



# A high-dimensional M-estimator framework for bi-level variable selection

Bin Luo<sup>1</sup> · Xiaoli Gao<sup>2</sup>

Received: 7 November 2020 / Revised: 12 May 2021 / Accepted: 17 August 2021 /  
Published online: 9 September 2021  
© The Institute of Statistical Mathematics, Tokyo 2021

## Abstract

In high-dimensional data analysis, bi-level sparsity is often assumed when covariates function group-wisely and sparsity can appear either at the group level or within certain groups. In such cases, an ideal model should be able to encourage the bi-level variable selection consistently. Bi-level variable selection has become even more challenging when data have heavy-tailed distribution or outliers exist in random errors and covariates. In this paper, we study a framework of high-dimensional M-estimation for bi-level variable selection. This framework encourages bi-level sparsity through a computationally efficient two-stage procedure. In theory, we provide sufficient conditions under which our two-stage penalized M-estimator possesses simultaneous local estimation consistency and the bi-level variable selection consistency if certain non-convex penalty functions are used at the group level. Both our simulation studies and real data analysis demonstrate satisfactory finite sample performance of the proposed estimators under different irregular settings.

**Keywords** Bi-level variable selection · Estimation consistency · High dimensionality · M-estimation · Non-convexity

---

Gao is partially supported by Simons Foundation Grant: SF359337.

---

✉ Bin Luo  
bin.luo2@duke.edu

<sup>1</sup> Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, Durham, NC 27705, United States

<sup>2</sup> Department of Mathematics and Statistics, The University of North Carolina at Greensboro, 116 Petty Building, Greensboro, NC 27402, United States

## 1 Introduction

Covariates often function group-wisely in many applications. For example, in gene expression analysis, genes from the same biological pathways may exhibit similar activities. In high-dimensional linear regression, penalized least squares approaches with penalties incorporating grouping structures have become very popular in recent decades. Yuan and Lin (2006) proposed the group Lasso, as a natural extension of the Lasso (Tibshirani, 1996), to select variables at the group level by applying the Lasso penalty on the  $\ell_2$ -norm of coefficients associated with each group of variables in penalized least squares regression (LS-GLasso). To address the bias and inconsistency of the group Lasso estimator in high-dimensional settings, several methods have been investigated, including the adaptive group Lasso (Wei & Huang, 2010), the  $\ell_2$ -norm MCP (Huang et al., 2012), the  $\ell_2$ -norm SCAD (Guo et al., 2015), among others. However, the above approaches encourage only “all-in” or “all-out” variable selection at the group level. To further encourage the sparsity within certain groups, extensive methods have been proposed to perform bi-level variable selection. See, for example, the group Bridge (Huang et al., 2009), the sparse group Lasso (Friedman et al., 2010; Simon et al., 2013), the concave  $\ell_1$ -norm group penalty (Jiang & Huang, 2014), the composite MCP (Breheny & Huang, 2009), the group exponential lasso (Breheny, 2015), among others. See Huang et al. (2012) for a complete review.

When the data dimensionality grows much faster than the sample size, irregular settings often appear, such as the response and a large number of variables are contaminated or heavy-tailed. It has been shown that the LS-GLasso is estimation consistent when the random errors are sub-Gaussian (Wei & Huang, 2010). However, the quadratic loss in LS-GLasso is non-robust to outliers and the estimator is no longer consistent if the random errors wildly deviate from sub-Gaussian distribution. In addition, the required restricted eigenvalue condition on the design matrix may not hold if the predictors are non-Gaussian.

To tackle the problem of heavy-tailed random errors in high-dimensional settings, a few robust penalized approaches have been recently studied. Lilly (2015) proposed the penalized least absolute deviation (LAD) estimator with the group Lasso penalty to relieve the model’s sensitivity due to the existence of outliers in random errors. This method was also extended to the weighted LAD group Lasso when some predictors are contaminated or heavy-tailed. Wang and Tian (2016) investigated a general penalized M-estimators framework using convex loss functions and concave  $\ell_2$ -norm penalties for the partially linear model with grouped covariates. However, those robust methods can only select variables at the group level and thus do not perform bi-level variable selection. In the example of gene expression study, while the data may be heavy-tailed or contaminated due to the complex data generation procedures, we may be still interested in selecting important genes as well as important groups.

Additionally, the above robust methods require the loss function to be convex. It is well known that the convex loss functions such as Huber loss and LAD loss do not downweight the very large residuals due to their convexity. Shevlyakov

et al. (2008) showed that re-descending M-estimators with non-convex loss function possess certain optimal robustness properties. In fact, there still lacks a systematic study of high-dimensional M-estimators that perform robust bi-level variable selection, allowing both loss and group penalty functions to be non-convex.

In this paper, we consider high-dimensional linear regression with grouped covariates, in irregular settings that the data (random errors and/or covariates) may be contaminated or heavy-tailed. In particular, we propose a novel high-dimensional bi-level variable selection method through a two-stage penalized M-estimator framework: penalized M-estimation with a concave  $\ell_2$ -norm penalty achieving the consistent group selection at the first stage and a post-hard-thresholding operator to achieve the within-group sparsity at the second stage. Our perspective at the first stage is different from Wang and Tian (2016) since we allow the loss function to be non-convex and thus it is more general. In addition, our proposed two-stage framework is able to separate the group selection and the individual variable selection efficiently, since the post-hard-thresholding operator at the second stage nearly poses no additional computational burden. More importantly, our framework includes a wide range of M-estimators with strong robustness if a re-descending loss function is adopted. Furthermore, we extend our framework to a more general setting by relaxing the sub-Gaussian assumption enforced on covariates.

Theoretically, we investigate the statistical properties of our proposed two-stage framework with weak assumptions on both random errors and covariates. We first show that with certain mild conditions on the loss function, a penalized M-estimator at the first stage has the local estimation consistency at the minimax rate enjoyed by the LS-GLasso. We further establish that with an appropriate group concave  $\ell_2$ -norm penalty, the estimator from our first stage has a group-level oracle property. We then show that these nice statistical properties can be carried over directly to the post-hard-thresholding estimators at the second stage and thus we establish its bi-level variable selection consistency. The theoretical results allow the dimensionality of data to grow with the sample size at an almost exponential rate.

Computationally, we propose to implement an efficient algorithm through a two-step optimization procedure. We compare the performance of estimators generated from different types of loss functions (e.g., the Huber loss and Cauchy loss) combined with a concave penalty (e.g., MCP penalty). Our numerical results demonstrate satisfactory finite sample performances of the proposed estimators under different settings. Additionally, it also suggests that a well-behaved two-stage M-estimator can be usually obtained by considering a re-descending loss (e.g., Cauchy loss) with a concave penalty, when the data are heavy-tailed or strongly contaminated.

The remainder of our paper is organized as follows. In Sect. 2, we introduce a basic setup for our two-stage penalized M-estimator framework. In Sect. 3, we present the statistical properties of our proposed bi-level M-estimators under some sufficient conditions. We discuss the implementation of the two-stage M-estimators in Sect. 4. In Sect. 5, we conduct some simulation studies to demonstrate the performance of the proposed estimators under different settings. We also apply the proposed estimators for NCI-60 data analysis and illustrate all results in Sect. 6. Section 7 concludes and summarizes the paper. All technical proofs are relegated to S2 in Supplement.

## 2 The two-stage M-estimator framework

Let us consider a high-dimensional data with  $p$  covariates from  $J$  non-overlapping groups. A linear regression model can be written as:

$$y_i = \sum_{j=1}^J \mathbf{x}_{ij}^T \boldsymbol{\beta}_j^* + \epsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where  $\epsilon_i$ s are independent and identically distributed (i.i.d) random errors,  $\mathbf{x}_{ij}$ s are i.i.d  $d_j$ -dimensional covariate vectors corresponding to the  $j$ th group and  $\boldsymbol{\beta}_j^*$  is the  $d_j$ -dimensional true regression coefficient vector of the  $j$ th group. Then,  $p = \sum_{j=1}^J d_j$ . Let  $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \dots, \mathbf{x}_{iJ}^T)^T$  and  $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \dots, \boldsymbol{\beta}_J^{*T})^T$ . We assume the independence between covariates  $\mathbf{x}_i$  and random errors  $\epsilon_i$  for the sake of simplicity. We also assume the group sparsity condition of the model: There exists  $S \subseteq \{1, \dots, J\}$  such that  $\boldsymbol{\beta}_j^* = \mathbf{0}$  for all  $j \notin S$ . Note that we allow the within-group sparsity on some  $\boldsymbol{\beta}_j^* \neq \mathbf{0}$  and thus there exists bi-level sparsity on the coefficient vector  $\boldsymbol{\beta}^*$ .

*Some More Notations* We use bold symbols to denote matrices or vectors. Let  $\beta_m$  be the  $m$ th element of  $\boldsymbol{\beta} \in \mathbb{R}^p$ . For any  $A \subseteq \{1, 2, \dots, p\}$ , we denote  $\boldsymbol{\beta}_A = (\beta_m, m \in A)^T$  a coefficient sub-vector with indexes in  $A$ . Define  $d_a := \max_{1 \leq j \leq J} d_j$ ,  $d_b := \min_{1 \leq j \leq J} d_j$ ,  $d := \sqrt{\frac{d_a}{d_b}}$ . Let  $I_j \subseteq \{1, 2, \dots, p\}$  denote the index set of coefficients in group  $j$ . Then,  $I_S := \bigcup_{j \in S} I_j$  includes all indexes of coefficients in those important groups. Let  $I_0 = \{m : \beta_m^* \neq 0, 1 \leq m \leq p\}$  and thus  $I_0 \subseteq I_S$ . Define  $\beta_{\min}^{*G} := \min_{j \in S} \|\boldsymbol{\beta}_j^*\|_2$  as the minimum group strength on  $\boldsymbol{\beta}^*$ , where  $\|\cdot\|_2$  is the  $\ell_2$  norm. Define  $\beta_{\min}^{*I} := \min_{m \in I_0} |\beta_m^*|$  as the minimum individual signal strength on  $\boldsymbol{\beta}^*$ . Let  $s = |S|$  and  $k = |I_S|$  be the number of important groups and number of variables among all important groups, respectively. We denote  $u_+ = \max(u, 0)$  for any  $u \in \mathbb{R}$ .

*Our Proposed M-estimator Framework for Bi-level Variable Selection* To perform an efficient bi-level variable selection with robustness for the existence of possible data contamination or heavy-tailed distribution between  $\epsilon_i$  and  $\mathbf{x}_i$ , we propose the following two-stage penalized M-estimator framework:

- *Group Penalization (GP) Stage.* First, we perform penalized M-estimation with a group concave penalty achieving the between-group sparsity:

$$\hat{\boldsymbol{\beta}} \in \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p, \|\boldsymbol{\beta}\|_1 \leq R} \left\{ \mathcal{L}_n(\boldsymbol{\beta}) + \sum_{j=1}^J \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j} \lambda) \right\}.$$

- *Hard-thresholding (HT) Stage.* Then, we apply a post-hard-thresholding operator on  $\hat{\boldsymbol{\beta}}$ :

$$\hat{\boldsymbol{\beta}}^h(\theta) = \hat{\boldsymbol{\beta}} \cdot I(|\hat{\boldsymbol{\beta}}| \geq \theta) \tag{2}$$

where “ $\cdot$ ” and “ $\geq$ ” in (2) are elementary-wise.

Here,  $\mathcal{L}_n$  is an empirical loss function that may produce a robust solution and  $\rho$  is a penalty function, which encourages the group sparsity in the solution. Note that  $\lambda$  and  $\theta$  are tuning parameters controlling the between-group and within-group sparsity, respectively. We include the side condition  $\|\boldsymbol{\beta}\|_1 \leq R$  for  $R \geq \|\boldsymbol{\beta}^*\|_1$  in the Group Penalization Stage in order to guarantee the existence of local/global optima, for the case where the loss or regularizer may be non-convex. In real applications,  $R$  can be a sufficiently large number.

Let  $l : \mathbb{R} \mapsto \mathbb{R}$  denote a residual function, or a loss function, defined on each observation pair  $(\mathbf{x}_i, y_i)$ . Then, the above Group Penalization Stage becomes

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n l(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \sum_{j=1}^J \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j} \lambda) \right\}. \tag{3}$$

With a well chosen  $l$ , the penalized M-estimator from (3) can be robust to heavy-tailed random error  $\epsilon_i$ . Some typical robust loss functions  $l$  include:

- *Huber loss*

$$l(u) = \begin{cases} \frac{u^2}{2} & \text{if } |u| \leq \alpha, \\ \alpha|u| - \frac{\alpha^2}{2} & \text{if } |u| \geq \alpha. \end{cases}$$

- *Tukey’s biweight loss*

$$l(u) = \begin{cases} \frac{\alpha^2}{6} (1 - (1 - \frac{u^2}{\alpha^2})^3) & \text{if } |u| \leq \alpha, \\ \frac{\alpha^2}{6} & \text{if } |u| \geq \alpha. \end{cases}$$

- *Cauchy loss*

$$l(u) = \frac{\alpha^2}{2} \log \left( 1 + \frac{u^2}{\alpha^2} \right).$$

The derivatives of the above three loss functions are bounded, and thus, they can mitigate the effect of larger residuals. In particular, the Tukey’s biweight loss and Cauchy loss produce re-descending M-estimators. From the robust regression literature, we call an M-estimator re-descending if there exists  $u_0 > 0$  such that  $|l'(u)| = 0$  or decrease to 0 smoothly, for all  $|u| \geq u_0$ . In that case, strong robustness is obtained by ignoring the large outliers completely. See more discussions in Müller (2004) and Shevlyakov et al. (2008).

Whereas the robust loss function in (3) takes into account the contamination or heavy-tailed distribution in error  $\epsilon_i$ , a single outlier in  $\mathbf{x}_i$  may still cause the corresponding estimator to perform arbitrarily badly. To downweight large values of  $\mathbf{x}_i$ , we extend the Group Penalization Stage in (3) to

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^p, \|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{w(\mathbf{x}_i)}{v(\mathbf{x}_i)} l((y_i - \mathbf{x}_i^T \boldsymbol{\beta})v(\mathbf{x}_i)) + \sum_{j=1}^J \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j} \lambda) \right\}, \tag{4}$$

where  $w, v$  are weight functions such that  $w, v > 0$ . A few options for choosing those weight functions can be found in Mallows (1975), Hill (1977), Merrill and Schuppe (1971) and Loh (2017).

Since  $\beta_j^* = \mathbf{0}$  for  $j \notin S$ , we need the Group Penalization Stage to generate sparse solutions between groups. In particular, we require the penalty function  $\rho$  in (4) to satisfy amendable properties listed in Assumption 1.

**Assumption 1** (Penalty Function Assumptions)  $\rho : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$  is a scalar function that satisfies the following conditions:

- (i) For any fixed  $t \in \mathbb{R}^+$ , the function  $\lambda \mapsto \rho(t, \lambda)$  is non-decreasing on  $\mathbb{R}^+$ .
- (ii) There exists a scalar function  $g : \mathbb{R}^+ \mapsto \mathbb{R}^+$  such that for any  $r \in [1, \infty)$ ,  $\frac{\rho(t, r\lambda)}{\rho(t, \lambda)} \leq g(r)$  for all  $t, \lambda \in \mathbb{R}^+$ .
- (iii) The function  $t \mapsto \rho(t, \lambda)$  is symmetric around zero and  $\rho(0, \lambda) = 0$ , given any fixed  $\lambda \in \mathbb{R}$ .
- (iv) The function  $t \mapsto \rho(t, \lambda)$  is non-decreasing on  $\mathbb{R}^+$ , given any fixed  $\lambda \in \mathbb{R}$ .
- (v) The function  $t \mapsto \frac{\rho(t, \lambda)}{t}$  is non-increasing on  $\mathbb{R}^+$ , given any fixed  $\lambda \in \mathbb{R}$ .
- (vi) The function  $t \mapsto \rho(t, \lambda)$  is differentiable for  $t \neq 0$ , given any fixed  $\lambda \in \mathbb{R}$ .
- (vii)  $\lim_{t \rightarrow 0^+} \frac{\partial \rho(t, \lambda)}{\partial t} = \lambda$ , given any fixed  $\lambda \in \mathbb{R}$ .
- (viii) There exists  $\mu > 0$  such that the function  $t \mapsto \rho(t, \lambda) + \frac{\mu}{2}t^2$  is convex, given any fixed  $\lambda \in \mathbb{R}$ .
- (ix) There exists  $\delta \in (0, \infty)$  such that  $\frac{\partial \rho(t, \lambda)}{\partial t} = 0$  for all  $t \geq \delta\lambda$ , given any fixed  $\lambda \in \mathbb{R}$ .

Properties (iii–ix) in Assumption 1 are related to the penalty functions studied in Loh (2017) and Loh and Wainwright (2015). Adopting the notation from Loh (2017), we consider  $\rho$  to be  $\mu$ -amenable if  $\rho$  satisfies conditions (i)–(viii). If  $\rho$  also satisfies condition (ix), we say that  $\rho$  is  $(\mu, \delta)$ -amenable. In general, if  $\rho$  is  $\mu$ -amenable,  $q(t, \lambda) := \lambda|t| - \rho(t, \lambda)$  is partially differentiable for  $t \neq 0$ . However, properties (iii) and (vii) imply that  $\lim_{t \rightarrow 0} \frac{\partial q(t, \lambda)}{\partial t} = 0$ . Therefore, we can define  $\frac{\partial q(t, \lambda)}{\partial t} |_{t=0} = 0$  and have  $q(t, \lambda)$  everywhere differentiable with respect to  $t$ . Define the vector version  $q_\lambda(\beta) := \sum_{j=1}^J q(\|\beta_j\|_2, \sqrt{d_j}\lambda)$  accordingly. It is easy to see that there exists  $\mu > 0$  such that  $\frac{\mu}{2}\|\beta\|_2^2 - q_\lambda(\beta)$  is convex. This property is important for both computational implementation and theoretical investigation of the group selection properties.

Some popular choices of amenable penalty functions include Lasso (Tibshirani, 1996), SCAD (Fan & Li, 2001), and MCP (Zhang, 2010). Below are the expressions of Lasso and MCP, which will be considered in our numerical analyses.

- The **Lasso** penalty  $\rho(t, \lambda) = \lambda|t|$  is 0-amenable, but not  $(0, \delta)$ -amenable for any  $\delta < \infty$ .
- The **MCP** penalty takes the form

$$\rho(t, \lambda) = \text{sign}(t)\lambda \int_0^{|t|} \left(1 - \frac{z}{\lambda b}\right)_+ dz,$$

where  $b > 0$  is fixed. The MCP penalty is  $(\mu, \delta)$ -amenable with  $\mu = \frac{1}{b}$  and  $\delta = b$ .

It has been shown that a folded concave penalty, such as the SCAD or MCP, often has better variable selection properties than the convex penalty including the Lasso.

### 3 Statistical properties

In this section, we present our theoretical results for the proposed two-stage penalized M-estimator framework. We begin with statistical properties of the estimator  $\hat{\beta}$  in program (4) generated from the Group Penalization Stage. On the one hand, we show a general non-asymptotic bound of the estimation error and establish the local estimation consistency of  $\hat{\beta}$  at the minimax rate enjoyed by the LS-GLasso, under certain mild conditions. On the other hand, we show that the estimator  $\hat{\beta}$  in fact equals the local oracle solution with the correct group support and thus obtain the group-level oracle properties. Finally, we show that those nice statistical properties of  $\hat{\beta}$  can be carried over to the hard-thresholding stage and thus we establish the bi-level variable selection consistency of  $\hat{\beta}^h$ . All proofs are given in S2 in Supplement.

As introduced in (4), the loss function in the two-stage penalized M-estimator framework takes the following form,

$$\mathcal{L}_n(\beta) = \frac{1}{n} \sum_{i=1}^n \frac{w(\mathbf{x}_i)}{v(\mathbf{x}_i)} l((y_i - \mathbf{x}_i^T \beta)v(\mathbf{x}_i)). \tag{5}$$

To obtain the estimation consistency, we make the following assumptions on the residual function  $l$ .

**Assumption 2** (Loss Function Assumptions)  $l : \mathbb{R} \mapsto \mathbb{R}$  is a scalar function with the existence of the first derivative  $l'$  everywhere and the second derivative  $l''$  almost everywhere. In addition,

- (i) there exists a constant  $0 < k_1 < \infty$  such that  $|l'(u)| \leq k_1$  for all  $u \in \mathbb{R}$ .
- (ii)  $l'$  is Lipschitz such that  $|l'(x) - l'(y)| \leq k_2|x - y|$ , for all  $x, y \in \mathbb{R}$  and some  $0 < k_2 < \infty$ .

Note that Assumption 2(i) requires bounded derivative of the loss function, which can limit the effect of large residuals and thus achieve certain robustness. Assumption 2(ii) indicates that  $|l''(u)| < k_2$  for all  $u \in \mathbb{R}$  where  $l''(u)$  exists. The above assumptions actually cover a wide range of loss functions, including Huber loss, Hampel loss, Tukey’s biweight and Cauchy loss.

We now make some assumptions on both random error  $\epsilon$  and covariate vector  $\mathbf{x}$ .

**Assumption 3** (Error and Covariate Assumptions) For  $w(\mathbf{x})$  and  $v(\mathbf{x})$  given in (5), the random error  $\epsilon$  with  $E[\epsilon] = 0$  and covariate vector  $\mathbf{x}$  with  $E[\mathbf{x}] = \mathbf{0}$  satisfy:

- (i) for any  $\mathbf{v} \in \mathbb{R}^p$ ,  $w(\mathbf{x})\mathbf{x}^T \mathbf{v}$  is sub-Gaussian with parameter at most  $k_0^2 \|\mathbf{v}\|_2^2$ .
- (ii) either (a)  $v(\mathbf{x}) = 1$  and  $E[w(\mathbf{x})\mathbf{x}] = \mathbf{0}$ , or (b)  $E[l'(v(\mathbf{x})\epsilon)|\mathbf{x}] = 0$ .

Note that Assumption 3(i) and (ii)(a) hold when  $\mathbf{x}_i^T \mathbf{v}$  is sub-Gaussian for any  $\mathbf{v} \in \mathbb{R}^p$  and  $w(\mathbf{x}) = 1$ . If covariate  $\mathbf{x}$  is contaminated or heavy-tailed, Assumption 3(i) nonetheless holds with some proper choices of  $w(\mathbf{x})$  (e.g.,  $w(\mathbf{x})\mathbf{x}^T \mathbf{v}$  is bounded for any  $\mathbf{v} \in \mathbb{R}$ ), which potentially relaxes the sub-Gaussian assumption on  $\mathbf{x}$ . Assumption 3(ii)(b) holds when  $l'$  is an odd function and  $\epsilon$  follows a symmetric distribution. Despite the possible mild condition of symmetry, the assumptions above are independent of the distribution of  $\epsilon$ , allowing the additive error  $\epsilon$  to be heavy-tailed or contaminated.

In order to obtain the estimation consistency for  $\hat{\boldsymbol{\beta}}$  in (4), we also require the loss function  $\mathcal{L}_n$  to satisfy the following local restricted strong convexity (RSC) condition. This RSC condition was also investigated in Loh and Wainwright (2015) and Loh (2017).

**Assumption 4** (RSC condition) There exist  $\gamma, \tau > 0$  and a radius  $r > 0$  such that the loss function  $\mathcal{L}_n$  in (5) satisfies

$$\langle \nabla \mathcal{L}_n(\boldsymbol{\beta}_1) - \nabla \mathcal{L}_n(\boldsymbol{\beta}_2), \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle \geq \gamma \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 - \tau \frac{\log p}{n} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_1^2, \tag{6}$$

where  $\boldsymbol{\beta}_j \in \mathbb{R}^p$  such that  $\|\boldsymbol{\beta}_j - \boldsymbol{\beta}^*\|_2 \leq r$  for  $j = 1, 2$ .

Note that the RSC assumption is only imposed on  $\mathcal{L}_n$  inside the ball of radius  $r$  centered at  $\boldsymbol{\beta}^*$ . Thus, the loss function used for robust regression can be wildly nonconvex, while it is away from the origin. The ball of radius  $r$  essentially specifies a local region around  $\boldsymbol{\beta}^*$  in which stationary points of program (4) are well behaved. We call such a region the RSC region.

We present the estimation consistency result concerning estimator  $\hat{\boldsymbol{\beta}}$  in Theorem 1.

**Theorem 1** Suppose the random error and covariates satisfy Assumption 3 and  $\mathcal{L}_n$  in (5) satisfies Assumption 2. Then, we have the following results.

- (i) It holds with probability at least  $1 - 2 \exp(-C_2 \log p)$  that  $\mathcal{L}_n$  satisfies

$$\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \leq C_1 \sqrt{\frac{\log p}{n}}. \tag{7}$$

- (ii) Suppose  $\mathcal{L}_n$  satisfies the RSC condition in Assumption 4 with  $\boldsymbol{\beta}_2 = \boldsymbol{\beta}^*$  and  $\rho$  is  $\mu$ -amenable with  $\frac{3}{4}\mu < \gamma$  in Assumption 1. Let  $\hat{\boldsymbol{\beta}}$  be a local estimator in (4) in the RSC region. Then, for  $n \geq Cr^{-2}d_a s \log p$  and  $\lambda \geq \max\{4\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty, 8\tau R \frac{\log p}{n}\}$ ,  $\hat{\boldsymbol{\beta}}$  exists and satisfies the bounds

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{6\sqrt{d_a} \lambda \sqrt{s}}{4\gamma - 3\mu} \text{ and } \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{6(1 + 3g(d))d_a \lambda s}{4\gamma - 3\mu}.$$

The statistical consistency result of Theorem 1 holds even when the random errors are heavy-tailed, and the regressors lack the sub-Gaussian assumption. Theorem 1(ii) essentially gives general deterministic bounds of the estimation error, provided that the loss function  $\mathcal{L}_n$  satisfies the RSC condition and the penalty function  $\rho$  is  $\mu$ -amenable. In particular, Theorem 1 shows that with high probability one can choose  $\lambda = \mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)$  such that  $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = \mathcal{O}_p\left(\sqrt{\frac{d_a s \log p}{n}}\right)$  and  $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}_p\left(g(d)d_a s \sqrt{\frac{\log p}{n}}\right)$ . Hence, if  $d_a$  is finite, the estimator  $\hat{\boldsymbol{\beta}}$  at the Group Penalization Stage is statistically consistent at the minimax rate enjoyed by the LS-GLasso under the sub-Gaussian assumption.

**Remark 1** Recall that  $\tilde{\boldsymbol{\beta}}$  is a stationary point of the optimization in (4) if

$$\langle \nabla \mathcal{L}_n(\tilde{\boldsymbol{\beta}}) + \nabla \rho_\lambda(\tilde{\boldsymbol{\beta}}), \boldsymbol{\beta} - \tilde{\boldsymbol{\beta}} \rangle \geq 0,$$

for all feasible  $\boldsymbol{\beta}$  in a neighbor of  $\tilde{\boldsymbol{\beta}}$ , where  $\rho_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^J \rho(\|\boldsymbol{\beta}_j\|_2, \sqrt{d_j} \lambda)$ . Note that stationary points include both the interior local maxima and all local and global minima. The proof of Theorem 1 in supplementary file reveals that the estimation consistency result also holds for the stationary points in program (4). Hence, Theorem 1 guarantees that all stationary points within the ball of radius  $r$  centered at  $\boldsymbol{\beta}^*$  have local statistical consistency at the minimax rate enjoyed by the LS-GLasso. To simplify the notation,  $\hat{\boldsymbol{\beta}}$  also denotes the stationary points of program (4).

Next, we establish the group-level oracle properties of estimator  $\hat{\boldsymbol{\beta}}$  in (4). Suppose  $I_S$  is given in advance, we define the group-level local oracle estimator as

$$\hat{\boldsymbol{\beta}}_{I_S}^\circ := \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{I_S} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq r} \{\mathcal{L}_n(\boldsymbol{\beta})\}. \tag{8}$$

Let  $\hat{\boldsymbol{\beta}}^\circ := (\hat{\boldsymbol{\beta}}_{I_S}^\circ, \mathbf{0}_{I_S^c})$ . The next theorem shows that when the penalty  $\rho$  is  $(\mu, \delta)$ -amenable and conditions in Theorem 1 are satisfied, the stationary point from (4) within the local neighborhood of  $\boldsymbol{\beta}^*$  is actually unique and agrees with the group oracle estimator in (8).

**Theorem 2** Suppose the penalty  $\rho$  is  $(\mu, \delta)$ -amenable and conditions in Theorem 1 hold. Suppose in addition that  $v(\mathbf{x})x_j$  is sub-Gaussian for all  $j = 1, \dots, p$ ,  $\|\boldsymbol{\beta}^*\|_1 \leq \frac{R}{2}$  for some  $R > \frac{12(1+3g(d))d_a s}{4\gamma-3\mu}$  and  $\boldsymbol{\beta}_{\min}^{*G} \geq C_8 \sqrt{\frac{d_a \log p}{n}}$ . Let  $\hat{\boldsymbol{\beta}}$  be a stationary point of program (4) in the RSC region. Then, for  $n \geq C_0 k \log p$ ,  $k^2 \log k = \mathcal{O}(\log p)$  and  $\lambda = C_6 \sqrt{\frac{\log p}{n}}$ ,  $\hat{\boldsymbol{\beta}}$  satisfies  $\operatorname{supp}(\hat{\boldsymbol{\beta}}) \subseteq I_S$  and  $\hat{\boldsymbol{\beta}}_{I_S} = \hat{\boldsymbol{\beta}}_{I_S}^\circ$  with probability at least  $1 - C_7 \exp(-C_4 \log p/k^2)$ .

Theorem 2 guarantees that the Group Penalization Stage in our proposed framework can recover the true group support with high probability, when the condition of minimum group signal strength is satisfied. Two most common  $(\mu, \delta)$ -amenable penalties are SCAD and MCP, as introduced in Sect. 2.

It has been shown that the GP Stage can select important covariates groups and provides consistent estimation for parameter  $\beta^*$ . We are now ready to establish statistical properties of  $\hat{\beta}^h$  after the HT stage in our proposed framework. We reveal in the following theorem that when the condition of minimum individual signal strength is satisfied, the estimates of the zero elements and the non-zero elements of  $\beta^*$  after the GP Stage can then be well separated. Hence, there exist some thresholds that are able to filter out those non-important covariates within the selected important groups, and thus, the HT stage can perform bi-level variable selection consistently.

**Theorem 3** *Suppose conditions of Theorem 2 hold and in addition that  $\beta_{\min}^{*I} \geq C_5 \sqrt{\frac{\log p}{kn}} + \theta$  and  $\theta > C_5 \sqrt{\frac{\log p}{kn}}$ . With probability at least  $1 - C_7 \exp(-C_4 \log p/k^2)$ , the hard-thresholding estimator  $\hat{\beta}^h(\theta)$  given in (2) satisfies  $\hat{\beta}^h = (\hat{\beta}_{I_0}^O, \mathbf{0}_{I_0^c})$  and  $\|\hat{\beta}^h - \beta^*\|_2 \leq C_5 \sqrt{\frac{\log p}{kn}}$ .*

Theorem 3 guarantees that the estimator  $\hat{\beta}^h$  in our proposed two-stage framework possesses estimation consistency and bi-level variable selection consistency, when conditions of Theorem 2 hold and the condition of minimum individual signal strength is satisfied. Note that such signal strength condition is fairly mild and the bound can decrease arbitrarily closed to 0 with the growth of sample size  $n$ .

### 4 Implementation

We discuss the implementation of the proposed two-stage M-estimator framework in this section, including finding a stationary point in program (4) for a fixed  $\lambda$  and the tuning parameters selection for  $\lambda$  and  $\theta$ .

Note that the optimization in (4) may not be a convex optimization problem since we allow both loss function  $\mathcal{L}_n$  and  $\rho$  to be non-convex. To obtain the corresponding stationary point, we use composite gradient descend algorithm (Nesterov, 2013). Recall  $q_\lambda(\beta) = \sum_{j=1}^J \sqrt{d_j} \lambda \|\beta_j\|_2 - \sum_{j=1}^J \rho(\|\beta_j\|_2, \sqrt{d_j} \lambda)$  and let  $\bar{L}_n(\beta) = \mathcal{L}_n(\beta) - q_\lambda(\beta)$ . We can rewrite the program as

$$\hat{\beta} \in \underset{\|\beta\|_1 \leq R}{\operatorname{argmin}} \left\{ \bar{L}_n(\beta) + \sum_{j=1}^J \sqrt{d_j} \lambda \|\beta_j\|_2 \right\}.$$

Then, the composition gradient iteration is given by

$$\beta^{t+1} \in \underset{\|\beta\|_1 \leq R}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\beta - (\beta^t - \frac{\nabla \bar{L}_n(\beta^t)}{\eta})\|_2^2 + \sum_{j=1}^J \lambda \eta \sqrt{d_j} \|\beta_j\|_2 \right\}, \tag{9}$$

where  $\eta > 0$  is the step size for the update and can be determined by the backtracking line search method described in Nesterov (2013). A simple calculation shows that the iteration in (9) takes the form

$$\beta_j^{t+1} = S_{\lambda\eta\sqrt{d_j}}\left(\left(\beta^t - \eta\nabla\bar{L}_n(\beta^t)\right)_j\right)$$

for  $j = 1, \dots, J$ , where  $S_{\sqrt{d_j}\lambda\eta}(\cdot)$  is the group soft-thresholding operator defined as

$$S_\delta(\mathbf{z}) := \left(1 - \frac{\delta}{\|\mathbf{z}\|_2}\right)_+ \mathbf{z}.$$

We further adopt the following two-step procedure discussed in Loh (2017) to guarantee the convergence to a stationary point of the non-convex optimization problem in (4).

*Step 1:* Run the composite gradient descent using a Huber loss function with convex group Lasso penalty to get an initial estimator.

*Step 2:* Run the composite gradient descent on program (4) using the initial estimator from Step 1.

As to the tuning parameters selection, the optimal values of tuning parameters  $\lambda$  and  $\theta$  are chosen from a two-dimensional grid search using the cross-validation. In particular, the searching grid is formed by partitioning a rectangle uniformly in the scale of  $(\theta, \log(\lambda))$ . Motivated by conditions of Theorem 1 and Theorem 3, the range of the rectangle can be chosen as  $C_{11}\sqrt{\frac{\log p}{n}} \leq \lambda \leq C_{12}\sqrt{\frac{\log p}{n}}$  and  $C_{21}\sqrt{\frac{\log p}{kn}} < \theta \leq C_{22}$ . The optimal values are then found by the combination that minimizes the cross-validated trimmed mean squared prediction error.

**Remark 2** When the data are contaminated or heavy-tailed, the conventional mean squared prediction error for the cross-validation is not resistant to outliers in validation sets and may provide a biased selection of tuning parameters. In our simulation studies, we found that a robust measurement of prediction errors for the cross-validation, such as the trimmed mean squared error, or the error computed using the Huber loss or Cauchy loss, is robust to outliers in the validation set and yields better results in estimation and group/variable selection. For all numerical studies in this paper, we adopt the trimmed mean squared prediction error for the sake of simplicity. As a comparison, we also report some simulation results using the mean squared prediction error for the cross-validation in supplementary file.

### 5 Simulation studies

In this section, we assess the performance of our two-stage M-estimator framework by considering different types of loss functions and penalty functions through various models. The data are generated from the following model:

$$y_i = \mathbf{x}_i^T \beta^* + \epsilon_i, \quad 1 \leq i \leq n.$$

The covariates vector  $\mathbf{x}_i$ s are generated from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance  $\Sigma$  independently. For covariance  $\Sigma = (\sigma_{ij})_{p \times p}$ , we choose

$$\sigma_{ij} = \begin{cases} 1 & \text{if } i = j, \\ (-1)^{i+j}a & \text{if } i \neq j \text{ and } i, j \text{ are in the same group,} \\ (-1)^{i+j}ab & \text{if } i \neq j \text{ and } i, j \text{ are in different groups,} \end{cases}$$

where  $a = 0.8$  or  $0.5$  and  $b = 0.8$  or  $0.5$ . Let  $\beta^* = \phi \cdot |\beta^*|$ , where  $\phi$  is a  $p$ -dimensional vector with the  $j$ th element being  $(-1)^{j+1}$ .

**Example 1 (Group-level Sparsity)** The number of observations  $n = 100$  and the number of variables  $p = 500$  with  $J = 100$  unequal-size groups. We choose  $a = 0.8$  and  $b = 0.5$ . The model includes only between-group sparsity with five relevant groups,  $|\beta_1^*| = |\beta_2^*| = \underbrace{(3, \dots, 3)}_4^T = 3_4^T$ ,  $|\beta_3^*| = |\beta_4^*| = 2_6$ ,  $|\beta_5^*| = 1.5_5$ ,  $\beta_6^* = \dots = \beta_{100}^* = \mathbf{0}_5$ . We

generate random error  $\epsilon_i$  from the following three scenarios: (a)  $N(0, 1)$ , (b)  $t_1$ , (c) Mix Cauchy (70% are from  $N(0, 1)$  and 30% are from standard Cauchy).

We consider bi-level penalized M-estimators with different types of loss functions (the  $\ell_2$  loss, Huber loss, Cauchy loss) and two types of penalty functions (the Lasso and MCP penalties). In particular, we evaluate the performance of non-group estimators, one-stage estimators and two-stage estimators. Without causing any confusion, let  $\hat{\beta}$  be any estimator of  $\beta^*$ . Its performances on both parameter estimation and group/variable selection were evaluated by the following eight measurements:

- (1)  $\ell_2$  error, which is defined as  $\|\hat{\beta} - \beta^*\|_2$ .
- (2)  $\ell_1$  error, which is defined as  $\|\hat{\beta} - \beta^*\|_1$ .
- (3) Model size (MS), the average number of selected covariates.
- (4) Group size (GS), the average number of selected groups.
- (5) False positives rate for individual variable selection (FPR), the percent of selected covariates which are actually unimportant variables.
- (6) False negatives rate for individual variable selection (FNR), the percent of non-selected covariates which are actually important variables.
- (7) False positives rate for group variable selection (GFPR), the percent of selected groups which are actually unimportant groups.
- (8) False negatives rate for group variable selection (GFNR), the percent of non-selected groups which are actually important groups.

Note that  $FPR = \frac{|\hat{I} \cap I_0^c|}{|\hat{I}^c|} \times 100\%$ ,  $FNR = \frac{|\hat{I}^c \cap I_0|}{|I_0|} \times 100\%$ ,  $GFPR = \frac{|\hat{S} \cap S^c|}{|\hat{S}|} \times 100\%$  and  $GFNR = \frac{|\hat{S}^c \cap S|}{|S|} \times 100\%$ , where  $\hat{I} = \{m : \hat{\beta}_m \neq 0, 1 \leq m \leq p\}$ ,  $I_0 = \{m : \beta_m^* \neq 0, 1 \leq m \leq p\}$ ,  $\hat{S} = \{j : \hat{\beta}_j \neq \mathbf{0}, 1 \leq j \leq J\}$  and  $S = \{j : \beta_j^* \neq \mathbf{0}, 1 \leq j \leq J\}$ .

The model considered in Example 1 contains only the between-group sparsity. We also assess the performance of the two-stage M-estimator framework under models with bi-level sparsity in the following example.

**Example 2 (Bi-level Sparsity)** The number of observations  $n = 100$ , and we generate the random error  $\epsilon$  following the same three scenarios described in Example 1.

- (i) The number of variables  $p = 500$  with  $J = 100$  unequal-size groups. We choose  $a = 0.8$  and  $b = 0.5$ . The model includes within-group sparsity among six relevant groups,  $|\beta_1^*| = (1.5, 2, 0, 2.5)^T$ ,  $|\beta_2^*| = (3, 2, 0, 0, 2)^T$ ,  $|\beta_3^*| = (1.5, 0, 2.5, 3, 0, 0)^T$ ,  $|\beta_4^*| = (2, 1.5, 0, \dots, 0)^T$ ,  $|\beta_5^*| = (2.5, 0, 0, 0)^T$ ,  $|\beta_6^*| = (3, 2.5, 2.5, 2, 1.5)^T$ ,  $\beta_7^* = \dots = \beta_{100}^* = \underbrace{(0, \dots, 0)^T}_5$ .
- (ii) Similar to (i) except that we choose  $a = 0.5$  and  $b = 0.8$ .
- (iii) The number of variables  $p = 1000$  with  $J = 100$  unequal-size groups. We choose  $a = 0.8$  and  $b = 0.5$ . The model includes within-group sparsity in among four relevant groups,  $|\beta_1^*| = (3, 2, 0, 0, 0)^T$ ,  $|\beta_2^*| = (1.5, 2, 2.5, 2.5, 3, 0, \dots, 0)^T$ ,  $|\beta_3^*| = (1.5, 0, 2.5, 3, 0, 3, 2, 1.5, 0, \dots, 0)^T$ ,  $|\beta_4^*| = (3, 3, 2.5, 2.5, 2, 2, 1.5, 1.5, 1.5, 1.5)^T$ ,  $\beta_5^* = \dots = \beta_{100}^* = \underbrace{(0, \dots, 0)^T}_{10}$ .

Finally, we design a simulation setting to evaluate the performance of the two-stage M-estimator framework when covariates are contaminated or not sub-Gaussian.

**Example 3 (Contamination on  $\mathbf{x}$ )** All the settings are similar to Example 2(i), except that we let  $n = 120$  and covariates be partially contaminated after the data generation. In particular, 20% of the observations in  $\mathbf{x}$  are replaced by data generated from  $\chi^2(10)$  first and then recentered to have mean zero.

We ran 100 simulations for each scenario described in Examples 1–3. While fixing  $v(\mathbf{x}) \equiv w(\mathbf{x}) \equiv 1$  for Examples 1 and 2, we consider the general two-stage M-estimator framework with  $v(\mathbf{x}) \equiv 1$  and  $w(\mathbf{x}) = \min \left\{ 1, \frac{4}{\|\mathbf{x}\|_\infty} \right\}$  in Example 3. As introduced in Sect. 4, we choose two tuning parameters  $\lambda$  and  $\theta$  optimally with 10-fold cross-validation, with  $\lambda$  ranging in  $(0.01 \sqrt{\frac{\log p}{n}}, 10 \sqrt{\frac{\log p}{n}})$  and  $\theta$  ranging in  $(0.01 \sqrt{\frac{\log p}{kn}}, 0.5)$ . The results from Examples 1 to 3 are reported in Tables 1, 2 and 3, respectively. Note that we consider the one-stage estimators with the Lasso penalty as the GLasso-type estimators. For the MCP penalty, we call the corresponding non-group estimators, one-stage estimators and two-stage estimators the MCP-type, GMCP-type and GMCP-HT-type estimators, respectively.

**Table 1** Simulation results under the model with only between-group sparsity in Example 1

		Group Lasso			Group MCP		
		LS	Huber	Cauchy	LS	Huber	Cauchy
N(0,1)	$\ell_2$ error	1.27	1.29	1.39	0.92	0.93	0.95
	$\ell_1$ error	6.3	6.59	6.93	3.75	3.77	3.85
	MS	55.9	66.21	66.01	31.43	33.23	33.09
	GS	11.18	13.24	13.2	6.29	6.65	6.62
	FPR	6.51	8.69	8.73	1.35	1.73	1.7
	FNR	0	0.36	1.76	0	0	0
	GFPR	6.51	8.69	8.73	1.36	1.74	1.71
	GFNR	0	0.4	1.8	0	0	0
$t_1$	$\ell_2$ error	13.77	2.02	1.82	24.96	2.72	2.46
	$\ell_1$ error	166.82	11.04	9.59	243.53	10.88	10.22
	MS	114.89	71.75	70.12	65.64	27.9	29.15
	GS	23	14.35	14.03	13.16	5.58	5.83
	FPR	19.26	9.94	9.59	9.11	0.61	0.87
	FNR	6.32	1.8	1.68	10.6	0	0
	GFPR	19.26	9.94	9.59	9.12	0.61	0.87
	GFNR	6	1.8	1.6	10	0	0
Mix Cauchy	$\ell_2$ error	12.84	1.42	1.36	16.92	1.46	1.36
	$\ell_1$ error	178.11	7.44	6.92	225.05	5.82	5.48
	MS	94.6	72.11	71.25	51.99	27.2	29.45
	GS	18.92	14.42	14.26	10.4	5.44	5.89
	FPR	14.75	9.94	9.79	5.84	0.46	0.94
	FNR	1.8	0.36	1	3.04	0	0
	GFPR	14.75	9.94	9.8	5.84	0.46	0.94
	GFNR	1.8	0.4	1	3	0	0

The mean  $\ell_2$  error,  $\ell_1$  error, MS, GS, FPR (%), FNR(%), GFPR (%) and GFNR (%) out of 100 iterations are displayed

We mainly evaluate the performance of one-stage estimators in Example 1 since there only exists the between-group sparsity. Table 1 shows that with the same loss function, while the GMCP-type estimators perform comparably to the GLasso-type estimators in the estimation, the former has better group/variable selection accuracy than the latter. This is consistent with the group oracle property stated in Theorem 2. As expected, for the estimators with the same penalty, while they behave similarly in the light-tailed setting ( $N(0, 1)$ ), estimators using Huber loss and Cauchy loss largely outperform the least squares estimator for the heavy-tailed settings ( $t_1$  and Mix Cauchy).

We compare the results of non-group estimators, one-stage estimators and two-stage estimators for Example 2. Note that here we only consider the MCP penalty since it has been shown to perform better than the Lasso penalty. We also omit the results of the least squares estimators since they are not robust to the

**Table 2** Simulation results under the model with bi-level sparsity in Example 2(i)

		MCP		GMCP		GMCP-HT	
		Huber	Cauchy	Huber	Cauchy	Huber	Cauchy
N(0,1)	$\ell_2$ error	6.98	7.45	1.67	1.66	1.57	1.6
	$\ell_1$ error	29.3	31.52	7.5	7.49	6.79	6.9
	MS	28.77	27.56	53.89	53.55	30.49	30.54
	GS	18.75	18	10.79	10.71	8.27	8.21
	FPR	3.54	3.37	7.64	7.57	2.79	2.8
	FNR	31.41	33.53	0	0	0	0
	GFPR	13.66	12.8	5.1	5.01	2.41	2.35
	GFNR	1.5	0.5	0	0	0	0
$t_1$	$\ell_2$ error	11.31	11.5	4.37	3.75	4.34	3.68
	$\ell_1$ error	46.5	47.37	19.45	16.68	19.28	16.08
	MS	12.61	9.85	52.57	47.53	34.7	31.44
	GS	9.99	8.51	10.51	9.5	8.28	7.74
	FPR	1.28	0.83	7.37	6.33	3.72	3.03
	FNR	62.29	65.53	0.18	0.18	1.59	1.12
	GFPR	4.94	3.4	4.83	3.76	2.51	1.9
	GFNR	10.83	11.5	0.5	0.5	1.33	0.83
Mix Cauchy	$\ell_2$ error	8.9	8.91	2.47	2.11	2.39	2.03
	$\ell_1$ error	37.94	38.24	11.14	9.42	10.29	8.62
	MS	18.87	18.91	47.4	48.94	29.24	29.06
	GS	13.84	13.97	9.48	9.79	7.24	7.42
	FPR	2.03	2.08	6.29	6.61	2.54	2.5
	FNR	46.65	48	0	0	0.06	0
	GFPR	8.48	8.59	3.7	4.03	1.32	1.51
	GFNR	2.17	1.67	0	0	0	0

The mean  $\ell_2$  error,  $\ell_1$  error, MS, GS, FPR (%), FNR(%), GFPR (%) and GFNR (%) out of 100 iterations are displayed

heavy-tailed settings. For Example 2(i), Table 2 shows that the GMCP-type estimators outperform the MCP-type estimators in all measurements, as the former incorporates the grouping structure in  $\mathbf{x}$ . By comparing the results of GMCP-type estimators and GMCP-HT-type estimators, we see that the extra hard-thresholding step in the two-stage estimators can effectively improve the estimation and group/variable selection performance. In addition, estimators using the Cauchy loss further outperform the one with Huber Loss for the heavy-tailed settings, showing that the re-descending estimators are more robust to outliers and more efficient for irregular settings. We observe similar patterns in the results of Example 2(ii)–(iii), and thus, we omit those results in this paper.

In Example 3, we only compare the performance of two-stage estimators with their weighted version. Table 3 indicates that the two-stage estimators with

**Table 3** Simulation results under the model with 20% contamination on X in Example 3

		GMCP-HT			WGMCP-HT		
		LS	Huber	Cauchy	LS	Huber	Cauchy
N(0,1)	$\ell_2$ error	7.49	7.52	7.54	6.74	6	4.97
	$\ell_1$ error	43.56	42.78	40.76	35.75	28.81	22.42
	MS	71.76	66.93	54.81	60.82	38.22	32.53
	GS	17.34	16.26	13.74	14.02	9.38	9.72
	FPR	11.64	10.63	8.15	9.28	4.6	3.38
	FNR	8.59	8.35	9.06	5.82	5.88	4.76
	GFPR	12.72	11.59	8.9	9.04	3.98	4.33
	GFNR	10.33	10.5	10.5	8	6	5.83
$t_1$	$\ell_2$ error	125.26	8.46	8.54	126.92	6.96	6.43
	$\ell_1$ error	2081.98	47.63	46.78	2099.27	33.87	30.84
	MS	96.16	61.26	52.46	86.38	37.42	35
	GS	22.76	15.2	14.12	19.26	9.07	9.56
	FPR	17.39	9.71	7.91	15.35	4.57	4.05
	FNR	28.29	15.65	16.12	27.94	9.71	9.18
	GFPR	19.76	10.93	9.77	16.06	4.02	4.56
	GFNR	30.17	17.83	17.67	30.67	11.83	12.17
Mix Cauchy	$\ell_2$ error	18.52	7.72	7.66	18.48	5.97	5.22
	$\ell_1$ error	211.8	43.23	39.92	210.62	28.68	24.67
	MS	75.62	64.31	48.28	63.31	37.8	35.46
	GS	18.17	15.69	12.39	13.98	9.41	10.16
	FPR	12.69	10.19	6.87	10.02	4.52	4
	FNR	15.76	11.18	11.29	12.29	6	4.94
	GFPR	14.09	11.16	7.59	9.41	4.09	4.74
	GFNR	17.83	13.33	12.33	14.5	7.17	5

The mean  $\ell_2$  error,  $\ell_1$  error, MS, GS, FPR (%), FNR(%), GFPR (%) and GFNR (%) out of 100 iterations are displayed

well-chosen  $w(\mathbf{x})$  perform better in all cases than the two-stage estimators with  $w(\mathbf{x}) = 1$ . Again when the errors are heavy-tailed ( $t_1$  and Mix Cauchy), the least squares estimator loses its efficiency and the re-descending estimators produced by Cauchy loss perform the best for all scenarios.

In summary, our simulation studies show that in the proposed two-stage M-estimator framework, (1) the GP Stage can utilize the grouping structure to yield satisfactory parameter estimation and group variable selection results for irregular settings, if a robust loss function (e.g., Huber loss and Cauchy loss) is used; (2) the HT Stage further improve the performance by filtering out the non-important selected variable from the first stage; and (3) the two-stage M-estimators with re-descending loss functions (e.g., Cauchy loss) and concave folded

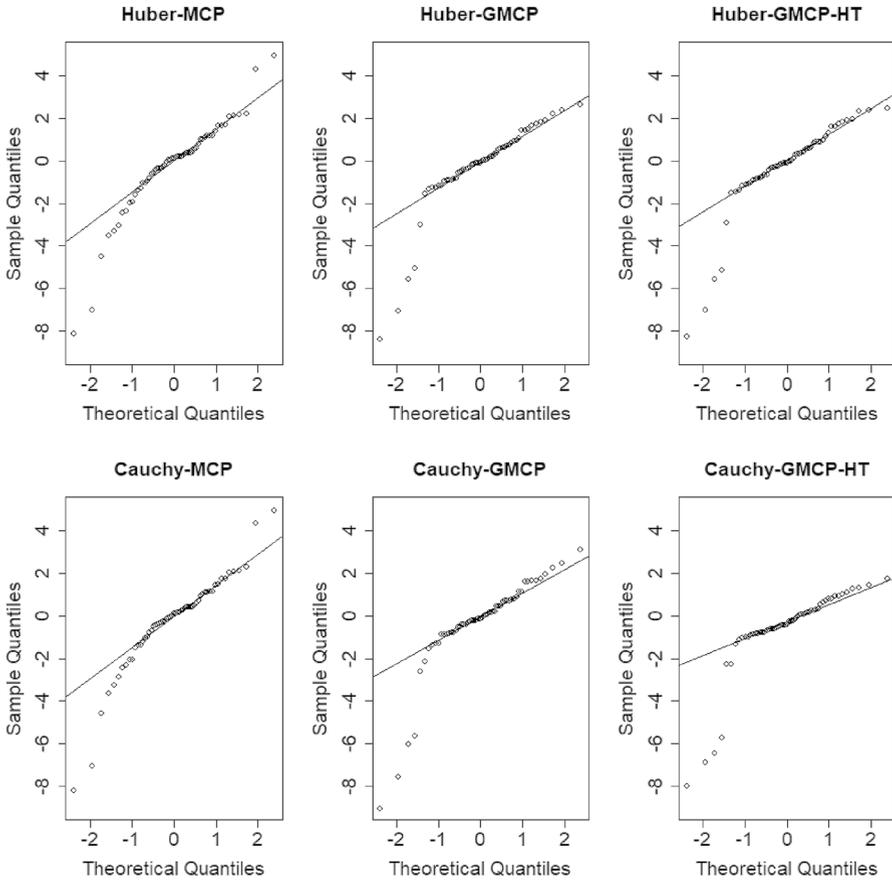
penalties consistently render more satisfactory results when data are heavy-tailed or strongly contaminated ( $t_1$  and Mix Cauchy).

## 6 Real data example

In this section, we use the NCI-60 data, a gene expression dataset collected from Affymetrix HG-U133A chip, to illustrate the performance of the proposed two-stage penalized M-estimators evaluated in Sect. 5. The NCI-60 data consist of data on 60 human cancer cell lines and can be downloaded via the Web application CellMiner (<http://discover.nci.nih.gov/cellminer/>). The study is to predict the protein expression on the KRT18 antibody from other gene expression levels. The expression level of the protein *keratin 18* is known to be persistently expressed in carcinomas (Oshima et al., 1996). After removing the missing data, there are  $n = 59$  samples with 21,944 genes in the dataset. One can refer Shankavaram et al. (2007) for more details.

We first perform some pre-screenings by keeping only 2000 genes with the largest variations and choosing 500 genes out of which are most correlated with the response variable. Then for each gene, we use B-spline with five bases to form a group with five variables. Thus, our final dataset has  $n = 59$  samples,  $p = 2500$  variables and  $J = 500$  groups. Similar to our simulation studies, we apply the non-group estimators, one-stage estimators and two-stage estimators to select important genes, with tuning parameter  $\lambda$  and  $\theta$  chosen from the 10-fold cross-validation with  $\lambda$  ranging in  $(0.01\sqrt{\frac{\log p}{n}}, 10\sqrt{\frac{\log p}{n}})$  and  $\theta$  ranging in  $(0.01, 1)$ . We also control the number of selected variables to be less than 75% of the sample size, in order to avoid overfitting. We report results from six methods: Huber-MCP, Cauchy-MCP, Huber-GMCP, Cauchy-GMCP, Huber-GMCP-HT, Cauchy-GMCP-HT.

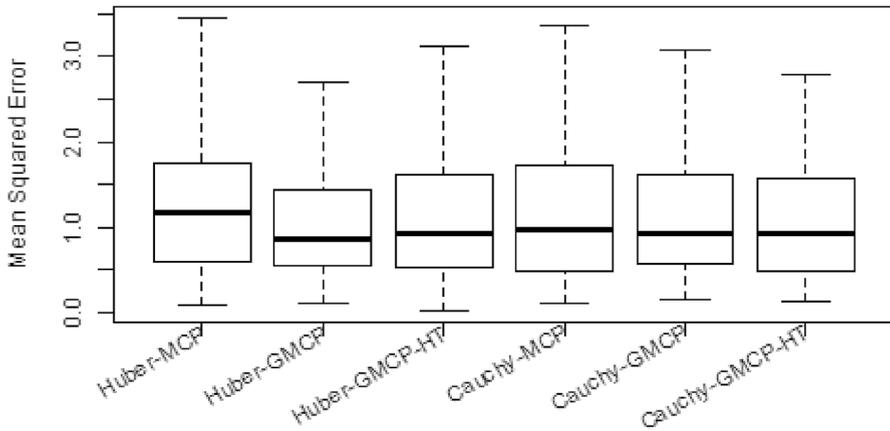
The QQ plots of the residuals generated from these six methods are shown in Fig. 1. It can be seen that each residual distribution has a longer tail on the left side, implying that the data may be contaminated or heavy-tailed. Table 4 displays the selected important genes. The numbers of selected genes from these methods are 9 (Huber-MCP), 3 (Huber-GMCP), 3 (Huber-GMCP-HT), 9 (Cauchy-MCP), 2 (Cauchy-GMCP) and 5 (Cauchy-GMCP-HT), respectively. It is not surprising that gene KRT8 is selected by all six methods due to its largest correlation with the response variable and the long history of being paired with KRT18 in cancer studies (Li & Zhou, 2016; Walker et al., 2007). Notice that gene INHBB is singled out by the four methods that incorporate grouping information. This gene has been shown to play essential roles in tumorigenesis and migration (Kita et al., 2017; Wijayarathna & De Kretser, 2016). Table 4 also shows that the Huber-MCP and Cauchy-MCP both select the same genes, which indicates that the contamination in the data may not be strong enough to cause different selection results between these two loss functions. In addition, it is reasonable to observe that the Huber-GMCP and Huber-GMCP-HT also select exactly the same genes, since there is no sparsity within each group in the data. It is also worth mentioning that the Cauchy-GMCP-HT selects more genes than the Cauchy-GMCP. It means that



**Fig. 1** QQ plots of the residuals from Huber–MCP, Cauchy–MCP, Huber–GMCP, Cauchy–GMCP, Huber–GMCP–HT, Cauchy–GMCP–HT

**Table 4** Selected genes by Huber–MCP, Cauchy–MCP, Huber–GMCP, Cauchy–GMCP, Huber–GMCP–HT, Cauchy–GMCP–HT

Huber–MCP	KRT8	NRN1	KRT18	GAS7	EPS8L2
	GPX3	TRIM29	LAD1	SEMA5A	
Huber–GMCP	KRT8	INHBB	PBX1		
Huber–GMCP–HT	KRT8	INHBB	PBX1		
Cauchy–MCP	KRT8	NRN1	KRT18	GAS7	EPS8L2
	GPX3	TRIM29	LAD1	SEMA5A	
Cauchy–GMCP	KRT8	INHBB			
Cauchy–GMCP–HT	KRT8	NR2F2	NOTCH3	INHBB	SIRPA



**Fig. 2** Boxplot of the mean squared error of predictions

the chosen  $\lambda$  of Cauchy–GMCP–HT is different from that of Cauchy–GMCP, as for the former the optimal  $\lambda$  and  $\theta$  are found together by the cross-validation.

For further investigation, we randomly choose six observations as the test set and applied those six methods to the rest patients to get the estimation of the coefficients and then compute the prediction error on the test set. We repeat the random splitting 100 times, and the boxplots of the mean squared error of predictions are shown in Fig. 2. It can be seen that the Huber–GMCP and Cauchy–GMCP perform better than the other methods. This is not surprising since there is only between-group sparsity in the dataset. In addition, Fig. 2 also shows that Cauchy-type estimators perform similarly to the corresponding Huber-type estimators, which indicates that when there exists only moderate contamination in the data, it may be sufficient to consider the convex Huber loss in our framework.

## 7 Discussion

Bi-level variable selection and parameter estimation are crucial when covariates function group-wisely in high dimensional settings. It has become even more challenging when data are contaminated or heavy-tailed. In this paper, we propose a two-stage penalized M-estimator framework for high-dimensional bi-level variable selection. This framework consists of two stages: penalized M-estimation with a concave  $\ell_2$ -norm penalty achieving the consistent group selection at the first stage, and a post-hard-thresholding operator to achieve the within-group sparsity at the second stage. The proposed framework is very general in that it covers a wide range of loss functions and penalty functions, allowing both functions to be non-convex. Thus, if the data are strongly contaminated, either in covariates or random error, we are still able to perform bi-level variable selection efficiently through the proposed framework.

Theoretically, we establish statistical properties of the proposed two-stage penalized M-estimator in ultra high-dimensional settings when  $p$  grows with  $n$  at an almost exponential rate. In particular, for the estimator at the Group Penalization Stage, we show its local estimation consistency at the minimax rate enjoyed by LS-GLasso and establish the local group selection consistency. For the post-hard-thresholding estimator at the second stage, we show that it naturally inherits all those nice statistical properties from the first stage and further possesses bi-level variable selection consistency. These theoretical results require weak assumptions on model settings and are applicable even though the random error and covariates are heavy-tailed or the dataset is contaminated by outliers.

Our framework is computationally efficient and is able to find a well-behaved local stationary point if a consistent initial such as Huber group Lasso is used. Our numerical studies show satisfactory finite sample performances of the two-stage penalized M-estimator under different irregular settings, which is consistent with our theoretical findings. In particular at the first stage, among some of the possible choices of loss and penalty functions that fit in the proposed framework, our numerical studies suggest considering a re-descending loss function, such as Cauchy loss or Tukey's biweight loss, with a group concave folded penalty, such as group MCP penalty, when the data are strongly contaminated.

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s10463-021-00809-z>) contains supplementary material, which is available to authorized users.

## References

- Brehehy, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, 71(3), 731–740.
- Brehehy, P., Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2(3), 369.
- Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.
- Friedman, J., Hastie, T., Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. arXiv preprint [arXiv:1001.0736](https://arxiv.org/abs/1001.0736).
- Guo, X., Zhang, H., Wang, Y., Wu, J.-L. (2015). Model selection and estimation in high dimensional regression models with group SCAD. *Statistics & Probability Letters*, 103, 86–92.
- Hill, R. W. (1977). Robust regression when there are outliers in the carriers. PhD thesis, Harvard University.
- Huang, J., Ma, S., Xie, H., Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2), 339–355.
- Huang, J., Brehehy, P., Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4), 481–499.
- Jiang, D., Huang, J. (2014). Concave 1-norm group selection. *Biostatistics*, 16(2), 252–267.
- Kita, A., Kasamatsu, A., Nakashima, D., Endo-Sakamoto, Y., Ishida, S., Shimizu, T., Kimura, Y., Miyamoto, I., Yoshimura, S., Shiiba, M., Tanzawa, H., Uzawa, K. (2017). Activin b regulates adhesion, invasiveness, and migratory activities in oral cancer: A potential biomarker for metastasis. *Journal of Cancer*, 8(11), 2033.
- Li, Z.-L., Zhou, S.-F. (2016). A silac-based approach elicits the proteomic responses to vancomycin-associated nephrotoxicity in human proximal tubule epithelial hk-2 cells. *Molecules*, 21(2), 148.
- Lilly, K. (2015). Robust variable selection methods for grouped data. PhD thesis, Auburn University.

- Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust  $m$ -estimators. *The Annals of Statistics*, 45(2), 866–896.
- Loh, P.-L., Wainwright, M. J. (2015). Regularized  $m$ -estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1), 559–616.
- Mallows, C. L. (1975). On some topics in robustness. Bell Telephone Laboratories. Unpublished memorandum.
- Merrill, H. M., Schweppe, F. C. (1971). Bad data suppression in power system static state estimation. *IEEE Transactions on Power Apparatus and Systems*, 6, 2718–2725.
- Müller, C. (2004). Redescending  $m$ -estimators in regression analysis, cluster analysis and image analysis. *Discussiones Mathematicae Probability and Statistics*, 24(1), 59–75.
- Nesterov, Y. (2013). Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1), 125–161.
- Oshima, R. G., Baribault, H., Caulin, C. (1996). Oncogenic regulation and function of keratins 8 and 18. *Cancer and Metastasis Reviews*, 15(4), 445–471.
- Shankavaram, U. T., Reinhold, W. C., Nishizuka, S., Major, S., Morita, D., Chary, K. K., Reimers, M. A., Scherf, J., Kahn, A., Dolginow, D., Cossman, J., Kaldjian, E. P., Scudiero, D. A., Petricoin, E., Liotta, L., Lee, J. K., Weinstein, J. N. (2007). Transcript and protein expression profiles of the nci-60 cancer cell panel: An integromic microarray study. *Molecular Cancer Therapeutics*, 6(3), 820–832.
- Shevlyakov, G., Morgenthaler, S., Shurygin, A. (2008). Redescending  $m$ -estimators. *Journal of Statistical Planning and Inference*, 138(10), 2906–2917.
- Simon, N., Friedman, J., Hastie, T., Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2), 231–245.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267–288.
- Walker, L. C., Harris, G. C., Hooloway, A. J., Mckenzie, G. W., Wells, J. E., Robinson, B. A., Morrisa, C. M. (2007). Cytokeratin krt8/18 expression differentiates distinct subtypes of grade 3 invasive ductal carcinoma of the breast. *Cancer Genetics and Cytogenetics*, 178(2), 94–103.
- Wang, M., Tian, G.-L. (2016). Robust group non-convex estimations for high-dimensional partially linear models. *Journal of Nonparametric Statistics*, 28(1), 49–67.
- Wei, F., Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli: Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, 16(4), 1369.
- Wijayarathna, R., De Kretser, D. M. (2016). Activins in reproductive biology and beyond. *Human Reproduction Update*, 22(3), 342–357.
- Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.
- Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2), 894–942.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.