal{G}^{\max}, \frac{\sqrt{\sigma^2 + 8\mathrm{Lip_L}^2} - \sigma}{2}\right\}$, then there exist two constants $c_1 > 0$ and $c_2 > 0$ such that*

$$\mathcal{L}(\mathbf{x}^{\tau+1}, \mathbf{z}^{\tau+1}; \mathbf{v}^{\tau+1}) - \mathcal{L}(\mathbf{x}^{\tau}, \mathbf{z}^{\tau}; \mathbf{v}^{\tau}) \leq -c_1\|\mathbf{x}^{\tau+1} - \mathbf{x}^{\tau}\|_2^2 - c_2\|\mathbf{z}^{\tau+1} - \mathbf{z}^{\tau}\|_2^2. \tag{30}$$

Proof. It follows from $\mathbf{v}$-update formula $\mathbf{v}^{\tau+1} - \mathbf{v}^{\tau} = \mathbf{x}^{\tau+1} - \mathbf{z}^{\tau+1}$ that

$$\begin{aligned}
&\mathcal{L}(\mathbf{x}^{\tau+1}, \mathbf{z}^{\tau+1}; \mathbf{v}^{\tau+1}) - \mathcal{L}(\mathbf{x}^{\tau+1}, \mathbf{z}^{\tau+1}; \mathbf{v}^{\tau}) \\
=&\langle \rho \mathbf{v}^{\tau+1}, \mathbf{x}^{\tau+1} - \mathbf{z}^{\tau+1}\rangle - \langle \rho \mathbf{v}^{\tau}, \mathbf{x}^{\tau+1} - \mathbf{z}^{\tau+1}\rangle \\
=&\rho\langle \mathbf{v}^{\tau+1} - \mathbf{v}^{\tau}, \mathbf{x}^{\tau+1} - \mathbf{z}^{\tau+1}\rangle \\
=&\rho\|\mathbf{v}^{\tau+1} - \mathbf{v}^{\tau}\|_2^2 \leq \frac{\mathrm{Lip_L}^2}{\rho}\|\mathbf{z}^{\tau+1} - \mathbf{z}^{\tau}\|_2^2,
\end{aligned}$$

where the last inequality holds by Lemma 9. By applying Lemmas 7-8, we achieve the desired inequality (30) with $c_1 = \frac{\bar{c}_1}{2}$ and $c_2 = \frac{\bar{c}_2}{2} - \frac{\mathrm{Lip_L}^2}{\rho}$, where $\bar{c}_1$ and $\bar{c}_2$ are defined in Lemma 7 and Lemma 8, respectively. The condition $\rho > \lambda \mathrm{Lip}_p \mathcal{G}^{\max}$ is required for Lemma 7. Additionally we require $\rho > \frac{\sqrt{\sigma^2 + 8\mathrm{Lip_L}^2} - \sigma}{2}$ such that $c_2 > 0$. □

Theorem 11 establishes the subsequence convergence of the iterates under an assumption that the objective function $F_\lambda(\cdot)$ is coercive; refer to Definition 3. If either the regularization function P or the loss function L is coercive, then the objective function is coercive, which guarantees the boundedness of the minimizing sequence.

Definition 3. *A function $f(\cdot)$ is coercive if $f(\mathbf{x}) \to \infty$ as $\|\mathbf{x}\|_2 \to \infty$.*

Theorem 11. *(convergence) Let Assumptions (P3), (A1)-(A3) hold. If either P or L is coercive, and ρ in (15) satisfies $\rho > \max\left\{\lambda \mathrm{Lip}_p \mathcal{G}^{\max}, \frac{\sqrt{\sigma^2 + 8\mathrm{Lip_L}^2} - \sigma}{2}\right\}$, then the sequence $\{(\mathbf{x}^{\tau}, \mathbf{z}^{\tau}, \mathbf{v}^{\tau})\}_{\tau=1}^{\infty}$ generated by (16) has a convergent subsequence. Moreover, its limit point is a stationary solution of the problem (1).*

Proof. We first show the convergence of the sequence $\{(\mathbf{x}^{\tau}, \mathbf{z}^{\tau}, \mathbf{v}^{\tau})\}_{\tau=1}^{\infty}$. By telescoping summation of (30) from $\tau = 0$ to T, we have

$$\begin{aligned}
&\mathcal{L}(\mathbf{x}^{T+1}, \mathbf{z}^{T+1}; \mathbf{v}^{T+1}) \\
\leq&\mathcal{L}(\mathbf{x}^{0}, \mathbf{z}^{0}; \mathbf{v}^{0}) - c_1\sum_{\tau=0}^{T}\|\mathbf{x}^{\tau+1} - \mathbf{x}^{\tau}\|_2^2 - c_2\sum_{\tau=0}^{T}\|\mathbf{z}^{\tau+1} - \mathbf{z}^{\tau}\|_2^2 \\
\leq&\mathcal{L}(\mathbf{x}^{0}, \mathbf{z}^{0}; \mathbf{v}^{0}), \forall\, T = 0, 1, \ldots
\end{aligned} \tag{31}$$

which implies that the sequence $\{\mathcal{L}(\mathbf{x}^{\tau}, \mathbf{z}^{\tau}; \mathbf{v}^{\tau})\}_{\tau=0}^{\infty}$ is upper-bounded. On the other hand, we can estimate a lower bound for $\forall\, T = 0, 1, \ldots$

$$\mathcal{L}(\mathbf{x}^{T+1}, \mathbf{z}^{T+1}; \mathbf{v}^{T+1})$$

$$
\begin{aligned}
=&L(\mathbf{z}^{T+1}) + F_\lambda(\mathbf{x}^{T+1}) - L(\mathbf{x}^{T+1}) + \rho\langle \mathbf{v}^{T+1}, \mathbf{x}^{T+1} - \mathbf{z}^{T+1}\rangle + \frac{\rho}{2}\|\mathbf{x}^{T+1} - \mathbf{z}^{T+1}\|_2^2 \\
\geq& F_\lambda(\mathbf{x}^{T+1}) + \langle \rho\mathbf{v}^{T+1} - \nabla L(\mathbf{z}^{T+1}), \mathbf{x}^{T+1} - \mathbf{z}^{T+1}\rangle + \frac{\rho - \mathrm{Lip_L}}{2}\|\mathbf{x}^{T+1} - \mathbf{z}^{T+1}\|_2^2 \\
=& F_\lambda(\mathbf{x}^{T+1}) + \frac{\rho - \mathrm{Lip_L}}{2}\|\mathbf{x}^{T+1} - \mathbf{z}^{T+1}\|_2^2, \qquad (32)
\end{aligned}
$$

where we use (28) and Assumption (A2). It follows from (32) that the sequence $\{F_\lambda(\mathbf{x}^\tau)\}_{\tau=0}^\infty$ is upper-bounded if $\rho > \mathrm{Lip_L}$ and thus $\{F_\lambda(\mathbf{x}^\tau)\}_{\tau=0}^\infty$ is bounded by Assumption (A1). Note that $\mathrm{Lip_L} \leq \frac{\sqrt{\sigma^2 + 8\mathrm{Lip_L}^2} - \sigma}{2}$, a constant stated in Lemma 10. Since $F_\lambda(\cdot)$ is coercive by the coerciveness of P or L, $\{\mathbf{x}^\tau\}_{\tau=1}^\infty$ is bounded, so is $\{\mathbf{z}^\tau\}_{\tau=1}^\infty$ by (32). To show the boundedness of $\{\mathbf{v}^\tau\}_{\tau=1}^\infty$, we consider

$$
\|\mathbf{v}^{T+1} - \mathbf{v}^0\|_2 \leq \frac{\mathrm{Lip_L}}{\rho}\|\mathbf{z}^{T+1} - \mathbf{z}^0\|_2, \ \forall T = 0, 1, \ldots
$$

which can be obtained similarly to (29). This implies that

$$
\|\mathbf{v}^{T+1}\|_2 \leq \|\mathbf{v}^0\| + \frac{\mathrm{Lip_L}}{\rho}(\|\mathbf{z}^{T+1}\|_2 + \|\mathbf{z}^0\|_2), \ \forall T = 0, 1, \ldots \qquad (33)
$$

Together with the boundedness of $\{\mathbf{z}^\tau\}_{\tau=1}^\infty$, we have $\{\mathbf{v}^\tau\}_{\tau=1}^\infty$ bounded. By Bolzano–Weierstrass theorem, the bounded sequence $\{(\mathbf{x}^\tau, \mathbf{z}^\tau, \mathbf{v}^\tau)\}_{\tau=1}^\infty$ has a convergent subsequence, denoted by $(\mathbf{x}^{\tau_j}, \mathbf{z}^{\tau_j}, \mathbf{v}^{\tau_j}) \to (\mathbf{x}^*, \mathbf{z}^*, \mathbf{v}^*)$ as $\tau_j \to \infty$.

With the boundedness, $\mathcal{L}(\mathbf{x}^\tau, \mathbf{z}^\tau; \mathbf{v}^\tau)$ converges due to the monotonic decreasing property shown in Lemma 10. By letting $T \to \infty$ in (31), we have $\sum_{\tau=0}^\infty \|\mathbf{x}^{\tau+1} - \mathbf{x}^\tau\|_2^2$ and $\sum_{\tau=0}^\infty \|\mathbf{z}^{\tau+1} - \mathbf{z}^\tau\|_2^2$ are finite. Therefore, $\mathbf{x}^{\tau+1} - \mathbf{x}^\tau \to 0$ and $\mathbf{z}^{\tau+1} - \mathbf{z}^\tau \to 0$ as $\tau \to \infty$. It further follows from Lemma 9 that $\mathbf{v}^{\tau+1} - \mathbf{v}^\tau \to 0$ as well. As $(\mathbf{x}^{\tau_j}, \mathbf{z}^{\tau_j}, \mathbf{v}^{\tau_j}) \to (\mathbf{x}^*, \mathbf{z}^*, \mathbf{v}^*)$, we have $(\mathbf{x}^{\tau_j+1}, \mathbf{z}^{\tau_j+1}, \mathbf{v}^{\tau_j+1}) \to (\mathbf{x}^*, \mathbf{z}^*, \mathbf{v}^*)$ and $\mathbf{x}^* = \mathbf{z}^*$ due to the $\mathbf{v}$-update.

We next show that $(\mathbf{x}^*, \mathbf{z}^*, \mathbf{v}^*)$ is a stationary solution of (1). By the iterative scheme (16), we have

$$
\begin{aligned}
\mathcal{L}(\mathbf{x}^{\tau_j+1}, \mathbf{z}^{\tau_j}, \mathbf{v}^{\tau_j}) \leq \mathcal{L}(\mathbf{x}, \mathbf{z}^{\tau_j}, \mathbf{v}^{\tau_j}) \quad \forall \mathbf{x} \\
\mathcal{L}(\mathbf{x}^{\tau_j+1}, \mathbf{z}^{\tau_j+1}, \mathbf{v}^{\tau_j}) \leq \mathcal{L}(\mathbf{x}^{\tau_j+1}, \mathbf{z}, \mathbf{v}^{\tau_j}) \quad \forall \mathbf{z}
\end{aligned}
$$

Let $\tau_j \to \infty$, we have $\mathcal{L}(\mathbf{x}^*, \mathbf{z}^*, \mathbf{v}^*) \leq \mathcal{L}(\mathbf{x}, \mathbf{z}^*, \mathbf{v}^*)$, $\forall \mathbf{x}$ and $\mathcal{L}(\mathbf{x}^*, \mathbf{z}^*, \mathbf{v}^*) \leq \mathcal{L}(\mathbf{x}^*, \mathbf{z}, \mathbf{v}^*)$, $\forall \mathbf{z}$, which implies that

$$
\lambda\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, p(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2) + \frac{\rho}{2}\|\mathbf{v}^*\|_2^2 \leq \lambda\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, p(\|\mathbf{x}_{\mathcal{G}_k}\|_2) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}^* + \mathbf{v}^*\|_2^2, \forall \mathbf{x} \qquad (34)
$$

and

$$
L(\mathbf{z}^*) + \frac{\rho}{2}\|\mathbf{v}^*\|_2^2 \leq L(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{x}^* - \mathbf{z} + \mathbf{v}^*\|_2^2, \forall \mathbf{z}. \qquad (35)
$$

Let us fix $\mathbf{x}$ in (34), and let $\mathbf{z} = \mathbf{x}$ in (35). As $\mathbf{x}^* = \mathbf{z}^*$, combining (34) and (35) yields

$$
\begin{aligned}
&L(\mathbf{x}^*) + \lambda\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, p(\|\mathbf{x}^*_{\mathcal{G}_k}\|_2) + \rho\|\mathbf{v}^*\|_2^2 \\
&\quad \leq L(\mathbf{x}) + \lambda\sum_{k=1}^m \sqrt{|\mathcal{G}_k|}\, p(\|\mathbf{x}_{\mathcal{G}_k}\|_2) + \frac{\rho}{2}(\|\mathbf{x} - \mathbf{x}^* + \mathbf{v}^*\|_2^2 + \|\mathbf{x}^* - \mathbf{x} + \mathbf{v}^*\|_2^2), \qquad (36)
\end{aligned}
$$

for any $\mathbf{x}$. Define $\widehat{F}_\lambda(\mathbf{x}) \triangleq F_\lambda(\mathbf{x}) + \rho\|\mathbf{x} - \mathbf{x}^*\|_2^2$. It follows from (36) that

$$
\widehat{F}_\lambda(\mathbf{x}^*) = F_\lambda(\mathbf{x}^*) \leq F_\lambda(\mathbf{x}) + \rho\|\mathbf{x} - \mathbf{x}^*\|_2^2 = \widehat{F}_\lambda(\mathbf{x}), \ \forall \mathbf{x},
$$

implying $\mathbf{x}^*$ is a global minimum for $\widehat{F}_\lambda(\mathbf{x})$ and hence a stationary point. We show F_λ and $\widehat{F}_\lambda$ have the same directional derivative at the point $\mathbf{x} = \mathbf{x}^*$ by the following calculations,

$$\begin{aligned}
\widehat{F}'_\lambda(\mathbf{x}^*; \mathbf{d}) &= \lim_{h \to 0^+} \frac{\widehat{F}_\lambda(\mathbf{x}^* + h\mathbf{d}) - \widehat{F}_\lambda(\mathbf{x}^*)}{h} \\
&= \lim_{h \to 0^+} \frac{F_\lambda(\mathbf{x}^* + h\mathbf{d}) + \rho\|\mathbf{x}^* + h\mathbf{d} - \mathbf{x}^*\|_2^2 - F_\lambda(\mathbf{x}^*) - \rho\|\mathbf{x}^* - \mathbf{x}^*\|_2^2}{h} \\
&= \lim_{h \to 0^+} \frac{F_\lambda(\mathbf{x}^* + h\mathbf{d}) - F_\lambda(\mathbf{x}^*) + \rho h^2\|\mathbf{d}\|_2^2}{h} \\
&= \lim_{h \to 0^+} \frac{F_\lambda(\mathbf{x}^* + h\mathbf{d}) - F_\lambda(\mathbf{x}^*)}{h} = F'_\lambda(\mathbf{x}^*; \mathbf{d}), \quad \forall \mathbf{d}.
\end{aligned}$$

Since $\mathbf{x}^*$ is a stationary point of $\widehat{F}_\lambda(\mathbf{x})$, it is also a stationary point of $F_\lambda(\mathbf{x})$. □

If neither L nor P is coercive, then we need to assume the sequence generated by the ADMM framework is bounded. In other words, the sequence either diverges or has a subsequence convergent to a stationary point; see Theorem 12.

Theorem 12. *(convergence without coerciveness) Let Assumptions (P3), (A1)-(A3) hold. Let ρ satisfy the condition given in Theorem 11. If the sequence $\{(\mathbf{x}^\tau, \mathbf{z}^\tau, \mathbf{v}^\tau)\}_{\tau=1}^\infty$ is bounded, then it has a subsequence convergent to a stationary point of problem (1).*

Proof. If $\{\mathbf{x}^\tau\}_{\tau=1}^\infty$ is bounded, then $\{\mathbf{z}^\tau\}_{\tau=1}^\infty$ and $\{\mathbf{v}^\tau\}_{\tau=1}^\infty$ are bounded due to (32) and (33), respectively. The rest of the proof is along the same lines as Theorem 11, thus is omitted. □

Note that the Poisson regression does not have a global Lipschitz constant owing to $\psi''(\theta_i) = \exp(\theta_i)$, and hence Assumption (A2) does not hold. As a result, Theorems 11-12 are not applicable to Poisson regression. Fortunately, we observe in the empirical studies of Section 5 that ADMM with the Poisson loss function generally converges.

5 Numerical experiments

In the numerical experiments, we examine group regularization methods that are widely used, including group LOG, group MCP, group SCAD, group $\ell_{1/2}$, group transformed ℓ_1, and group LASSO. The definitions of the corresponding univariate regularization functions and the references for their proximal operators are listed in Table 1.

A comparison to LASSO is added as a baseline method for feature selection without the group structure. We consider three loss functions for linear regression, Poisson regression, and logistic regression in Sections 5.1-5.3, respectively. We generate synthetic data for linear regression and Poisson regression, while applying logistic regression on a real dataset that involves prostate cancer gene expression levels.

5.1 Synthetic data for linear regression

We generate 50 triplets of a dataset that consists of 200 features and 200 observations $(A, \mathbf{x}^*, \mathbf{b})$ for linear regression, where $A \in \mathbb{R}^{200\times 200}$ is called the feature matrix, $\mathbf{x}^* \in \mathbb{R}^{200}$ is the ground-truth vector, and $\mathbf{b} \in \mathbb{R}^{200}$ is the response vector. Each row of the feature matrix A, denoted by $A_i \in \mathbb{R}^{200}$, is randomly generated from multivariate Gaussian distribution with zero mean $\mathbf{0} \in \mathbb{R}^{200}$ and covariance matrix $\Sigma \in \mathbb{R}^{200\times 200}$ independently, i.e, $A_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$. We set the variances of all the features to be 1, i.e. $\Sigma_{jj} = 1$, $\forall j$. The off-diagonal elements of the covariances $\Sigma_{jj'}$, $j \neq j'$ are set as one of the following cases:

Case 1. $\Sigma_{jj'} = 0$.
Case 2. $\Sigma_{jj'} = 0.2$ when $j, j' \in \mathcal{G}_k$, $\Sigma_{jj'} = 0$ otherwise.
Case 3. $\Sigma_{jj'} = 0.5$ when $j, j' \in \mathcal{G}_k$, $\Sigma_{jj'} = 0.2$ otherwise.

Name	Definition	Proximal operator
LASSO	$p(t) \triangleq \lvert t\rvert$	[20]
MCP $(\tilde{\lambda} > 0, a > 1)$	$p(t;\tilde{\lambda},a) \triangleq \begin{cases} \tilde{\lambda}\lvert t\rvert - \frac{t^2}{2a}, & \lvert t\rvert \le a\tilde{\lambda} \\ \frac{1}{2}a\tilde{\lambda}^2, & \lvert t\rvert > a\tilde{\lambda}, \end{cases}$	[20]
SCAD $(\tilde{\lambda} > 0, a > 2)$	$p(t;\tilde{\lambda},a) \triangleq \begin{cases} \tilde{\lambda}\lvert t\rvert, & \lvert t\rvert \le \tilde{\lambda} \\ \frac{2a\tilde{\lambda}\lvert t\rvert - t^2 - \tilde{\lambda}^2}{2(a-1)}, & \tilde{\lambda} < \lvert t\rvert \le a\tilde{\lambda} \\ \frac{(a+1)\tilde{\lambda}^2}{2}, & \lvert t\rvert > a\tilde{\lambda} \end{cases}$	[20]
Transformed ℓ_1 $(a > 0)$	$p(t;a) \triangleq \frac{(a+1)\lvert t\rvert}{a+\lvert t\rvert},$	[65]
$\ell_{1/2}$	$p(t) \triangleq \lvert t\rvert^{1/2},$	[61]
LOG $(\epsilon > 0)$	$p(t;\varepsilon) \triangleq \log(\sqrt{t^2+\varepsilon} + \lvert t\rvert)$	Appendix B

Table 1: A list of sparsity-promoting regularization functions and their proximal operators

The three cases for the covariance matrix Σ consider various levels of correlations among the features. Specifically, Case 1 considers features that are completely independent of each other, while Cases 2-3 consider the features positively correlated in part or in whole. Positive correlations are introduced within a group in Case 2. In Case 3, all feature pairs have positive correlations, and the within-group correlations are stronger than the across-group correlations.

The ground-truth $\mathbf{x}^* \in \mathbb{R}^d$ $(d = 200)$ consists of 40 equal-size groups with 5 coefficients in each group that are simultaneously zero or nonzero. We assume without loss of generality that $x_1^*, \ldots, x_s^*$ are nonzero coefficients $(s < d)$, whose indices $1, \ldots, s$ are grouped into the first m' distinct groups $\mathcal{G}_1, \ldots, \mathcal{G}_{m'}$ $(m' < m)$. The indices in $\mathcal{G}_{m'+1}, \ldots, \mathcal{G}_m$ correspond to zero coefficients. We set the number of nonzero groups to be one, three, or five for Cases 1-3. The coefficients in the nonzero group(s) are randomly generated from uniform distribution between -5 and 5 independently, i.e, $\mathbf{x}_j^* \overset{iid}{\sim} U[-5, 5]$. The response vector $\mathbf{b}$ is generated by a linear regression model

$$\mathbf{b} = A\mathbf{x}^* + \mathbf{e}, \tag{37}$$

where the noise $\mathbf{e} \in \mathbb{R}^{200}$ follows Gaussian distribution, $e_i \overset{iid}{\sim} \mathcal{N}(0, \tilde{\sigma}^2)$. Here, $\tilde{\sigma}^2$ is the empirical version of $\mathrm{Var}(A\mathbf{x}^*)$ such that $\tilde{\sigma}^2 = \sum_i (A_i\mathbf{x}^* - \bar{A}\mathbf{x}^*)^2/199$, where $\bar{A} = \sum_i A_i/200$.

Let $\hat{\mathbf{x}} \in \mathbb{R}^d$ be a reconstructed solution from any method with its support $\hat{\mathcal{S}} = \{j : \hat{x}_j \neq 0\}$. The complement of the support is denoted as $\hat{\mathcal{S}}^c = \{j : \hat{x}_j = 0\}$. Let $\hat{\mathcal{M}} = \{k : \hat{x}_j \neq 0, \forall j \in \mathcal{G}_k\}$ and $\hat{\mathcal{N}} = \{k : \hat{x}_j = 0, \forall j \in \mathcal{G}_k\}$ denote the index set of groups in which coefficients being estimated as nonzero and the index set of groups whose coefficients being estimated as zero, respectively. To quantitatively evaluate the performance of each regularization method, we consider the following standard metrics:

1. Relative error of $\hat{\mathbf{x}} \triangleq \dfrac{\|\hat{\mathbf{x}} - \mathbf{x}^*\|_2}{\|\mathbf{x}^*\|_2}$.
2. Precision of $\hat{\mathbf{x}} \triangleq \dfrac{|\hat{\mathcal{S}} \cap \{1, \ldots, s\}|}{|\hat{\mathcal{S}}|}$.
3. Recall of $\hat{\mathbf{x}} \triangleq \dfrac{|\hat{\mathcal{S}} \cap \{1, \ldots, s\}|}{s}$.
4. Element accuracy of $\hat{\mathbf{x}} \triangleq \dfrac{|\hat{\mathcal{S}} \cap \{1, \ldots, s\}| + |\hat{\mathcal{S}}^c \cap \{s+1, \ldots, d\}|}{d}$.

	Fixed parameters	Tuned parameters
Group LOG	$\varepsilon = 0.01$	λ from 10^{-4} to 10
Group MCP	$\lambda = 1, a = 2$	$\tilde{\lambda}$ from 10^{-4} to 10
Group SCAD	$\lambda = 1, a = 3.7$	$\tilde{\lambda}$ from 10^{-4} to 10
Group $\ell_{1/2}$	-	λ from 10^{-4} to 10
Group Transformed ℓ_1	$a = 1$	λ from 10^{-4} to 10
Group LASSO	-	λ from 10^{-4} to 10
LASSO	-	λ from 10^{-4} to 10

Table 2: Hyperparameter settings for different penalty functions

5. Group accuracy of $\hat{\mathbf{x}} \triangleq \dfrac{|\hat{\mathcal{M}} \cap \{1, \ldots, m'\}| + |\hat{\mathcal{N}} \cap \{m'+1, \ldots, m\}|}{m}$.

For tuning the hyperparameter λ, we split the dataset $(A, \mathbf{b})$ into two equal-size datasets: a training dataset $(A_{tr}, \mathbf{b}_{tr}) \in \mathbb{R}^{100\times 200} \times \mathbb{R}^{100}$ and a validation dataset $(A_v, \mathbf{b}_v) \in \mathbb{R}^{100\times 200} \times \mathbb{R}^{100}$. We solve the following optimization problem with the training dataset for different penalty functions $p(\cdot)$:

$$\min_{\mathbf{x}, x_0} \frac{1}{100}\|A_{tr}\mathbf{x} + x_0\mathbf{1} - \mathbf{b}_{tr}\|_2^2 + \lambda \sum_{k=1}^{40} \sqrt{|\mathcal{G}_k|}\, p(\|\mathbf{x}_{\mathcal{G}_k}\|_2). \tag{38}$$

For the linear regression problem (38), there is a closed-form solution for the $\mathbf{z}$-subproblem as detailed in Appendix A. The hyperparameter settings for the different penalty functions are summarized in Table 2. Specifically for group LOG, group $\ell_{1/2}$, group transformed ℓ_1 , group LASSO, and LASSO, we tune the hyperparameter λ with 50 logarithmically spaced values (generated by Matlab function `logspace`) from 10^{-4} to 10. For group MCP and group SCAD, we tune the log-spaced hyperparameter $\tilde{\lambda}$ from 10^{-4} to 10 with $\lambda = 1$ in order to produce the estimates in agreement with [14, 64]. Although we do not have the intercept x_0 in generating data from (37), we include it in the estimation to mimic the reality with no prior information about the intercept value. Given a specific value of hyperparameters, we obtain an estimated vector $\hat{\mathbf{x}}$ and an intercept $\hat{x}_0$ that can be used to compute the mean squared error (MSE) on the validation set, $\text{MSE}_{\text{lin}} \triangleq \|A_v\hat{\mathbf{x}} + \hat{x}_0\mathbf{1} - \mathbf{b}_v\|_2^2/100$. The optimal value of the hyperparameter can be found with the smallest MSE among the preset range of the hyperparameters.

Table 3 summarizes the results of all the eight methods (TL1 stands for transformed ℓ_1) for the datasets with Case 1 covariance. The overall performance of the various models does not depend on the number of nonzero groups. We additionally consider the oracle method that sets the zero coefficients in $\mathbf{x}^*$ to be zero a priori and obtains the solution by least-squares minimization without any regularization. The nonconvex regularizers outperform the convex regularizers such as group LASSO and LASSO. The average relative errors of the nonconvex regularizers are very close to those of the oracle method. In particular, group LOG is successful in recovering the sparsity and group structure and has smaller relative errors than the oracle method on average. Although the sparsity and group structure of the oracle method perfectly aligns with the ground-truth, the group LOG outperforms the oracle solution which is undermined by a large signal-to-noise ratio. Both LASSO and group LASSO have low average precision values but high average recall values, which indicates that both methods tend to incorrectly estimate the zero coefficients at the ground-truth as nonzero. LASSO is not successful in recovering the group structure as the group structure is not incorporated in the model formulation. Note that the union of the group index sets $\hat{\mathcal{M}}$ and $\hat{\mathcal{N}}$ for the group regularizers and the oracle method is the whole group index set $\{1, \ldots, m\}$ while LASSO may have some group indices not in any of the two sets. The estimation with a larger number of nonzero groups is a more challenging task, as indicated by the increase in the relative errors and the decrease in the precision, accuracy, and group accuracy metrics. Similar patterns are obtained in the results with Case 2 and Case 3 covariances. Refer to Tables 6-7 in Appendix C for more details.

5.2 Synthetic data for Poisson regression

Poisson regression is a generalized linear model that takes count data as its response. We generate 50 triples of $(A, \mathbf{x}^*, \mathbf{b})$, where $\mathbf{x}^* \in \mathbb{R}^{200}$ is the ground-truth vector and $\mathbf{b} \in \mathbb{R}^{200}$ is the response vector generated by

Method	Relative error	Precision	Recall	Accuracy	Group accuracy
1 nonzero group					
Group LOG	0.0209(0.0073)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group MCP	0.0220(0.0084)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group SCAD	0.0222(0.0084)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group $\ell_{1/2}$	0.0225(0.0084)	0.9900(0.0707)	1.0000(0.0000)	0.9995(0.0035)	0.9995(0.0035)
Group TL1	0.0237(0.0093)	0.8348(0.3202)	1.0000(0.0000)	0.9705(0.0851)	0.9705(0.0851)
Group LASSO	0.0396(0.0124)	0.2769(0.2526)	1.0000(0.0000)	0.8685(0.0982)	0.8685(0.0982)
LASSO	0.0567(0.0157)	0.3293(0.1588)	0.9800(0.0606)	0.9352(0.0364)	0.7230(0.1360)
Oracle	0.0219(0.0087)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
3 nonzero groups					
Group LOG	0.0420(0.0097)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group MCP	0.0425(0.0103)	0.8956(0.1631)	1.0000(0.0000)	0.9875(0.0216)	0.9875(0.0216)
Group SCAD	0.0437(0.0111)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group $\ell_{1/2}$	0.0436(0.0096)	0.9950(0.0354)	1.0000(0.0000)	0.9995(0.0035)	0.9995(0.0035)
Group TL1	0.0433(0.0098)	0.7975(0.2461)	1.0000(0.0000)	0.9670(0.0486)	0.9670(0.0486)
Group LASSO	0.0782(0.0144)	0.2299(0.1408)	1.0000(0.0000)	0.6785(0.1513)	0.6785(0.1513)
LASSO	0.1099(0.0199)	0.2848(0.0621)	0.9480(0.0607)	0.8042(0.0634)	0.3645(0.1133)
Oracle	0.0424(0.0097)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
5 nonzero groups					
Group LOG	0.0567(0.0103)	0.9967(0.0236)	1.0000(0.0000)	0.9995(0.0035)	0.9995(0.0035)
Group MCP	0.0575(0.0122)	0.9411(0.1199)	1.0000(0.0000)	0.9890(0.0263)	0.9890(0.0263)
Group SCAD	0.0574(0.0121)	0.9943(0.0404)	1.0000(0.0000)	0.9990(0.0071)	0.9990(0.0071)
Group $\ell_{1/2}$	0.0599(0.0108)	0.9843(0.0558)	1.0000(0.0000)	0.9975(0.0091)	0.9975(0.0091)
Group TL1	0.0587(0.0116)	0.8902(0.1838)	1.0000(0.0000)	0.9745(0.0524)	0.9745(0.0524)
Group LASSO	0.1128(0.0208)	0.2225(0.0481)	1.0000(0.0000)	0.5400(0.1132)	0.5400(0.1132)
LASSO	0.1723(0.0382)	0.3225(0.0519)	0.9344(0.0377)	0.7369(0.0598)	0.2525(0.0855)
Oracle	0.0569(0.0113)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)

Table 3: Results for synthetic linear regression datasets with Case 1 covariance matrix. The averages are presented along with their standard deviations in parentheses.

Poisson distribution, i.e.,

$$\mathbf{b}_i \overset{iid}{\sim} \text{Poisson}(A_i\mathbf{x}^*).$$

The feature matrix $A \in \mathbb{R}^{200\times 200}$ is generated in the same manner as in Section 5.1 with Case 1, 2, 3 covariance settings. One, three, and five groups of 5 coefficients in the ground-truth $\mathbf{x}^*$ are nonzero and the rest of the coefficients are zero. The nonzero coefficients are randomly generated from uniform distribution between -0.5 and 0.5, i.e., $\mathbf{x}_j^* \overset{iid}{\sim} U[-0.5, 0.5]$. Same as in Section 5.1, we randomly split the dataset $(A, \mathbf{b})$ into training and validation sets. For the training dataset $(A_{tr}, \mathbf{b}_{tr})$, we consider the following optimization problem with different penalty functions $p(\cdot)$:

$$\min_{\mathbf{x},x_0} \frac{1}{100}\sum_{i=1}^{100}\Big[(b_{tr})_i\Big((A_{tr})_i\mathbf{x}+x_0\Big)-\exp\Big((A_{tr})_i\mathbf{x}+x_0\Big)\Big] + \lambda\sum_{k=1}^{40}\sqrt{|\mathcal{G}_k|}p(\|\mathbf{x}_{\mathcal{G}_k}\|_2).$$

We adopt Newton's method, described in Appendix A, to solve for the $\mathbf{z}$-subproblem. We use 50 different hyperparameter values from 10^{-4} to 10 (λ for group LOG, group $\ell_{1/2}$, group transformed ℓ_1, group LASSO, LASSO, and $\tilde{\lambda}$ for group MCP, group SCAD). Given a specific hyperparameter value, we obtain the estimated coefficients $\hat{\mathbf{x}}$ and the intercept $\hat{x}_0$. For hyperparameter selection, we use mean squared error (MSE) on the validation set $(A_v, \mathbf{b}_v)$ for Poisson regression defined as

$$\text{MSE}_{\text{pois}} \triangleq \|\exp(A_v\hat{\mathbf{x}}+\hat{x}_0\mathbf{1})-\mathbf{b}_v\|_2^2,$$

where the exp of a vector is a componentwise operation. The hyperparameter value giving the smallest MSE is chosen. With the known ground-truth, the evaluation metrics are the same as in linear regression.

Method	Relative error	Precision	Recall	Accuracy	Group accuracy
1 nonzero group					
Group LOG	0.3857(0.2139)	0.7918(0.3120)	0.9800(0.1414)	0.9800(0.0410)	0.9800(0.0410)
Group MCP	0.4108(0.2248)	0.7542(0.3183)	0.9800(0.1414)	0.9815(0.0311)	0.9815(0.0311)
Group SCAD	0.4338(0.2797)	0.9040(0.2344)	0.9200(0.2740)	0.9905(0.0231)	0.9905(0.0231)
Group $\ell_{1/2}$	0.3949(0.2249)	0.9358(0.1931)	0.9600(0.1979)	0.9945(0.0154)	0.9945(0.0154)
Group TL1	0.4726(0.2060)	0.3558(0.2959)	0.9800(0.1414)	0.9005(0.0983)	0.9005(0.0983)
Group LASSO	0.5486(0.1888)	0.1754(0.1927)	0.9800(0.1414)	0.8050(0.1204)	0.8050(0.1204)
LASSO	0.6776(0.1685)	0.2526(0.1845)	0.5960(0.2725)	0.9220(0.0442)	0.7095(0.1618)
Oracle	0.3673(0.2025)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
3 nonzero groups					
Group LOG	0.3849(0.1423)	0.8335(0.1998)	0.9533(0.1168)	0.9755(0.0288)	0.9755(0.0288)
Group MCP	0.4143(0.1591)	0.8049(0.1883)	0.9467(0.1406)	0.9730(0.0266)	0.9730(0.0266)
Group SCAD	0.4051(0.1676)	0.9537(0.1095)	0.9467(0.1234)	0.9915(0.0148)	0.9915(0.0148)
Group $\ell_{1/2}$	0.4143(0.1503)	0.9090(0.1465)	0.9467(0.1406)	0.9865(0.0197)	0.9865(0.0197)
Group TL1	0.4910(0.1406)	0.4262(0.2012)	0.9867(0.0660)	0.8575(0.1043)	0.8575(0.1043)
Group LASSO	0.5668(0.1161)	0.2207(0.0717)	0.9867(0.0660)	0.7046(0.1115)	0.7045(0.1115)
LASSO	0.7057(0.1022)	0.2823(0.0672)	0.6173(0.1318)	0.8430(0.0474)	0.4680(0.1485)
Oracle	0.3618(0.1255)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
5 nonzero groups					
Group LOG	0.3789(0.1248)	0.8296(0.1834)	0.9640(0.0875)	0.9621(0.0423)	0.9620(0.0423)
Group MCP	0.4165(0.1373)	0.8393(0.1215)	0.9600(0.0808)	0.9690(0.0245)	0.9690(0.0245)
Group SCAD	0.4223(0.1533)	0.9445(0.1127)	0.9160(0.1462)	0.9815(0.0271)	0.9815(0.0271)
Group $\ell_{1/2}$	0.4274(0.1439)	0.9166(0.1538)	0.9400(0.1161)	0.9770(0.0357)	0.9770(0.0357)
Group TL1	0.4907(0.1295)	0.4112(0.1577)	0.9800(0.0606)	0.7831(0.1123)	0.7830(0.1124)
Group LASSO	0.5735(0.1085)	0.2554(0.0643)	0.9800(0.0606)	0.6113(0.1220)	0.6110(0.1222)
LASSO	0.7090(0.1130)	0.3238(0.0634)	0.5616(0.1527)	0.7911(0.0470)	0.3690(0.1436)
Oracle	0.3787(0.1001)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)

Table 4: Results for synthetic Poisson regression datasets with Case 1 covariance matrix. The averages are presented along with their standard deviations in parentheses.

Table 4 summarizes the results of the eight approaches with Case 1 covariance setting. The five nonconvex regularizers show great performance, among which group LOG attains the smallest relative error. Group LASSO and LASSO show poor performance in terms of relative errors, precision, and group accuracy. Tables 8-9 in Appendix C present the results for the Case 2 and Case 3 covariance settings, reporting a similar conclusion.

5.3 Prostate cancer gene expression dataset for logistic regression

Prostate cancer is the most commonly diagnosed non–skin cancer and the second leading cause of cancer death among men in the United States [7]. In order to identify prostate cancer risk genes, we analyze gene expression levels of prostate samples collected from 102 study participants [48]. Genetic expression levels of 6033 genes were measured from 52 prostate cancer patients and 50 normal controls. For our model fittings, we choose the top 50 genes with large median absolute deviation in the participants' gene expression levels. The response vector consists of binary values representing `0 = cancer patient` and `1 = normal control`. We generate a feature matrix of dimension 102×150 for a cubic B-spline basis function of the 50 genes, where each row is for one man and each of the three columns corresponds to one gene.

To evaluate the performance of the seven regularization methods for the logistic regression model, we perform 50 random splits of the dataset into a training set $(A_{tr}, \mathbf{b}_{tr}) \in \mathbb{R}^{82\times 150} \times \mathbb{R}^{82}$ and test set $(A_{test}, \mathbf{b}_{test}) \in \mathbb{R}^{20\times 150} \times \mathbb{R}^{20}$. The model coefficients $\hat{\mathbf{x}}$ and $\hat{x}_0$ are estimated from the logit loss function with regularization, i.e.,

$$\min_{\mathbf{x},x_0} \frac{1}{82}\sum_{i=1}^{82}\left[\log\left\{1+\exp\left((A_{tr})_i\mathbf{x}+x_0\right)\right\} - (b_{tr})_i\exp\left((A_{tr})_i\mathbf{x}+x_0\right)\right] + \lambda\sum_{k=1}^{50}\sqrt{|\mathcal{G}_k|}p(\|\mathbf{x}_{\mathcal{G}_k}\|_2).$$

As the Hessian matrix for this problem is nearly singular, we apply the standard gradient descent for solving

the $\mathbf{z}$-subproblem. The oracle method is not available for this experiment since the ground-truth is not known. The performance metrics of each method are given by

1. Prediction error $\triangleq \frac{1}{20}\sum_{i=1}^{20}\left|\mathbf{1}\left((\hat{b}_{test})_i > 0.5\right) - (b_{test})_i\right|$, where

$$(\hat{b}_{test})_i = \frac{1}{1+\exp(-\hat{x}_0 - (A_{test})_i\hat{\mathbf{x}})}$$

   and $\mathbf{1}(\Omega)$ is the indicator function that returns 1 if condition Ω holds and 0 otherwise.

2. Area under a receiver operating characteristic curve (AUC) $\triangleq \int_0^1 TPR(t)dt$, where

$$TPR(t) \triangleq \frac{TP(t)}{TP(t)+FN(t)}$$
$$TP(t) \triangleq \sum_{i=1}^{20}\mathbf{1}\left((\hat{b}_{test})_i > t\right)(b_{test})_i$$
$$FN(t) \triangleq \sum_{i=1}^{20}\mathbf{1}\left((\hat{b}_{test})_i \leq t\right)(b_{test})_i.$$

3. Coefficient selection rate $\triangleq \frac{1}{d}\sum_{j=1}^{d}\mathbf{1}(\hat{x}_j \neq 0)$.

4. Group selection rate $\triangleq \frac{1}{m}\sum_{k=1}^{m}\mathbf{1}(\hat{\mathbf{x}}_{\mathcal{G}_k} \neq \mathbf{0})$, where $\hat{\mathbf{x}}_{\mathcal{G}_k}$ is a coefficient subsequence corresponding to the k-th group.

For all methods, we perform 5-fold cross-validation on the 50 training sets to tune the hyperparameters (λ for group LOG, group $\ell_{1/2}$, group transformed ℓ_1, group LASSO, LASSO, and $\tilde{\lambda}$ for group MCP, group SCAD) by maximizing AUC. The other parameters are fixed as in Table 2, except for $\varepsilon =$ `1e-4` of group LOG. For group LOG, we tune the log-spaced hyperparameter λ from 10^{-4} to 10^{-2}. For group $\ell_{1/2}$, group transformed ℓ_1 and group LASSO, we tune the log-spaced hyperparameter λ from 10^{-3} to 10^{-1}. For group MCP and group SCAD, we tune the log-spaced hyperparameter $\tilde{\lambda}$ from 10^{-3} to 10^{-1} while fixing the hyperparameter $\lambda = 1$. With the selected hyperparamter by cross-validation, the model coefficients $\hat{\mathbf{x}}$ and $\hat{x}_0$ are estimated on the whole training set and their performances are evaluated on the corresponding test set.

Method	Prediction error	AUC	Coefficient selection rate	Group selection rate
Group LOG	0.4240(0.1089)	0.6308(0.1344)	0.5224(0.2456)	0.5224(0.2456)
Group MCP	0.4130(0.1024)	0.6467(0.1251)	0.9984(0.0055)	0.9984(0.0055)
Group SCAD	0.4080(0.0804)	0.6528(0.1112)	0.9944(0.0162)	0.9944(0.0162)
Group $\ell_{1/2}$	0.4310(0.0931)	0.6248(0.1149)	0.4968(0.3302)	0.4968(0.3302)
Group TL1	0.4530(0.0992)	0.5965(0.1238)	0.3580(0.2188)	0.3580(0.2188)
Group LASSO	0.4670(0.1053)	0.5982(0.1086)	0.3264(0.2791)	0.3264(0.2791)
LASSO	0.4410(0.1077)	0.6324(0.1175)	0.1891(0.1535)	0.4172(0.3083)

Table 5: Results for prostate cancer gene dataset. The averages are presented along with their standard deviations in parentheses.

Table 5 exhibits the performance of the seven regularization methods with logistic regression. The nonconvex group penalties except for group transformed ℓ_1 have smaller prediction errors and higher AUCs than group LASSO. As seen in the coefficient and group selection rates, both group MCP and group SCAD result in non-sparse coefficients with marginally higher AUC values. LASSO, on average, has the smallest coefficient selection rate, but it cannot retain the group structure by nature. Taking both prediction accuracy and group sparsity recovery into consideration, group LOG shows satisfactory results compared to the other methods.

6 Conclusion

We generalized the formulation of the non-overlapping group selection problem, which encompasses many existing works by choosing a specific set of loss functions and sparsity-promoting functions. We analyzed the properties of a stationary solution to our proposed model, demonstrating its global optimality under certain conditions and providing a bound of its distance to a reference point which is a proxy of the ground-truth in the view of probability. We applied the ADMM framework with a proximal operator to iteratively minimize the generalized formulation that is commonly nonconvex and nonsmooth. We also proved the subsequence convergence of the algorithm to a stationary point. The global convergence, which requires more conditions on the loss function, will be left for future work. In numerical experiments, we tested our algorithm on synthetic datasets with linear and Poisson regression analysis, showing that nonconvex group regularization methods often outperform the convex approaches with respect to the recovery of the ground-truth. The analysis of prostate cancer gene expression data confirmed that a solution with group sparsity structure is successfully produced by our proposed model, in which nonconvex group regularization methods outperform group LASSO.

Acknowledgements

Ahn and Ke were supported by NSF grant CRII IIS-1948341. Shin was supported in part by NSF grant DMS-2113674 and by POSTECH Basic Science Research Institute Fund, whose Korean NRF grant number is 2021R1A6A1A1004294. Lou was partially supported by NSF grant DMS-1846690.

References

[1] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, vol. 55, US Government printing office, 1964.

[2] M. Ahn, *Consistency bounds and support recovery of d-stationary solutions of sparse sample average approximations*, Journal of Global Optimization, 78 (2019), pp. 397–422.

[3] M. Ahn, J.-S. Pang, and J. Xin, *Difference-of-convex learning: Directional stationarity, optimality, and sparsity*, SIAM Journal on Optimization, 27 (2017), pp. 1637–1665.

[4] F. Bach, R. Jenatton, J. Mairal, G. Obozinski, et al., *Optimization with sparsity-inducing penalties*, Foundations and Trends® in Machine Learning, 4 (2012), pp. 1–106.

[5] A. Beck, *First-Order Methods in Optimization*, Society for Industrial and Applied Mathematics, Oct. 2017.

[6] S. Boyd, N. Parikh, and E. Chu, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, Now Publishers Inc, 2011.

[7] O. W. Brawley, *Trends in prostate cancer in the United States*, Journal of the National Cancer Institute Monographs, 45 (2012), pp. 152–156.

[8] P. Breheny and J. Huang, *Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors*, Statistics and computing, 25 (2015), pp. 173–187.

[9] E. J. Candés, M. B. Wakin, and S. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, Journal of Fourier Analysis and Applications, 14 (2008), pp. 877–905.

[10] R. Chartrand and B. Wohlberg, *A nonconvex ADMM algorithm for group sparsity with sparse groups*, in 2013 IEEE international conference on acoustics, speech and signal processing, pp. 6009–6013.

[11] F. E. Curtis, Y. Dai, and D. P. Robinson, *A subspace acceleration method for minimization involving a group sparsity-inducing regularizer*, arXiv preprint arXiv:2007.14951, (2020).

[12] W. Deng, W. Yin, and Y. Zhang, *Group sparse optimization by alternating direction method*, in Wavelets and Sparsity XV, vol. 8858, International Society for Optics and Photonics, 2013, p. 88580R.

[13] L. Euler, *Of a new method of resolving equations of the fourth degree*, in Elements of Algebra, Springer, 1972, pp. 282–288.

[14] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association, 96 (2001), pp. 1348–1360.

[15] Y. Gu, J. Fan, L. Kong, S. Ma, and H. Zou, *ADMM for high-dimensional sparse penalized quantile regression*, Technometrics, 60 (2018), pp. 319–331.

[16] I. Guyon and A. Elisseeff, *An introduction to variable and feature selection*, Journal of Machine Learning Research, 3 (2003), pp. 1157–1182.

[17] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, vol. 2, Springer, 2009.

[18] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, RC Press Taylor & Francis Group, 2015.

[19] M. Hong, Z.-Q. Luo, and M. Razaviyayn, *Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems*, SIAM Journal on Optimization, 26 (2016), pp. 337–364.

[20] J. Huang, P. Breheny, and S. Ma, *A selective review of group selection in high-dimensional models*, Statistical Science, 27 (2012).

[21] Y. Jiao, B. Jin, and X. Lu, *Group sparse recovery via the $l_0(l_2)$ penalty: Theory and algorithm*, IEEE Transactions on Signal Processing, 65 (2016), pp. 998–1012.

[22] C. Ke, M. Ahn, S. Shin, and Y. Lou, *Iteratively reweighted group lasso based on log-composite regularization*, SIAM Journal on Scientific Computing, 43 (2021), pp. S655–S678.

[23] K. Khamaru and M. J. Wainwright, *Convergence guarantees for a class of non-convex and non-smooth optimization problems*, Journal of Machine Learning Research, 20 (2019), pp. 1–52.

[24] T. Kronvall and A. Jakobsson, *Hyperparameter selection for group-sparse regression: A probabilistic approach*, Signal Processing, 151 (2018), pp. 107–118.

[25] M.-J. Lai, Y. Xu, and W. Yin, *Improved iteratively reweighted least squares for unconstrained smoothed lq minimization*, SIAM Journal on Numerical Analysis, 51 (2013), pp. 927–957.

[26] G. Lanckriet and B. K. Sriperumbudur, *On the convergence of the concave-convex procedure*, Advances in Neural Information Processing Systems, 22 (2009), pp. 1759–1767.

[27] F. Lauer and H. Ohlsson, *Finding sparse solutions of systems of polynomial equations via group-sparsity optimization*, Journal of Global Optimization, 62 (2015), pp. 319–349.

[28] T. Lipp and S. Boyd, *Variations and extension of the convex–concave procedure*, Optimization and Engineering, 17 (2015), pp. 263–287.

[29] P.-L. Loh and M. J. Wainwright, *Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima*, Journal of Machine Learning Research, 16 (2015), pp. 559–616.

[30] Y. Lou and M. Yan, *Fast $\ell_1-\ell_2$ minimization via a proximal operator*, Journal of Scientific Computing, 74 (2018), pp. 767–785.

[31] Y. Lou, P. Yin, Q. He, and J. Xin, *Computing sparse representation in a highly coherent dictionary based on difference of ℓ_1 and ℓ_2*, J. Sci. Comput., 64 (2015), pp. 178–196.

[32] Z. Lu, Z. Zhou, and Z. Sun, *Enhanced proximal DC algorithms with extrapolation for a class of structured nonsmooth DC minimization*, Mathematical Programming, 176 (2018), pp. 369–401.

[33] P. McCullagh and J. A. Nelder, *Generalized linear models*, Routledge, 2019.

[34] L. Meier, S. Van De Geer, and P. Bühlmann, *The group lasso for logistic regression*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70 (2008), pp. 53–71.

[35] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*, Statistical Science, 27 (2012), pp. 538–557.

[36] M. Nikolova, *Local strong homogeneity of a regularized estimator*, SIAM Journal on Applied Mathematics, 61 (2000), pp. 633–658.

[37] L. Pan and X. Chen, *Group sparse optimization for images recovery using capped folded concave functions*, SIAM Journal on Imaging Sciences, 14 (2021), pp. 1–25.

[38] J.-S. Pang, M. Razaviyayn, and A. Alvarado, *Computing b-stationary points of nonsmooth DC programs*, Mathematics of Operations Research, 42 (2017), pp. 95–118.

[39] J.-S. Pang and M. Tao, *Decomposition methods for computing directional stationary solutions of a class of nonsmooth nonconvex optimization problems*, SIAM Journal on Optimization, 28 (2018), pp. 1640–1669.

[40] N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends in Optimization, 1 (2014), pp. 127–239.

[41] B. Peng and L. Wang, *An iterative coordinate descent algorithm for high-dimensional nonconvex penalized quantile regression*, Journal of Computational and Graphical Statistics, 24 (2015), pp. 676–694.

[42] T. Pham Dinh and H. Le Thi, *Convex analysis approach to DC programming: Theory, algorithms and applications*, ACTA Mathematica Vietnamica, 22 (1997), pp. 289–355.

[43] D. N. Phan and H. A. Le Thi, *Group variable selection via $l_{p,0}$ regularization and application to optimal scoring*, Neural Networks, 118 (2019), pp. 220–234.

[44] Y. Rahimi, C. Wang, H. Dong, and Y. Lou, *A scale-invariant approach for sparse signal recovery*, SIAM J. Sci. Comput., 41 (2019), pp. A3649–A3672.

[45] A. Rakotomamonjy, *Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms*, Signal Processing, 91 (2011), pp. 1505–1526.

[46] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*, SIAM Publications, 2009.

[47] X. Shen, L. Chen, Y. Gu, and H.-C. So, *Square-root lasso with nonconvex regularization: An ADMM approach*, IEEE Signal Processing Letters, 23 (2016), pp. 934–938.

[48] D. Singh et al., *Gene expression correlates of clinical prostate cancer behavior*, Cancer Cell, 1 (2002), pp. 203–209.

[49] M. R. Spiegel, *Mathematical Handbook of Formulas and Tables*, McGraw-Hill, 1968.

[50] S. Tian, Y. Yu, and H. Guo, *Variable selection and corporate bankruptcy forecasts*, Journal of Banking & Finance, 52 (2015), pp. 89–100.

[51] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58 (1996), pp. 267–288.

[52] R. Tibshirani, *The lasso method for variable selection in the cox model*, Statistics in Medicine, 16 (1997), pp. 385–395.

[53] W. van Ackooij, S. Demassey, P. Javal, H. Morais, W. de Oliveira, and B. Swaminathan, *A bundle method for nonsmooth DC programming with application to chance-constrained problems*, Computational Optimization and Applications, 78 (2020), pp. 451–490.

[54] C. Wang, M. Yan, Y. Rahimi, and Y. Lou, *Accelerated schemes for the $l1/l2$ minimization*, IEEE Trans. Signal Process., 68 (2020), pp. 2660–2669.

[55] L. Wang, G. Chen, and H. Li, *Group SCAD regression analysis for microarray time course gene expression data*, Bioinformatics, 23 (2007), pp. 1486–1494.

[56] Y. Wang, W. Yin, and J. Zeng, *Global convergence of ADMM in nonconvex nonsmooth optimization*, Journal of Scientific Computing, 78 (2019), pp. 29–63.

[57] Y. Wang and L. Zhu, *Coordinate majorization descent algorithm for nonconvex penalized regression*, Journal of Statistical Computation and Simulation, (2021), pp. 1–15.

[58] F. Wei and H. Zhu, *Group coordinate descent algorithms for nonconvex penalized regression*, Computational Statistics & Data Analysis, 56 (2012), pp. 316–326.

[59] Y. Xie and U. V. Shanbhag, *Tractable ADMM schemes for computing KKT points and local minimizers for ℓ_0-minimization problems*, Computational Optimization and Applications, 78 (2020), pp. 43–85.

[60] J. Xu, E. Chi, and K. Lange, *Generalized linear model regression under distance-to-set penalties*, Advances in Neural Information Processing Systems, 30 (2017), pp. 1385–1395.

[61] Z. Xu, X. Chang, F. Xu, and H. Zhang, *$l_{1/2}$ regularization: A thresholding representation theory and a fast solver*, IEEE Transactions on Neural Networks and Learning Systems, 23 (2012), pp. 1013–1027.

[62] P. Yin, Y. Lou, Q. He, and J. Xin, *Minimization of ℓ_{1-2} for compressed sensing*, SIAM J. Sci. Comput., 37 (2015), pp. A536–A563.

[63] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68 (2006), pp. 49–67.

[64] C.-H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, The Annals of statistics, 38 (2010), pp. 894–942.

[65] S. Zhang and J. Xin, *Minimization of transformed ℓ_1 penalty: Closed form representation and iterative thresholding algorithms*, Communications in Mathematical Sciences, 15 (2017), p. 511 – 537.

[66] S. Zhang and J. Xin, *Minimization of transformed ℓ_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing*, Mathematical Programming, 169 (2018), pp. 307–336.

[67] Y. Zhang, N. Zhang, D. Sun, and K.-C. Toh, *An efficient hessian based algorithm for solving large-scale sparse group lasso problems*, Mathematical Programming, 179 (2020), pp. 223–263.

[68] Y. Zhou, J. Han, X. Yuan, Z. Wei, and R. Hong, *Inverse sparse group lasso model for robust object tracking*, IEEE Transactions on Multimedia, 19 (2017), pp. 1798–1810.

[69] Y. Zhu, *An augmented ADMM algorithm with application to the generalized lasso problem*, Journal of Computational and Graphical Statistics, 26 (2017), pp. 195–204.

A Common approaches to solving the z-subproblem

This section details three approaches regarding how to solve for the z-subproblem in (16).

- A *closed-form solution* can be derived for the least squares loss

$$L^{\text{lin}}(\mathbf{z}) \triangleq \frac{1}{n}\sum_{i=1}^{n}(b_i - A_i\mathbf{z})^2 .$$

  We shall consider an intercept, denoted by x_0, and hence the least squares loss can be expressed as

$$L^{\text{lin}}(\mathbf{z}, x_0) \triangleq \frac{1}{n}\sum_{i=1}^{n}(b_i - A_i\mathbf{z} - x_0)^2 .$$

  The $\mathbf{z}$-update is given by

$$\mathbf{z}^{\tau+1} = \Big(\frac{1}{n}A^TA + \rho I_d\Big)^{-1}\Big(\frac{1}{n}A^T(x_0^\tau\mathbf{1} + \mathbf{b}) + \rho(\mathbf{x}^{\tau+1} + \mathbf{v}^\tau)\Big),$$

  where $\mathbf{1}$ denotes the all-ones vector and I_d is the $d \times d$ identity matrix. The x_0-update is made by

$$x_0^{\tau+1} = \frac{1}{n}\sum_{i=1}^{n}(b_i - A_i\mathbf{z}^{\tau+1}).$$

- The *Newton's method* is often used when any GLM loss has a continuous second-order derivative. It is especially useful when there is no closed-form solution of $\mathbf{z}$, such as logistic regression and Poisson regression. The Newton's method at the s-th inner iteration is given by

$$\begin{aligned}\mathbf{z}_{s+1} &= \mathbf{z}_s - \delta_s\Big\{\nabla^2_{\mathbf{z}_s}L^{\text{glm}}(\mathbf{z}_s) + \rho I_d\Big\}^{-1} \\ &\qquad \Big\{\nabla_{\mathbf{z}_s}L^{\text{glm}}(\mathbf{z}_s) + \rho(\mathbf{z}_s - \mathbf{x}^{\tau+1} - \mathbf{v}^\tau)\Big\}, \\ &= \mathbf{z}_s - \delta_s\Big\{\psi^{''}(A_i\mathbf{z}_s)A_i^TA_i + \rho I_d\Big\}^{-1} \\ &\qquad \Big[\{\psi^{'}(A_i\mathbf{x}) - b_i\}A_i^T + \rho(\mathbf{z}_s - \mathbf{x}^{\tau+1} - \mathbf{v}^\tau)\Big],\end{aligned}$$

  where $\delta_s > 0$ is a step size. We define the logit loss as follows,

$$L^{\text{logit}}(\mathbf{x}, x_0) \triangleq \frac{1}{n}\sum_{i=1}^{n}\left[\log\left\{1 + \exp(A_i\mathbf{x} + x_0)\right\} - b_i(A_i\mathbf{x} + x_0)\right].$$

  Its first and second derivatives with respect to each component of $\mathbf{x}$ can be obtained by

$$\begin{aligned}\frac{\partial L^{\text{logit}}}{\partial x_j} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{a_{ij}\exp(A_i\mathbf{x} + x_0)}{1 + \exp(A_i\mathbf{x} + x_0)} - b_ia_{ij}\right], \\ \frac{\partial^2 L^{\text{logit}}}{\partial x_j\partial x_k} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{a_{ij}a_{ik}\exp(A_i\mathbf{x} + x_0)}{1 + \exp(A_i\mathbf{x} + x_0)} - a_{ij}a_{ik}\left(\frac{\exp(A_i\mathbf{x} + x_0)}{1 + \exp(A_i\mathbf{x} + x_0)}\right)^2\right],\end{aligned}$$

  for $j, k = 1, \ldots, d$. Its derivatives with respect to x_0 are

$$\begin{aligned}\frac{\partial L^{\text{logit}}}{\partial x_0} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\exp(A_i\mathbf{x} + x_0)}{1 + \exp(A_i\mathbf{x} + x_0)} - b_i\right], \\ \frac{\partial^2 L^{\text{logit}}}{\partial x_0^2} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\exp(A_i\mathbf{x} + x_0)}{1 + \exp(A_i\mathbf{x} + x_0)} - \left(\frac{\exp(A_i\mathbf{x} + x_0)}{1 + \exp(A_i\mathbf{x} + x_0)}\right)^2\right],\end{aligned}$$

For the Poisson loss, which is defined by

$$L^{\text{pois}}(\mathbf{x}) \triangleq -\frac{1}{n}\sum_{i=1}^{n}\left\{b_i(A_i\mathbf{x}+x_0)-\exp(A_i\mathbf{x}+x_0)\right\},$$

its first and second derivatives with respect to each component of $\mathbf{x}$ are

$$\frac{\partial L^{\text{pois}}}{\partial x_j} = \frac{1}{n}\sum_{i=1}^{n}\left[a_{ij}\exp(A_i\mathbf{x}+x_0)-b_i a_{ij}\right]$$

$$\frac{\partial^2 L^{\text{pois}}}{\partial x_j \partial x_k} = \frac{1}{n}\sum_{i=1}^{n} a_{ij}a_{ik}\exp(A_i\mathbf{x}+x_0).$$

Its first and second derivatives with respect to x_0 are given by

$$\frac{\partial L^{\text{pois}}}{\partial x_0} = \frac{1}{n}\sum_{i=1}^{n}\left[\exp(A_i\mathbf{x}+x_0)-b_i\right],$$
$$\frac{\partial^2 L^{\text{pois}}}{\partial x_0^2} = \frac{\partial^2 L^{\text{pois}}}{\partial x_j \partial x_k} = \frac{1}{n}\sum_{i=1}^{n}\exp(A_i\mathbf{x}+x_0),$$

- *Gradient descent* is considered when computing the Hessian matrix is infeasible or inefficient. It is useful for analysis of high dimensional datasets or employing loss functions that are not twice differentiable. The gradient descent at the s-th inner iteration is given by

$$\begin{aligned}\mathbf{z}_{s+1} &= \mathbf{z}_s - \delta_s\left[\nabla_{\mathbf{z}_s} L^{\text{glm}}(\mathbf{z}_s)+\rho(\mathbf{z}_s-\mathbf{x}^{\tau+1}-\mathbf{v}^{\tau})\right]\\ &= \mathbf{z}_s - \delta_s\left[\{\psi^{'}(A_i\mathbf{x})-b_i\}A_i^T+\rho(\mathbf{z}_s-\mathbf{x}^{\tau+1}-\mathbf{v}^{\tau})\right]\end{aligned}$$

  where $\delta_s > 0$ is a step size.

B Example: group LOG

Group LOG penalty was recently developed in [22] that can be solved by an iterative reweighted algorithm. The high computational costs due to the double loop of the iterative scheme motivated us to derive the proximal operator of group LOG, followed by ADMM leading to a single-loop algorithm. We derive a closed-form solution of the proximal operator for group LOG under certain conditions and demonstrate the ADMM scheme equipped with this proximal operator significantly reduces the computational time compared to our previous iterative reweighted approach [22].

Let $p_{log}(x)$ denote the group LOG penalty, i.e, $p_{log}(x) = \log(\sqrt{x^2+\varepsilon}+|x|)$. The penalty function satisfies Assumptions (P1)-(P3). We define an objective function $f_y(x)$, $x \in \mathbb{R}$, corresponding to the LOG penalty function, i.e.,

$$\begin{aligned}\mathbf{prox}_{log}(y;\mu) \in \operatorname{argmin} f_y(x) &\triangleq \mu p_{log}(x)+\frac{1}{2}(x-y)^2\\ &= \mu\log(\sqrt{x^2+\varepsilon}+|x|)+\frac{1}{2}(x-y)^2.\end{aligned}$$

We are interested in the stationary points of $f_y(x)$, which can be 0 or any point $x^* \neq 0$ such that $f_y'(x^*)=0$. Since the first and second terms of f_y are symmetric about the vertical axis and $y=x$, a minimizer of f_y must have the same sign as y. The first and second order derivative of $f_y(x)$ in $\mathbb{R}\setminus\{0\}$ are given by

$$f_y'(x) = \frac{\mu\,\operatorname{sign}(x)}{\sqrt{x^2+\varepsilon}}+x-y, \quad x\neq 0,$$

$$f_y''(x) = -\frac{\mu|x|}{(x^2+\varepsilon)^{3/2}} + 1, \quad x \neq 0.$$

Instead of directly solving for $f_y'(x) = 0$ to derive the proximal operator, we simply find real roots of the quartic equation $f_y'(x)g_y(x) = 0$ with

$$g_y(x) \triangleq \frac{\mu\,\mathrm{sign}(x)}{\sqrt{x^2+\varepsilon}} - (x-y).$$

Specifically we first examine the quartic equation,

$$f_y'(x)g_y(x) = x^4 - 2yx^3 + (y^2+\varepsilon)x^2 - 2y\varepsilon x + y^2\varepsilon - \mu^2 = 0, \tag{39}$$

followed by the discussion on which of these roots corresponds to the solution of $f_y'(x) = 0$. According to [13, 49], the quartic equation (39) have the following four roots:

$$\begin{aligned}
x_1 &= \mathrm{sign}(y)\left(\sqrt{t_1} - (\sqrt{t_2} + \sqrt{t_3})\right) + \frac{y}{2},\\
x_2 &= \mathrm{sign}(y)\left(\sqrt{t_1} + (\sqrt{t_2} + \sqrt{t_3})\right) + \frac{y}{2},\\
x_3 &= \mathrm{sign}(y)\left(-\sqrt{t_1} - (\sqrt{t_2} - \sqrt{t_3})\right) + \frac{y}{2},\\
x_4 &= \mathrm{sign}(y)\left(-\sqrt{t_1} + (\sqrt{t_2} - \sqrt{t_3})\right) + \frac{y}{2},
\end{aligned} \tag{40}$$

where

$$\begin{aligned}
t_1 &= \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} - \frac{1}{3}\left(\frac{\varepsilon}{2} - \frac{y^2}{4}\right),\\
t_2 &= \frac{-1+\sqrt{3}i}{2}\sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}\\
&\qquad + \left(\frac{-1+\sqrt{3}i}{2}\right)^2 \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} - \frac{1}{3}\left(\frac{\varepsilon}{2} - \frac{y^2}{4}\right),\\
t_3 &= \frac{-1-\sqrt{3}i}{2}\sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}\\
&\qquad + \left(\frac{-1-\sqrt{3}i}{2}\right)^2 \sqrt[3]{-\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} - \frac{1}{3}\left(\frac{\varepsilon}{2} - \frac{y^2}{4}\right),\\
p &= \left(\frac{\mu^2}{4} + \frac{\varepsilon^2}{16} - \frac{y^2\varepsilon}{8}\right) - \frac{1}{3}\left(\frac{\varepsilon}{2} - \frac{y^2}{4}\right)^2,\\
q &= \frac{2}{27}\left(\frac{\varepsilon}{2} - \frac{y^2}{4}\right)^3 - \frac{1}{3}\left(\frac{\varepsilon}{2} - \frac{y^2}{4}\right)\left(\frac{\mu^2}{4} + \frac{\varepsilon^2}{16} - \frac{y^2\varepsilon}{8}\right) - \frac{y^2\varepsilon^2}{64}.
\end{aligned}$$

In what follows, we show that the first solution (x_1) given in (40) is the stationary solution of $f_y(x)$ when certain conditions hold.

Lemma 13. *Given $\varepsilon > 0$, if $\mu < \frac{3\sqrt{3}}{2}\varepsilon$ and $\frac{\mu}{\sqrt{\varepsilon}} < |y|$, then x_1 given in (40) satisfies $f_y'(x_1) = 0$.*

Proof. By examining the derivative of f_y'', we verify that

$$\inf_{x\neq 0} f_y''(x) = f_y''(\sqrt{\varepsilon/2}) = -\frac{2\mu}{3\sqrt{3}\varepsilon} + 1 > 0, \tag{41}$$

when $\mu < \frac{3\sqrt{3}}{2}\varepsilon$. This implies that $f_y'(a) < f_y'(b)$ for any $a < b < 0$ or $b > a > 0$. From the assumption of $|y| > \dfrac{\mu}{\sqrt{\varepsilon}}$, we discuss two cases: $y > \frac{\mu}{\sqrt{\varepsilon}}$ and $y < -\frac{\mu}{\sqrt{\varepsilon}}$.

In the first case, we have $f_y'(0^-), f_y'(0^+) < 0$. Since $\lim_{x\to\infty} f_y'(x) = \infty$ and f_y' is strictly increasing on $(0,\infty)$ by (41), there exists exactly one root $\bar{x}_1 > 0$ such that $f_y'(\bar{x}_1) = 0$. Similarly, we verify that g_y has exactly one root $\bar{x}_2 > 0$, i.e., $g(\bar{x}_2) = 0$. This implies that the equation (39) has exactly two positive solutions. Furthermore, by examining $f_y'(\bar{x}_1) = \frac{\mu}{\sqrt{\bar{x}_1^2+\varepsilon}} + \bar{x}_1 - y = 0$ and $g(\bar{x}_2) = \frac{\mu}{\sqrt{\bar{x}_2^2+\varepsilon}} - \bar{x}_2 + y = 0$, we deduce $\bar{x}_1 < y < \bar{x}_2$.

Next we identify $\bar{x}_1$ among the candidates shown in (40). Referring to [13, 49], t_1, t_2 and t_3 in (40) are three roots of the cubic equation

$$t^3 + \left(\frac{\varepsilon}{2} - \frac{y^2}{4}\right)t^2 + \left(\frac{\mu^2}{4} + \frac{\varepsilon^2}{16} - \frac{y^2\varepsilon}{8}\right)t - \frac{y^2\varepsilon^2}{64} = 0.$$

Since $-\dfrac{y^2\varepsilon^2}{64} < 0$, there are only three cases for the roots of the above equation: (1) $t_1, t_2, t_3 > 0$; (2) $t_1 > 0$ and $t_2, t_3 < 0$; (3) $t_1 > 0$, t_2 and t_3 are complex conjugate. As only two of the solutions (39) are in $\mathbb{R}$, we must have the third case. Referring to [1], we verify that $\sqrt{t_2} + \sqrt{t_3} > 0$. To see this, let $t_2 = a + bi$ and $t_3 = a - bi$, where $a, b \in \mathbb{R}$ with $b \neq 0$, then we have

$$\begin{aligned}\sqrt{t_2} + \sqrt{t_3} = &\sqrt{\frac{\sqrt{a^2+b^2}+a}{2}} + i\,\mathrm{sign}(b)\sqrt{\frac{\sqrt{a^2+b^2}-a}{2}} \\ &+ \sqrt{\frac{\sqrt{a^2+b^2}+a}{2}} - i\,\mathrm{sign}(b)\sqrt{\frac{\sqrt{a^2+b^2}-a}{2}} \\ = &\sqrt{2(\sqrt{a^2+b^2}+a)} > 0.\end{aligned}$$

Similarly, we can show that $\sqrt{t_2} - \sqrt{t_3} \notin \mathbb{R}$. Since $\bar{x}_1 < \bar{x}_2$, we have

$$\begin{aligned}\bar{x}_1 &= \mathrm{sign}(y)\left(\sqrt{t_1} - (\sqrt{t_2} + \sqrt{t_3})\right) + \frac{y}{2} \text{ with } f_y'(\bar{x}_1) = 0, \\ \bar{x}_2 &= \mathrm{sign}(y)\left(\sqrt{t_1} + (\sqrt{t_2} + \sqrt{t_3})\right) + \frac{y}{2} \text{ with } g_y(\bar{x}_2) = 0,\end{aligned}$$

showing $\bar{x}_1$ is a stationary solution of $f_y(x)$. The proof for the remaining case of $y < -\frac{\mu}{\sqrt{\varepsilon}}$ can be shown similarly.

□

Theorem 14. *If $\mu < \dfrac{3\sqrt{3}}{2}\varepsilon$, then the proximal operator of group LOG is given by*

$$\mathbf{prox}_{log}(y;\mu) = \begin{cases} 0, & \text{if } |y| \le \dfrac{\mu}{\sqrt{\varepsilon}}, \\ \mathrm{sign}(y)\left(\sqrt{t_1} - (\sqrt{t_2} + \sqrt{t_3})\right) + \dfrac{y}{2}, & \text{if } |y| > \dfrac{\mu}{\sqrt{\varepsilon}}. \end{cases}$$

Proof. The case of $|y| > \frac{\mu}{\sqrt{\varepsilon}}$ is shown by Lemma 13. For the other case, the definition of f_y' together with (41) yields $f_y'(x_1) < f_y'(0^-) \le 0 \le f_y(0^+) < f_y'(x_2), \forall x_1 < 0 < x_2$. Hence f_y obtains the minimum at 0. □

Here we summarize the procedure regarding how to numerically compute the proximal operator $\mathbf{prox}_{log}(y;\mu)$, which can be either one point among the solutions in (40) or 0. For any $x \neq 0$, $f_y'(x) = 0$ is equivalent to $y - x = \frac{\mu\,\mathrm{sign}(x)}{\sqrt{\bar{x}_1^2+\varepsilon}}$. This implies that, when $y > 0$, a stationary solution $\bar{x}$ of f_y satisfies $\bar{x} \in (0, y)$. Similarily, when $y < 0$, a stationary solution $\bar{x}$ of f_y satisfies $\bar{x} \in (y, 0)$. With the intervals for the real roots of $f_y'(x) = 0$, we present the following process to compute $\mathbf{prox}_{log}(y;\mu)$:

1. Compute x_1, x_2, x_3, x_4 by (40) and define the set of roots for the quartic equation (39) $Q \triangleq \{x_1, x_2, x_3, x_4\}$;

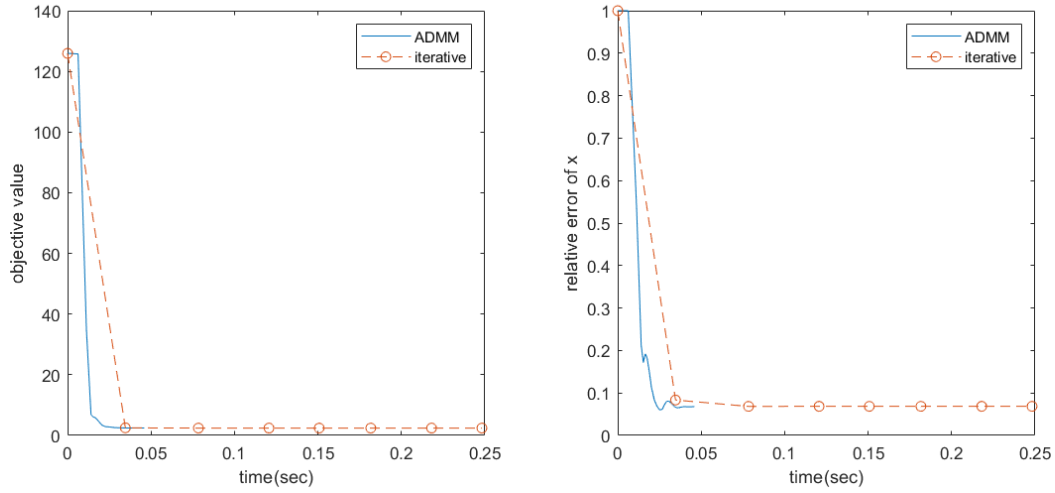

Figure 1: Convergence comparison of the ADMM algorithm for group LOG penalty as opposed to an iterative scheme [22] in terms of objective values (left) and the relative errors to the ground-truth (right) with respect to computation time.

2. Let $P \triangleq \{x \in Q \,|\, x \in \mathbb{R}, |x| \in (0, |y|)\}$, which is the set of real roots for equation $f_y'(x) = 0$.
3. Find $\bar{x}$ such that $f_y(\bar{x}) \leq f_y(x)$ for any $x \in P \cup \{0\}$, which is $\mathbf{prox}_{log}(y; \mu)$.

Next, we compare the ADMM algorithm based on its proximal operator and the iterative algorithm proposed in [22] for the group LOG penalty. We implement both algorithms on an identical computation environment for a synthetic dataset following the linear regression setting in Section 5.1 with case 1 covariance matrix and 3 nonzero groups in ground-truth. Figure 1 plots the decrease in the objective values and the relative errors of iterates with respect to computation time. Although both methods reach the same level of objective value and relative error in the end, the ADMM algorithm is about 5 times faster than the iterative scheme.

C Additional tables

Method	Relative error	Precision	Recall	Accuracy	Group accuracy
1 nonzero group					
Group LOG	0.0245(0.0106)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group MCP	0.0253(0.0113)	0.9473(0.1860)	1.0000(0.0000)	0.9945(0.0228)	0.9945(0.0228)
Group SCAD	0.0250(0.0115)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group $\ell_{1/2}$	0.0272(0.0142)	0.9700(0.1199)	1.0000(0.0000)	0.9985(0.0060)	0.9985(0.0060)
Group TL1	0.0268(0.0109)	0.8310(0.3314)	1.0000(0.0000)	0.9680(0.0753)	0.9680(0.0753)
Group LASSO	0.0506(0.0163)	0.1621(0.1283)	1.0000(0.0000)	0.7855(0.1440)	0.7855(0.1440)
LASSO	0.0729(0.0213)	0.2383(0.1418)	0.9760(0.0657)	0.8927(0.0610)	0.5860(0.1832)
Oracle	0.0253(0.0114)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
3 nonzero groups					
Group LOG	0.0545(0.0136)	0.9900(0.0495)	1.0000(0.0000)	0.9990(0.0049)	0.9990(0.0049)
Group MCP	0.0540(0.0142)	0.8777(0.1904)	1.0000(0.0000)	0.9830(0.0325)	0.9830(0.0325)
Group SCAD	0.0538(0.0140)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group $\ell_{1/2}$	0.0569(0.0143)	0.9520(0.1054)	1.0000(0.0000)	0.9950(0.0113)	0.9950(0.0113)
Group TL1	0.0561(0.0142)	0.7331(0.2945)	1.0000(0.0000)	0.9390(0.0944)	0.9390(0.0944)
Group LASSO	0.1015(0.0260)	0.1709(0.0558)	1.0000(0.0000)	0.5966(0.1352)	0.5965(0.1352)
LASSO	0.1410(0.0356)	0.2781(0.0571)	0.9480(0.0560)	0.7991(0.0627)	0.3355(0.1087)
Oracle	0.0538(0.0138)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
5 nonzero groups					
Group LOG	0.0780(0.0245)	0.9900(0.0400)	1.0000(0.0000)	0.9985(0.0060)	0.9985(0.0060)
Group MCP	0.0776(0.0236)	0.9583(0.1030)	1.0000(0.0000)	0.9925(0.0197)	0.9925(0.0197)
Group SCAD	0.0777(0.0233)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group $\ell_{1/2}$	0.0817(0.0245)	0.9892(0.0576)	1.0000(0.0000)	0.9980(0.0111)	0.9980(0.0111)
Group TL1	0.0791(0.0246)	0.8175(0.2146)	1.0000(0.0000)	0.9565(0.0636)	0.9565(0.0636)
Group LASSO	0.1462(0.0363)	0.2071(0.0393)	1.0000(0.0000)	0.5005(0.1158)	0.5005(0.1158)
LASSO	0.2170(0.0648)	0.3244(0.0447)	0.9200(0.0511)	0.7444(0.0492)	0.2450(0.0825)
Oracle	0.0779(0.0236)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)

Table 6: Results for synthetic linear regression datasets with Case 2 covariance matrix. The averages are presented along with their standard deviations in parentheses.

Method	Relative error	Precision	Recall	Accuracy	Group accuracy
1 nonzero group					
Group LOG	0.0215(0.0073)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group MCP	0.0222(0.0074)	0.8880(0.2486)	1.0000(0.0000)	0.9905(0.0247)	0.9905(0.0247)
Group SCAD	0.0219(0.0069)	0.9867(0.0943)	1.0000(0.0000)	0.9990(0.0071)	0.9990(0.0071)
Group $\ell_{1/2}$	0.0231(0.0078)	0.9900(0.0707)	1.0000(0.0000)	0.9995(0.0035)	0.9995(0.0035)
Group TL1	0.0237(0.0078)	0.8307(0.3294)	1.0000(0.0000)	0.9695(0.0761)	0.9695(0.0761)
Group LASSO	0.0413(0.0098)	0.2977(0.2899)	1.0000(0.0000)	0.8640(0.1104)	0.8640(0.1104)
LASSO	0.0584(0.0125)	0.2874(0.1479)	0.9960(0.0283)	0.9140(0.0587)	0.6690(0.1679)
Oracle	0.0218(0.0070)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
3 nonzero groups					
Group LOG	0.0407(0.0086)	0.9950(0.0354)	1.0000(0.0000)	0.9995(0.0035)	0.9995(0.0035)
Group MCP	0.0422(0.0098)	0.8477(0.2097)	1.0000(0.0000)	0.9790(0.0317)	0.9790(0.0317)
Group SCAD	0.0417(0.0102)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group $\ell_{1/2}$	0.0418(0.0085)	0.9790(0.0858)	1.0000(0.0000)	0.9975(0.0104)	0.9975(0.0104)
Group TL1	0.0417(0.0079)	0.8262(0.2235)	1.0000(0.0000)	0.9745(0.0383)	0.9745(0.0383)
Group LASSO	0.0741(0.0135)	0.2356(0.1004)	1.0000(0.0000)	0.7067(0.1326)	0.7065(0.1325)
LASSO	0.1051(0.0217)	0.3119(0.0760)	0.9533(0.0508)	0.8249(0.0616)	0.4040(0.1391)
Oracle	0.0416(0.0091)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
5 nonzero groups					
Group LOG	0.0563(0.0109)	0.9933(0.0330)	1.0000(0.0000)	0.9990(0.0049)	0.9990(0.0049)
Group MCP	0.0562(0.0110)	0.9611(0.0898)	1.0000(0.0000)	0.9935(0.0158)	0.9935(0.0158)
Group SCAD	0.0561(0.0107)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
Group $\ell_{1/2}$	0.0598(0.0114)	0.9479(0.1044)	1.0000(0.0000)	0.9910(0.0201)	0.9910(0.0201)
Group TL1	0.0576(0.0106)	0.8418(0.2174)	1.0000(0.0000)	0.9595(0.0730)	0.9595(0.0730)
Group LASSO	0.1114(0.0170)	0.2344(0.0602)	1.0000(0.0000)	0.5630(0.1199)	0.5630(0.1199)
LASSO	0.1654(0.0339)	0.3239(0.0640)	0.9304(0.0449)	0.7362(0.0647)	0.2575(0.1044)
Oracle	0.0560(0.0107)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)

Table 7: Results for synthetic linear regression datasets with Case 3 covariance matrix. The averages are presented along with their standard deviations in parentheses.

Method	Relative error	Precision	Recall	Accuracy	Group accuracy
1 nonzero group					
Group LOG	0.4667(0.2512)	0.6767(0.3629)	0.9200(0.2740)	0.9735(0.0366)	0.9735(0.0366)
Group MCP	0.4730(0.2430)	0.6477(0.3650)	0.9200(0.2740)	0.9720(0.0345)	0.9720(0.0345)
Group SCAD	0.5383(0.2897)	0.8370(0.3172)	0.8400(0.3703)	0.9905(0.0159)	0.9905(0.0159)
Group $\ell_{1/2}$	0.4825(0.2487)	0.9621(0.1432)	0.8800(0.3283)	0.9950(0.0113)	0.9950(0.0113)
Group TL1	0.5951(0.2318)	0.3303(0.3088)	0.9200(0.2740)	0.8955(0.0906)	0.8955(0.0906)
Group LASSO	0.6708(0.2127)	0.1442(0.1505)	0.9000(0.3030)	0.8095(0.1320)	0.8095(0.1320)
LASSO	0.7662(0.1874)	0.2150(0.1438)	0.4680(0.2638)	0.9321(0.0450)	0.7475(0.1687)
Oracle	0.4841(0.3168)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
3 nonzero groups					
Group LOG	0.4889(0.1759)	0.7051(0.2399)	0.8867(0.1976)	0.9490(0.0492)	0.9490(0.0492)
Group MCP	0.5131(0.1955)	0.7152(0.2591)	0.8467(0.2448)	0.9510(0.0434)	0.9510(0.0434)
Group SCAD	0.5297(0.2046)	0.8753(0.2081)	0.8000(0.2694)	0.9700(0.0357)	0.9700(0.0357)
Group $\ell_{1/2}$	0.5069(0.1755)	0.9102(0.1638)	0.8667(0.2130)	0.9800(0.0253)	0.9800(0.0253)
Group TL1	0.5480(0.1500)	0.3559(0.1910)	0.9467(0.1234)	0.8206(0.1047)	0.8205(0.1047)
Group LASSO	0.5991(0.1313)	0.1933(0.0668)	0.9533(0.1168)	0.6656(0.1104)	0.6655(0.1105)
LASSO	0.7521(0.1357)	0.2983(0.0895)	0.5347(0.1663)	0.8576(0.0514)	0.5215(0.1674)
Oracle	0.4152(0.1461)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
5 nonzero groups					
Group LOG	0.4923(0.1601)	0.7152(0.2315)	0.9040(0.1628)	0.9220(0.0708)	0.9220(0.0708)
Group MCP	0.5475(0.1610)	0.7324(0.1915)	0.8600(0.1773)	0.9365(0.0480)	0.9365(0.0480)
Group SCAD	0.5632(0.1859)	0.9075(0.1556)	0.8320(0.2035)	0.9650(0.0361)	0.9650(0.0361)
Group $\ell_{1/2}$	0.5186(0.1589)	0.8785(0.1768)	0.8640(0.1871)	0.9640(0.0392)	0.9640(0.0392)
Group TL1	0.5537(0.1288)	0.4053(0.1207)	0.9480(0.1199)	0.7926(0.1025)	0.7925(0.1027)
Group LASSO	0.6187(0.1097)	0.2450(0.0593)	0.9560(0.0929)	0.6006(0.1164)	0.6005(0.1164)
LASSO	0.7725(0.0961)	0.3421(0.0951)	0.5144(0.1582)	0.8014(0.0493)	0.3850(0.1689)
Oracle	0.4606(0.1251)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)

Table 8: Results for synthetic Poisson regression datasets with Case 2 covariance matrix. The averages are presented along with their standard deviations in parentheses.

Method	Relative error	Precision	Recall	Accuracy	Group accuracy
1 nonzero group					
Group LOG	0.4084(0.2395)	0.7899(0.2997)	0.9400(0.2399)	0.9855(0.0243)	0.9855(0.0243)
Group MCP	0.4246(0.2448)	0.7630(0.3104)	0.9400(0.2399)	0.9825(0.0304)	0.9825(0.0304)
Group SCAD	0.4620(0.2887)	0.9167(0.2230)	0.8600(0.3505)	0.9930(0.0134)	0.9930(0.0134)
Group $\ell_{1/2}$	0.4158(0.2543)	0.9420(0.1994)	0.9000(0.3030)	0.9950(0.0124)	0.9950(0.0124)
Group TL1	0.4976(0.2137)	0.4364(0.3366)	0.9600(0.1979)	0.9200(0.0851)	0.9200(0.0851)
Group LASSO	0.5776(0.1868)	0.1873(0.1404)	0.9600(0.1979)	0.8335(0.1144)	0.8335(0.1144)
LASSO	0.7383(0.1654)	0.2282(0.1607)	0.5680(0.2810)	0.9163(0.0571)	0.6855(0.2018)
Oracle	0.3660(0.1685)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
3 nonzero groups					
Group LOG	0.3439(0.1184)	0.8405(0.1978)	0.9533(0.1168)	0.9770(0.0285)	0.9770(0.0285)
Group MCP	0.3437(0.1195)	0.8020(0.2062)	0.9600(0.1094)	0.9715(0.0350)	0.9715(0.0350)
Group SCAD	0.3540(0.1253)	0.9633(0.0925)	0.9333(0.1347)	0.9915(0.0130)	0.9915(0.0130)
Group $\ell_{1/2}$	0.3796(0.1259)	0.9456(0.1356)	0.9333(0.1347)	0.9885(0.0203)	0.9885(0.0203)
Group TL1	0.4459(0.1237)	0.4827(0.2140)	0.9733(0.0913)	0.8836(0.0906)	0.8835(0.0907)
Group LASSO	0.5271(0.1011)	0.2158(0.0709)	0.9800(0.0800)	0.6950(0.1249)	0.6950(0.1249)
LASSO	0.6757(0.0825)	0.2817(0.0631)	0.6240(0.1180)	0.8438(0.0449)	0.4555(0.1275)
Oracle	0.3181(0.0904)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
5 nonzero groups					
Group LOG	0.3715(0.1013)	0.8119(0.1824)	0.9680(0.0741)	0.9585(0.0427)	0.9585(0.0427)
Group MCP	0.4261(0.1446)	0.8217(0.1744)	0.9400(0.1010)	0.9590(0.0390)	0.9590(0.0390)
Group SCAD	0.4169(0.1516)	0.9455(0.1203)	0.9240(0.1271)	0.9810(0.0279)	0.9810(0.0279)
Group $\ell_{1/2}$	0.4031(0.1134)	0.9133(0.1451)	0.9560(0.0837)	0.9790(0.0296)	0.9790(0.0296)
Group TL1	0.4740(0.1150)	0.4271(0.1561)	0.9880(0.0480)	0.7930(0.1191)	0.7930(0.1191)
Group LASSO	0.5476(0.0936)	0.2487(0.0677)	0.9960(0.0283)	0.5895(0.1356)	0.5895(0.1356)
LASSO	0.7113(0.1095)	0.3382(0.0774)	0.6016(0.1364)	0.7929(0.0509)	0.3440(0.1404)
Oracle	0.3735(0.1074)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)

Table 9: Results for synthetic Poisson regression datasets with Case 3 covariance matrix. The averages are presented along with their standard deviations in parentheses.