Electronic Journal of Statistics
Vol. 17 (2023) 2912–2961
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2167

A convex-nonconvex strategy for grouped variable selection

Xiaoqian Liu*

Department of Bioinformatics and Computational Biology,
The University of Texas MD Anderson Cancer Center,
Houston, TX 77030, USA
e-mail: xliu31@mdanderson.org

Aaron J. Molstad

School of Statistics,
University of Minnesota,
Minneapolis, MN 55455, USA
e-mail: amolstad@umn.edu

Eric C. Chi

Department of Statistics,
Rice University,
Houston, TX 77005, USA
e-mail: echi@rice.edu

Abstract: This paper deals with the grouped variable selection problem. A widely used strategy is to augment the negative log-likelihood function with a sparsity-promoting penalty. Existing methods include the group Lasso, group SCAD, and group MCP. The group Lasso solves a convex optimization problem but suffers from underestimation bias. The group SCAD and group MCP avoid this estimation bias but require solving a nonconvex optimization problem that may be plagued by suboptimal local optima. In this work, we propose an alternative method based on the generalized minimax concave (GMC) penalty, which is a folded concave penalty that maintains the convexity of the objective function. We develop a new method for grouped variable selection in linear regression, the group GMC, that generalizes the strategy of the original GMC estimator. We present a primal-dual algorithm for computing the group GMC estimator and also prove properties of the solution path to guide its numerical computation and tuning parameter selection in practice. We establish error bounds for both the group GMC and original GMC estimators. A rich set of simulation studies and a real data application indicate that the proposed group GMC approach outperforms existing methods in several different aspects under a wide array of scenarios.

Keywords and phrases: Sparse linear regression, convex optimization, convex-nonconvex penalization, high-dimensional data analysis.

Received February 2022.

*Corresponding author.

1. Introduction

Consider the classical linear regression setting where the data have been generated according to the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{\star} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a response vector, $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a fixed design matrix whose columns are p covariate variables, and $\boldsymbol{\epsilon}$ is a vector of independent noise variables with mean zero and variance σ^2. In modern statistical applications, we often have $p \gg n$ where the ordinary least squares estimator is not well defined.

A natural strategy to address this issue is to assume that $\boldsymbol{\beta}^{\star}$ is sparse so as to improve both the prediction accuracy and the interpretation of the model. To that end, Tibshirani [32] developed the least absolute shrinkage and selection operator (Lasso) which performs both coefficient estimation and variable selection. The Lasso estimator is a solution to

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}}\, \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1,$$

where the objective function is a sum of the squared error loss, which represents the lack-of-fit, and the l_1-norm penalty, which encourages sparsity in the estimated model. The l_1-norm is able to perform variable selection because of its singularity at the origin. The nonnegative tuning parameter λ balances the trade-off between the goodness-of-fit and the model complexity.

The Lasso is one of the most popular penalized regression formulations for selecting individual variables but cannot immediately deal with certain types of structured sparsity. For example, in many statistical applications, variables may have a natural group structure. A classic example in regression is the encoding of a single categorical variable using a group of dummy variables. In such case, what we need is a method for selecting, or not selecting, the entire set of dummy variables. The most prominent work in grouped variable selection is the group Lasso [39], which is a natural extension of Lasso and solves the following penalized least squares problem:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}}\, \frac{1}{2n}\left\|\mathbf{y} - \sum_{j=1}^{J}\mathbf{X}_{\cdot,j}\boldsymbol{\beta}_j\right\|_2^2 + \lambda\sum_{j=1}^{J}K_j\|\boldsymbol{\beta}_j\|_2, \tag{1.2}$$

where the p covariates are divided into J groups, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathsf{T}},\ldots,\boldsymbol{\beta}_J^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^p$ with $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$ and $\sum_{j=1}^{J} p_j = p$. The matrix $\mathbf{X}_{\cdot,j}$ is the submatrix of $\mathbf{X}$ whose columns correspond to the variables in the j-th group. The K_j's are nonnegative weights used to adjust for the group sizes. A typical choice of K_j is $\sqrt{p_j}$. The group Lasso employs the l_2-norm of the group coefficients as a component of the penalty function. One can also view the penalty as applying the l_1-norm to the vector of l_2-norms of the groups, which enforces sparsity at the group level while encouraging ridge regression-like shrinkage within a group.

Since the introduction of the group Lasso, many variants and generalizations have been proposed and investigated. Kim, Kim and Kim [20] designed a blockwise sparse regression method to extend the idea of the group Lasso to general loss functions but used the same penalty as the group Lasso. Meier, Van De Geer and Bühlmann [24] derived the group Lasso for logistic regression and presented an efficient algorithm for fitting related generalized linear models. Zhao et al. [43] developed a family of composite absolute penalties for grouped and hierarchical variable selection, which includes the group Lasso as a special case. Wei and Huang [37] generalized the adaptive Lasso to an adaptive group Lasso method to improve the variable selection performance. The grouping information included in the models discussed above has led to improvements in both estimation accuracy and model interpretability. Many applications can be found in the corresponding references.

Nonetheless, despite its many desirable characteristics, the group Lasso and its variants suffer from the same drawback as the Lasso, namely, they tend to underestimate large magnitude coefficients due to applying the same amount of shrinkage on all coefficients. Nonconvex penalties, such as the smoothly clipped absolute deviation (SCAD) [10] and the minimax concave penalty (MCP) [41], have been developed as alternatives to the Lasso that can diminish the estimation bias in the Lasso estimator. By applying such nonconvex penalties to the l_2-norms of the group coefficients, it is natural to obtain grouped variable selection via nonconvex penalization, such as the group SCAD and group MCP [35, 36, 18].

Grouped variable selection models with nonconvex penalties are not without their disadvantages, however. The nonconvex penalty, though beneficial for the estimation of coefficients, leads to a nonconvex optimization problem. Objective functions in nonconvex programming typically possess multiple local optima which are not global optima, and algorithms for solving nonconvex optimization problems such as gradient or coordinate descent may be trapped in local optima. Early work on statistical theory for SCAD or MCP penalized least squares estimators focused on either error bounds for global optima [42] or local optima obtained through specific initialization schemes and algorithms [11]. More recently, Loh and Wainwright [23] established statistical properties which apply to all stationary points of SCAD or MCP penalized least squares objective functions (though their results do not apply directly to group SCAD or MCP penalized estimators). However, empirical results from [11] and [23], among others, suggest that in practice some stationary points perform much better than others, especially when the overall objective function is highly nonconvex, e.g., see the Remark on (α_1, μ) and Figure 4 of [23].

To overcome the drawbacks of nonconvex optimization, one line of research, commonly referred to as the convex-nonconvex strategy, has been studied in the field of signal processing. This strategy adopts the so-called convexity-preserving nonconvex penalization, namely, the penalties are nonconvex but capable of maintaining the convexity of the whole objective function. The idea of convexity-preserving nonconvex penalties was introduced by [5], [26], and [27], and then further investigated in [4], [29], and [44]. In particular, Selesnick [30] proposed

a novel nonconvex penalty function for the regularized least squares problem, which they call the generalized minimax concave (GMC) penalty. The GMC penalty is expressed as

$$\psi_{\mathbf{B}}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 - \min_{\mathbf{v}\in\mathbb{R}^p} \left\{ \|\mathbf{v}\|_1 + \frac{1}{2}\|\mathbf{B}(\boldsymbol{\beta}-\mathbf{v})\|_2^2 \right\}, \tag{1.3}$$

which can guarantee the convexity of the optimization problem under a suitable condition on the matrix parameter $\mathbf{B}$. Lanza et al. [21] and Selesnick et al. [31] further developed a general strategy to construct parametric noncovex nonseparable regularizers for linear inverse problems. Abe, Yamagishi and Yamada [1] extended the idea of the GMC to solving linearly involved sparse estimation problems, such as piecewise constant signal recovery. Abe, Yamagishi and Yamada [2] further extended the GMC framework to more general regularizers, and their follow-up work [38] investigated linearly involved sparse estimates under additional constraints. Liu and Chi [22] revisited the linearly involved GMC penalty, proposing a new method for choosing the matrix parameter $\mathbf{B}$ in the penalty and providing an additional algorithm to compute the solution path of the linearly involved GMC model. The convex-nonconvex strategy bridges the gap between convex and nonconvex approaches in prior work. The nonconvex penalty can greatly mitigate the estimation bias for large magnitude coefficients, while the convexity of the optimization problem guarantees that all local minima are global minima and opens the door to using many efficient algorithms that are available for convex optimization problems.

In this paper, we focus on grouped variable selection in linear regression and propose a new generalization of the GMC penalty, called the group GMC penalty, which is also a convexity-preserving nonconvex penalty. We present the convexity-preserving condition for the group GMC model and some properties of its solution path. To solve the proposed optimization problem, we cast it as a saddle-point problem and provide a primal-dual algorithm for iteratively computing its saddle point. Theoretically, we establish error bounds for both the group GMC penalized least squares estimator and, as a special case, the GMC estimator of [30], which to the best of our knowledge have not been established yet. We evaluate the effectiveness of the proposed approach by comparing it with existing grouped variable selection methods through a bunch of simulation experiments and a real data application.

The rest of this paper is organized as follows. In Section 2, we first review the GMC penalty and its relation to existing folded concave penalties. Then we formulate the group GMC penalty and the corresponding optimization problem for grouped variable selection in linear regression. We also include the convexity-preserving condition and some theoretical properties of the solution path in this section. In Section 3, we present in detail how to solve the proposed optimization problem with a first-order primal-dual algorithm. In Section 4, we study the statistical properties of both the group GMC estimator and the original GMC estimator by establishing l_2-norm error bounds. In Sections 5 and 6, we report on numerical experiments and a real data application. We close with a discussion in Section 7. All proofs in this paper are included in Appendix A.

2. Group GMC

2.1. GMC and MCP

We first review the GMC penalty to help readers understand the relationship between the GMC and the MCP in [41]. For that purpose, we have to recall the definition of the infimal convolution and the Huber function.

The infimal convolution of two functions f and g is

$$(f\square g)(\boldsymbol{\beta}) = \inf_{\mathbf{v}\in\mathbb{R}^p} \{f(\mathbf{v}) + g(\boldsymbol{\beta}-\mathbf{v})\}.$$

The Huber function [19] is defined as

$$h(\beta) = \begin{cases} \frac{1}{2}\beta^2, & \text{if } |\beta| \le 1, \\ |\beta| - \frac{1}{2}, & \text{if } |\beta| > 1, \end{cases}$$

which can be equivalently expressed as the infimal convolution

$$h(\beta) = \min_{v\in\mathbb{R}} \left\{ |v| + \frac{1}{2}(\beta - v)^2 \right\},$$

where the infimum is replaced by a minimum since the infimum in the definition is attained.

Selesnick [30] defined the scaled version of the Huber function as

$$h_b(\beta) = h(b^2\beta)/b^2 = \min_{v\in\mathbb{R}} \left\{ |v| + \frac{1}{2}b^2(\beta - v)^2 \right\},$$

where $b \neq 0$ is a scalar parameter. In the special case where $b = 0$, $h_b(\beta) = 0$. Based on the scaled Huber function, the scaled minimax concave (MC) penalty is given by

$$\phi_b(\beta) = |\beta| - h_b(\beta). \tag{2.1}$$

After defining the scaled Huber function in the univariate case, Selesnick [30] proposed a natural multivariate generalization. Given a matrix parameter $\mathbf{B}$, the generalized Huber function $H_{\mathbf{B}} : \mathbb{R}^p \to \mathbb{R}$ is written as

$$H_{\mathbf{B}}(\boldsymbol{\beta}) = \inf_{\mathbf{v}\in\mathbb{R}^p} \left\{ \|\mathbf{v}\|_1 + \frac{1}{2}\|\mathbf{B}(\boldsymbol{\beta}-\mathbf{v})\|_2^2 \right\},$$

which is a convex function; the infimum is attained since the l_1-norm is coercive. Mimimicking the univariate scaled minimax MC penalty, the generalized MC (GMC) penalty is defined as the difference of the l_1-norm and the generalized Huber function.

$$\psi_{\mathbf{B}}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 - H_{\mathbf{B}}(\boldsymbol{\beta}), \tag{2.2}$$

which coincides with (1.3). Since the difference of two convex functions is not necessarily convex, the GMC penalty function is in general nonconvex. But as

mentioned in the introduction, the GMC penalty can maintain the convexity of the penalized least squares problem by suitably choosing $\mathbf{B}$. Details can be found in [30].

Recall that the MCP defined in [41] is expressed as

$$P_{\lambda,\gamma}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \rho_{\lambda,\gamma}(|\beta_j|), \tag{2.3}$$

where the univariate MCP function defined on $[0, \infty)$ is

$$\rho_{\lambda,\gamma}(|\beta_j|) = \begin{cases} \lambda|\beta_j| - \frac{\beta_j^2}{2\gamma}, & \text{if } |\beta_j| \le \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |\beta_j| > \gamma\lambda, \end{cases} \tag{2.4}$$

where $\lambda \ge 0$ is the tuning parameter controlling the degree of penalization, and $\gamma > 1$ is a hyper-parameter that determines the degree of concavity of the MCP. The MCP function converges pointwise to the l_1-norm as $\gamma \to \infty$ and to the l_0-norm as $\gamma \to 1$, therefore it provides a continuum of penalties by varying the value of γ.

Now let us have a closer look at the similarities and differences between the GMC penalty (2.2) and the MCP (2.3). In the univariate case, the GMC penalty coincides with the scaled MC penalty (2.1) and the MCP (2.3) reduces to $P_{\lambda,\gamma}(\beta) = \rho_{\lambda,\gamma}(|\beta|)$. If we set $b^2 = 1/\gamma\lambda$, then $\rho_{\lambda,\gamma}(|\beta|) = \lambda\phi_b(\beta)$. In other words, (2.2) is equivalent to (2.3) up to a factor of λ. The difference between the GMC penalty (2.2) and the MCP (2.3) lies in how they are generalized from the univariate case to the multivariate one. The MCP (2.3) takes an additive form from the univariate MCP function (2.4), while the GMC penalty (2.2) is derived from the scaled MC penalty (2.1) via an infimal convolution, thus leading to a non-separable penalty function whenever $\mathbf{B}^\mathsf{T}\mathbf{B}$ is non-diagonal.

The implications of expressing the MC penalty as an infimal convolution are non-trivial and lead to intrinsic differences with the standard MCP. It is well known that in the classic low-dimensional case where $n > p$, there exists a suitable hyper-parameter γ choice for MCP that leads to a convex objective function but that no such γ exists when $n < p$. In contrast, we will see that it is always possible to find a matrix $\mathbf{B}$ for GMC that leads to a convex objective function for any n and p. Thus, the GMC function enables the application of folded concave penalties in the high-dimensional case where $n \ll p$ without sacrificing convexity, opening the door to methods that can enjoy the best of both convex and nonconvex worlds.

2.2. The group GMC model

Based on the form of the GMC penalty (1.3) and mimicking the generalization from the Lasso to the group Lasso, we define the group GMC penalty as

$$\Phi_{\mathbf{B}}(\boldsymbol{\beta}) = \sum_{j=1}^{J} K_j\|\boldsymbol{\beta}_j\|_2 - \min_{\mathbf{v}\in\mathbb{R}^p} \left\{ \sum_{j=1}^{J} K_j\|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}(\boldsymbol{\beta} - \mathbf{v})\|_2^2 \right\}, \tag{2.5}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\mathsf{T}, \ldots, \boldsymbol{\beta}_J^\mathsf{T})^\mathsf{T} \in \mathbb{R}^p$ and $\mathbf{v} = (\mathbf{v}_1^\mathsf{T}, \ldots, \mathbf{v}_J^\mathsf{T})^\mathsf{T} \in \mathbb{R}^p$. For each $j \in [J]$ where $[J] = \{1, \ldots, J\}$ for a positive integer J, $\boldsymbol{\beta}_j, \mathbf{v}_j \in \mathbb{R}^{p_j}$ with $\sum_{j=1}^J p_j = p$, and K_j is the same as that in the group Lasso model (1.2). Here we insert a multiplier $1/n$ in the squared term of the group GMC penalty to put it on the same scale with the squared error loss term in (1.2).

Therefore, the group GMC model for grouped variable selection and coefficient estimation in linear regression is cast as the following optimization problem:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \Phi_{\mathbf{B}}(\boldsymbol{\beta}), \tag{2.6}$$

where $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \|\mathbf{y} - \sum_{j=1}^J \mathbf{X}_{\cdot,j}\boldsymbol{\beta}_j\|_2^2$ as in (1.2). Here $\lambda \geq 0$ is again a tuning parameter that controls the degree of penalization, while $\mathbf{B}$ is a matrix parameter that controls the concavity of the group GMC penalty. Note that in our paper, we refer to λ as the tuning parameter of the group GMC and treat the matrix $\mathbf{B}$ as a hyper-parameter.

Similar to the GMC approach, the basic idea of the group GMC method is to maintain the convexity of the optimization problem while using a nonconvex penalty, which can be realized with an appropriate choice of the matrix hyper-parameter $\mathbf{B}$. The next proposition specifies the condition that $\mathbf{B}$ has to satisfy to guarantee the convexity of problem (2.6). Recall that for two matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \succeq \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is positive semi-definite; similarly, $\mathbf{A} \succ \mathbf{B}$ means $\mathbf{A} - \mathbf{B}$ is positive definite.

Proposition 2.1. *Let* $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times p}$, *and* $\lambda \geq 0$. *Define* $F : \mathbb{R}^p \to \mathbb{R}$ *as in* (2.6)

$$F(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \Phi_{\mathbf{B}}(\boldsymbol{\beta}), \tag{2.7}$$

where $\Phi_{\mathbf{B}} : \mathbb{R}^p \to \mathbb{R}$ *is the group GMC penalty* (2.5). *If*

$$\mathbf{X}^\mathsf{T}\mathbf{X} \succeq \lambda \mathbf{B}^\mathsf{T}\mathbf{B}, \tag{2.8}$$

then F *is a convex function. We call* (2.8) *the convexity-preserving condition for the group GMC problem* (2.6).

Note that the convexity-preserving condition (2.8) can hold without any restriction on the problem dimension p and the sample size n, namely, it can hold for both the low-dimensional case ($n \geq p$) and the high-dimensional case ($n < p$). To satisfy the convexity-preserving condition (2.8), an intuitive and simple choice for $\mathbf{B}$ is

$$\mathbf{B} = \sqrt{\alpha/\lambda}\mathbf{X}, \quad 0 \leq \alpha \leq 1. \tag{2.9}$$

We refer to α as the convexity-preserving parameter of the group GMC model since α controls the nonconvexity of the group GMC penalty. Setting $\alpha = 0$ reduces the group GMC penalty to the group Lasso penalty. And setting $\alpha = 1$ gives a maximally nonconvex penalty which can maintain the convexity of the optimization problem (2.6). The convexity-preserving parameter α is another

hyper-parameter of the group GMC method and needs to be chosen by users. We recommend a range of $0.4 < \alpha < 1$ based on our simulation studies in Section 5.

The following proposition establishes the relationship between the group GMC and group MCP. It also clarifies the relationship between the GMC and MCP as a by-product.

Proposition 2.2. *The group GMC method is equivalent to the group MCP method when* $\mathbf{B}^\mathsf{T}\mathbf{B}$ *is diagonal and the diagonal elements are suitably designed. This equivalence also holds for the GMC and MCP.*

We write the group GMC estimator, namely a minimizer to problem (2.6), as $\hat{\boldsymbol{\beta}}(\lambda)$ which explicitly represents the dependency of the solution to (2.6) on the tuning parameter λ. We next discuss two properties of the solution path $\hat{\boldsymbol{\beta}}(\lambda)$, that expedite the numerical computation in practice.

Theorem 2.1. *Suppose* $\mathbf{X}^\mathsf{T}\mathbf{X} \succ \lambda\mathbf{B}^\mathsf{T}\mathbf{B}$*, then the solution path* $\hat{\boldsymbol{\beta}}(\lambda)$ *to the group GMC problem* (2.6) *exists, is unique, and is continuous in* λ*.*

Theorem 2.1 tells us that the optimization problem (2.6) is well-posed. Moreover, continuity of $\hat{\boldsymbol{\beta}}(\lambda)$ opens the door to a homotopy strategy to reduce computation time when solving a sequence of problems over a grid of λ values. Namely, we use the solution to the problem at the previous value of λ to initialize, or warm start, the next iterate for computing the solution at the next λ value.

Intuitively, we may expect that all groups are excluded from the model when the tuning parameter λ is sufficiently large. The following theorem confirms our intuition.

Theorem 2.2. *The group GMC problem* (2.6) *has a unique solution* $\hat{\boldsymbol{\beta}}(\lambda) = \mathbf{0}_p$ *for all* λ *greater than* $\lambda_0 = \max_j \left\{\|(\mathbf{X}_{\cdot,j})^\mathsf{T}\mathbf{y}\|_2/(nK_j)\right\}$*, where* $\mathbf{X}_{\cdot,j}$ *and* K_j *are as defined in* (1.2) *for* $j \in [J]$*.*

This second property is practically useful since it gives a range of λ, $[0, \lambda_0]$, to sample the full dynamic range of group sparse models, and as an added benefit the computation of λ_0 is straightforward.

We close this section with a few remarks. First, the group GMC penalty (2.5) depends on $\mathbf{B}^\mathsf{T}\mathbf{B}$, not $\mathbf{B}$ itself. Therefore, there is no need to express $\mathbf{B}$ explicitly when computing the solution path $\hat{\boldsymbol{\beta}}(\lambda)$. Second, the two properties of the solution path hold for any matrix $\mathbf{B}$ satisfying the convexity-preserving condition and are independent of how $\hat{\boldsymbol{\beta}}(\lambda)$ is computed, as they are intrinsic to the group GMC problem. Finally, Theorem 2.1 applies only in the classic setting where $n > p$. This is a more stringent condition than what is required to ensure the uniqueness of the Lasso [33]. The proof of the uniqueness of the Lasso solution hinged on the Karush-Kuhn-Tucker (KKT) conditions of the Lasso optimization problem. The KKT conditions for the group GMC problem, however, are more complicated than the KKT conditions for the Lasso. Generalization of the proof used in the Lasso case to the group GMC is not straightforward due to the more complicated KKT conditions of the latter. Nonetheless, we conjecture that relaxed conditions similar to those that ensure the uniqueness of the Lasso

Algorithm 1 Basic PDHG steps for problem (3.1)

Set $\mathbf{x}_0 \in \mathbb{R}^N, \mathbf{y}_0 \in \mathbb{R}^M, \sigma_k > 0, \tau_k > 0$
repeat
  $\hat{\mathbf{x}}_{k+1} = \mathbf{x}_k - \tau_k \mathbf{A}^\mathsf{T} \mathbf{y}_k$
  $\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x}) + \frac{1}{2\tau_k}\|\mathbf{x} - \hat{\mathbf{x}}_{k+1}\|_2^2$
  $\hat{\mathbf{y}}_{k+1} = \mathbf{y}_k + \sigma_k \mathbf{A}(2\mathbf{x}_{k+1} - \mathbf{x}_k)$
  $\mathbf{y}_{k+1} = \arg\min_{\mathbf{y}\in\mathbb{R}^M} g(\mathbf{y}) + \frac{1}{2\sigma_k}\|\mathbf{y} - \hat{\mathbf{y}}_{k+1}\|_2^2$
until convergence

solution can be established and leave establishing these conditions for future work.

3. Algorithms

3.1. Algorithm for the group GMC model

In this subsection, we focus on the computation of the solution path $\hat{\boldsymbol{\beta}}(\lambda)$ to the group GMC model (2.6). We first review the Primal-Dual Hybrid Gradient (PDHG) algorithm [9, 7] for computing the solution to non-smooth saddle-point problems. Then we formulate problem (2.6) as an instance of the kind of saddle-point problem that the PDHG algorithm can solve.

The PDHG method, also known as the Chambolle-Pock method, is widely used to solve the following saddle-point problem:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \; f(\mathbf{x}) + \mathbf{y}^\mathsf{T}\mathbf{A}\mathbf{x} - g(\mathbf{y}), \tag{3.1}$$

where f and g are closed and convex functions, $\mathbf{A} \in \mathbb{R}^{M\times N}$ is a matrix, and $\mathcal{X} \subset \mathbb{R}^N$ and $\mathcal{Y} \subset \mathbb{R}^M$ are convex sets. A wide range of problems in statistics and machine learning can be cast as a instance of (3.1), such as the scaled Lasso and total variation denoising.

Algorithm 1 summarizes the basic PDHG steps for problem (3.1), where σ_k and τ_k are step-size parameters for updating $\mathbf{x}$ and $\mathbf{y}$, respectively. One can choose constant step-sizes, $\tau_k = \tau$ and $\sigma_k = \sigma$ with $\tau\sigma < \|\mathbf{A}^\mathsf{T}\mathbf{A}\|^{-1}$, to guarantee the convergence of the PDHG algorithm. Note that we use $\|\mathbf{A}\|$ to denote the spectral norm of a matrix $\mathbf{A}$.

We now recast the optimization problem (2.6) as a saddle-point problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \max_{\mathbf{v}\in\mathbb{R}^p} \; f(\boldsymbol{\beta}) + \mathbf{v}^\mathsf{T}\mathbf{Z}\boldsymbol{\beta} - g(\mathbf{v}), \tag{3.2}$$

where

$$f(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{J} K_j\|\boldsymbol{\beta}_j\|_2 - \frac{\lambda}{2n}\|\mathbf{B}\boldsymbol{\beta}\|_2^2, \tag{3.3}$$

and

$$g(\mathbf{v}) = \frac{\lambda}{2n}\|\mathbf{B}\mathbf{v}\|_2^2 + \lambda\sum_{j=1}^{J} K_j\|\mathbf{v}_j\|_2, \tag{3.4}$$

and $\mathbf{Z} = \frac{\lambda}{n}\mathbf{B}^\mathsf{T}\mathbf{B} \in \mathbb{R}^{p\times p}$ is a symmetric matrix. In addition, both (3.3) and (3.4) are convex functions under the convexity-preserving condition (2.8). It is straightforward to see that problem (3.2) is under the framework of (3.1) and thereby can be solved by the PDHG algorithm.

The basic PDHG method can be slow to converge, however. In this paper, we employ a faster adaptive PDHG algorithm [15, 12], to solve the group GMC problem. We provide details about the adaptive PDHG for problem (3.2) and discuss its convergence guarantees in Appendix B.

3.2. Algorithm for the PDHG updates

The PDHG algorithm for solving the group GMC problem requires solving two subordinate optimization problems for updating $\boldsymbol{\beta}_{k+1}$ and $\mathbf{v}_{k+1}$ which we describe next.

We first introduce an efficient algorithm, Fast Adaptive Shrinkage/Thresholding Algorithm (FASTA) [13, 14], for solving optimization problems of the form

$$\underset{\mathbf{x}\in\mathbb{R}^N}{\text{minimize}}\ m(\mathbf{x}) + h(\mathbf{x}), \tag{3.5}$$

where m is convex and Lipschitz differentiable, h is proper, lower semi-continuous and convex, and $m+h$ is coercive. FASTA provides a simple framework for implementing the forward-backward splitting (FBS) method, also known as the proximal gradient method [3], to efficiently compute the solution to problem (3.5). Problems under the framework of (3.5) include the Lasso, noisy matrix completion, and many other regularized regression problems.

Algorithm 2 shows pseudocode of the basic FBS steps in FASTA for solving (3.5), where t_k is a positive step-size parameter and plays an important role in the convergence rate of the algorithm. The proximal operator of h is given by

$$\text{prox}_h(\mathbf{u}) = \underset{\mathbf{x}\in\mathbb{R}^N}{\text{argmin}}\left(h(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2\right).$$

The proximal operator exists and is unique if h is convex and lower semi-continuous. The key computation in FBS is the proximal mapping, and many regularizers h in sparse learning admit proximal operators which either have an explicit formula or can be evaluated by an efficient algorithm. For instance, the proximal operator of the l_2-norm can be explicitly expressed as

$$\text{prox}_{\lambda\|\cdot\|_2}(\mathbf{u}) = \left(1 - \frac{\lambda}{\|\mathbf{u}\|_2}\right)_+ \mathbf{u},$$

Algorithm 2 Basic FBS steps in FASTA for problem (3.5)

Set $\mathbf{x}_0 \in \mathbb{R}^N, t_k > 0$
1: **repeat**
2: $\quad \hat{\mathbf{x}}_{k+1} = \mathbf{x}_k - t_k \nabla m(\mathbf{x}_k)$
3: $\quad \mathbf{x}_{k+1} = \text{prox}_{t_k h}(\hat{\mathbf{x}}_{k+1}) = \text{argmin}_{\mathbf{x}\in\mathbb{R}^N}\, t_k h(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \hat{\mathbf{x}}_{k+1}\|_2^2$
4: **until** convergence

where $(\cdot)_+ = \max\{0, \cdot\}$. The efficiency of computing the proximal operators of many commonly used penalty functions makes the FBS method popular in practice.

We now rewrite the two optimization problems for updating $\boldsymbol{\beta}_{k+1}$ and $\mathbf{v}_{k+1}$ in the PDHG updates in the form of (3.5). First, the optimization problem for updating $\boldsymbol{\beta}_{k+1}$ can be written as

$$\begin{aligned}\boldsymbol{\beta}_{k+1} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}}\, f(\boldsymbol{\beta}) + \frac{1}{2\tau_k}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k+1}\|_2^2 \\ &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 - \frac{\lambda}{2n}\|\mathbf{B}\boldsymbol{\beta}\|_2^2 + \frac{1}{2\tau_k}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k+1}\|_2^2 \right\} + \lambda \sum_{j=1}^{J} K_j \|\boldsymbol{\beta}_j\|_2. \end{aligned} \tag{3.6}$$

Similarly, we write the optimization problem for updating $\mathbf{v}_{k+1}$ as

$$\begin{aligned}\mathbf{v}_{k+1} &= \underset{\mathbf{v}\in\mathbb{R}^p}{\text{argmin}}\, g(\mathbf{v}) + \frac{1}{2\sigma_k}\|\mathbf{v} - \hat{\mathbf{v}}_{k+1}\|_2^2 \\ &= \underset{\mathbf{v}\in\mathbb{R}^p}{\text{argmin}} \left\{ \frac{\lambda}{2n}\|\mathbf{B}\mathbf{v}\|_2^2 + \frac{1}{2\sigma_k}\|\mathbf{v} - \hat{\mathbf{v}}_{k+1}\|_2^2 \right\} + \lambda \sum_{j=1}^{J} K_j \|\mathbf{v}_j\|_2. \end{aligned} \tag{3.7}$$

Under the convexity-preserving condition (2.8), both (3.6) and (3.7) satisfy the conditions on m and h in (3.5). Therefore, we can compute $\boldsymbol{\beta}_{k+1}$ and $\mathbf{v}_{k+1}$ by using Algorithm 2.

Both the gradient of m and proximal operator of the l_2 norm required by FBS admit simple analytical expressions, ensuring the tractability of both (3.6) and (3.7). One of the primary difficulties with FBS is that users must carefully choose the step-size. Fortunately, many variants of FBS are available in FASTA for adaptively choosing the step-size and accelerating convergence. In this paper, we use the strategies adopted in the R package `fasta` to implement FASTA to compute the solutions to (3.6) and (3.7).

4. Statistical properties

4.1. Main results

In this section, we consider the statistical properties of the group GMC estimator obtained by solving (2.6). First, we demonstrate that the group GMC

estimator achieves an error bound of the same asymptotic order as existing estimators. Second, we discuss how the choice of $\mathbf{B}$, or $\mathbf{B}^\mathsf{T}\mathbf{B}$ to be exact, affects the error bound. We will also contrast our assumptions and error bounds with $\mathbf{B} = \sqrt{\alpha/\lambda}\mathbf{X}$ to those under the group Lasso penalization.

We now define a number of important quantities. First, define

$$\mathbf{v}^\star \in \operatorname*{argmin}_{\mathbf{v}\in\mathbb{R}^p} \left\{ \sum_{j=1}^{J} K_j \|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}(\boldsymbol{\beta}^\star - \mathbf{v})\|_2^2 \right\}, \tag{4.1}$$

where $\boldsymbol{\beta}^\star$ is the true vector of coefficients and $\mathbf{v}_j$ has the same dimension as $\boldsymbol{\beta}_j^\star$ for each $j \in [J]$. Implicitly, $\mathbf{v}^\star$ is a function of $\mathbf{B}, \boldsymbol{\beta}^\star, n$, and K_j: we avoid notation indicating this dependence for improved readability. We can then define the sets $\mathcal{S} = \{j : \|\boldsymbol{\beta}_j^\star\|_2 \neq 0, j \in [J]\}$ and $\mathcal{S}^c = [J] \setminus \mathcal{S}$ and use $|\mathcal{S}|$ to denote the cardinality of $\mathcal{S}$. We also define

$$\nu_j = \begin{cases} K_j + n^{-1}\|[\mathbf{B}^\mathsf{T}\mathbf{B}]_{j,\cdot}(\boldsymbol{\beta}^\star - \mathbf{v}^\star)\|_2, & j \in \mathcal{S}, \\ K_j - n^{-1}\|[\mathbf{B}^\mathsf{T}\mathbf{B}]_{j,\cdot}(\boldsymbol{\beta}^\star - \mathbf{v}^\star)\|_2, & j \in \mathcal{S}^c, \end{cases}$$

where $[\mathbf{B}^\mathsf{T}\mathbf{B}]_{j,\cdot} \in \mathbb{R}^{p_j \times p}$ is the submatrix of $\mathbf{B}^\mathsf{T}\mathbf{B}$ with rows corresponding to the indices of $\boldsymbol{\beta}^\star$ defining the j-th group for each $j \in [J]$. The fact that $\mathbf{v}^\star$ may not be uniquely defined is inconsequential to the definition of ν_j because $\mathbf{B}\mathbf{v}^\star$ is identical for all $\mathbf{v}^\star$ which satisfy (4.1). Finally, let $\bar{\nu} = \max_{j\in\mathcal{S}} \nu_j$ and $\underline{\nu} = \min_{k\in\mathcal{S}^c} \nu_k$. Both $\bar{\nu}$ and $\underline{\nu}$ play important roles in our error bounds. In brief, our results indicate that a good choice of $\mathbf{B}$ is one in which $\bar{\nu}$ is minimized and $\underline{\nu}$ is maximized, while also, the ν_j for $j \in \mathcal{S}$ are large and ν_k for $k \in \mathcal{S}^c$ are small. For the sake of illustration, we will derive closed form expressions for the ν_j under a particular choice of $\mathbf{B}$ in the next subsection.

Our results require a number of conditions and assumptions. Our first condition is that the submatrices of $\mathbf{X}$ satisfy a simple scaling condition [25]. Specifically, we assume that $\|\mathbf{X}_{\cdot,j}\| \leq \sqrt{n}$ for all $j \in [J]$. Such a condition was used in, for example, Corollary 4 of [25]. In the case that $p_j = 1$, this simplifies to the standard column-wise scaling condition that each column of $\mathbf{X}$ has squared Euclidean norm no greater than n (e.g., see Example 11.1 of [16]). We also assume the following:

- **A1.** (Subgaussian errors). The data are generated from (1.1) where $\boldsymbol{\epsilon} \in \mathbb{R}^n$ has independent entries which are each σ-subgaussian random variables for $0 < \sigma < \infty$. That is, $\mathbb{E}(\epsilon_i) = 0$ and for all $t \in \mathbb{R}$, $\mathbb{E}\{\exp(t\epsilon_i)\} \leq \exp(t^2\sigma^2/2)$ for each $i \in [n]$.
- **A2.** (Convexity) The matrix $\mathbf{B}$ is chosen so that $\mathbf{X}^\mathsf{T}\mathbf{X} \succeq \lambda\mathbf{B}^\mathsf{T}\mathbf{B}$.
- **A3.** (Sample size) The sample size n is sufficiently large such that there exists a constant ξ where $\underline{\nu} \geq \xi > 0$.

Assumption **A3** implicitly requires that $K_j \geq \xi > 0$ for all $j \in \mathcal{S}^c$. In finite sample settings, assumption **A3** could be replaced with an assumption on the magnitude of the K_j. Finally, we require a version of the well-known restricted

eigenvalue condition which depends on the design matrix $\mathbf{X}$ and the matrix parameter $\mathbf{B}$.

A4. (Restricted eigenvalue condition) For a fixed $c > 1$, define

$$\mathbb{C}_n(\mathcal{S}, \nu, c) = \Big\{ \boldsymbol{\Delta} \in \mathbb{R}^p : \boldsymbol{\Delta} \neq \mathbf{0}, \sum_{k \in \mathcal{S}^c} \left(\nu_k - \frac{\xi}{c} \right) \|\boldsymbol{\Delta}_k\|_2 \leq \sum_{j \in \mathcal{S}} \left(\nu_j + \frac{\xi}{c} \right) \|\boldsymbol{\Delta}_j\|_2 \Big\},$$

where we use a single notation ν to indicate the dependency on $\bar{\nu}, \underline{\nu}$, and ν_j for $j \in [J]$. We assume there exists a constant $k > 0$ such that for all n and p,

$$0 < k \leq \kappa_{\mathbf{B}}(\mathcal{S}, c) = \inf_{\boldsymbol{\Delta} \in \mathbb{C}_n(\mathcal{S}, \nu, c)} \frac{\boldsymbol{\Delta}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X} - \lambda \mathbf{B}^{\mathsf{T}}\mathbf{B})\boldsymbol{\Delta}}{2n\|\boldsymbol{\Delta}\|_2^2}.$$

We will discuss the modified restricted eigenvalue condition after stating our main result, which we prove in Appendix A.

Theorem 4.1 (Error bound for group GMC)**.** *Let $c > 1$ and $k_1 > 1$ be fixed constants. If assumptions **A1**–**A4** hold and*

$$\lambda = \frac{4c\sigma}{\xi} \left(\max_{j \in [J]} \sqrt{\frac{p_j}{n}} + \sqrt{\frac{k_1 \log(J)}{n}} \right),$$

then with probability at least $1 - 2\exp\{-(k_1 - 1)\log(J)\}$,

$$\|\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^\star\|_2 \leq \frac{4c\sigma}{\kappa_{\mathbf{B}}(\mathcal{S}, c)} \left(\frac{\bar{\nu}}{\xi} + \frac{1}{c} \right) \left\{ \left(\max_{j \in [J]} \sqrt{\frac{|\mathcal{S}| p_j}{n}} \right) + \sqrt{\frac{|\mathcal{S}| k_1 \log(J)}{n}} \right\},$$

where $\hat{\boldsymbol{\beta}}(\lambda)$ is the group GMC estimator obtained from (2.6).

Perhaps unsurprisingly, the group GMC estimator achieves the same asymptotic error rate as the group Lasso penalized least squares estimator. Where an improvement over other convex estimators could be realized is through judicious choice of $\mathbf{B}$ such that $\kappa_{\mathbf{B}}(\mathcal{S}, c)$ is large and $\bar{\nu}/\xi$ is small. We discuss this further in the next subsection.

As a consequence of our proof technique, we also establish an error bound for the original GMC estimator in [30]: this is a special case of the group GMC estimator with each group consisting of a single coefficient. This is the first error bound for the GMC estimator that we are aware of.

Theorem 4.2 (Error bound for GMC)**.** *Let $c > 1$ and $k_2 \in (0, 1/2)$ be fixed constants. Let $p_j = 1$ for $j \in [p]$ so that $\mathcal{S} = \{j : \beta_j^\star \neq 0, j \in [p]\}$. If assumptions **A1**–**A4** hold and $\lambda = (c\sigma/\xi)\sqrt{2\log(p/k_2)/n}$, then with probability at least $1 - 2k_2$,*

$$\|\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^\star\|_2 \leq \frac{c\sigma}{\kappa_{\mathbf{B}}(\mathcal{S}, c)} \left(\frac{\bar{\nu}}{\xi} + \frac{1}{c} \right) \sqrt{\frac{2|\mathcal{S}| \log(p/k_2)}{n}},$$

where $\hat{\boldsymbol{\beta}}(\lambda)$ is the corresponding GMC estimator.

Like the group-penalized version, the GMC penalized estimator achieves the same well-known asymptotic $\sqrt{|\mathcal{S}|\log p/n}$ rate as its l_1-norm penalized counterpart as long as $\bar{\nu} = O(\xi)$. However, like the group GMC penalized estimator, an improvement over Lasso in finite sample settings may be realized through an inflation of the restricted eigenvalue $\kappa_{\mathbf{B}}(\mathcal{S}, c)$, and through the role of the ν_j's.

4.2. Additional insights

The restricted eigenvalue $\kappa_{\mathbf{B}}(\mathcal{S}, c)$ in **A4** differs from the analogous condition under the l_2-norm group penalization, which may partly explain the difference in performance observed in Section 5. Specifically, to establish error bounds for the group Lasso analog of (2.6), the corresponding restricted eigenvalue condition posits a lower bound on $\inf_{\boldsymbol{\Delta}\in\mathbb{D}_n(\mathcal{S},c)} \frac{\boldsymbol{\Delta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\Delta}}{2n\|\boldsymbol{\Delta}\|_2^2}$, where

$$\mathbb{D}_n(\mathcal{S}, c) = \left\{ \boldsymbol{\Delta} \in \mathbb{R}^p : \boldsymbol{\Delta} \neq \mathbf{0}, \sum_{k\in\mathcal{S}^c} K_k\|\boldsymbol{\Delta}_k\|_2 \leq \left(\frac{c+1}{c-1}\right) \sum_{j\in\mathcal{S}} K_j\|\boldsymbol{\Delta}_j\|_2 \right\},$$

when the tuning parameter is chosen to be at least as large as $c\max_j \|\boldsymbol{\epsilon}^\mathsf{T}\mathbf{X}_{\cdot,j}\|_2/n$. The difference between restricted eigenvalue conditions comes both in terms of the function the infimum is taken with respect to and in terms of the set over which the infimum is taken. For example, when we take $\mathbf{B} = \sqrt{\alpha/\lambda}\mathbf{X}$ for $\alpha \in (0, 1)$, we see that

$$\kappa_{\mathbf{B}}(\mathcal{S}, c) = (1-\alpha)\left\{ \inf_{\boldsymbol{\Delta}\in\mathbb{C}_n(\mathcal{S},\nu,c)} \frac{\boldsymbol{\Delta}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\Delta}}{2n\|\boldsymbol{\Delta}\|_2^2} \right\},$$

which would at first glance seem to imply a $(1-\alpha)$ factor decrease in the restricted eigenvalue relative to that for the group Lasso estimator. However, the benefit comes through the potential reduction in volume of the set $\mathbb{C}_n(\mathcal{S}, \nu, c)$ relative to $\mathbb{D}_n(\mathcal{S}, c)$. If, for example, each $\nu_j + \xi/c \geq (c+1)K_j$ for $j \in \mathcal{S}$ and each $\nu_k - \xi/c \leq (c-1)K_k$ for $k \in \mathcal{S}^c$, this would guarantee the set $\mathbb{C}_n(\mathcal{S}, \nu, c)$ has volume no greater than $\mathbb{D}_n(\mathcal{S}, c)$. More generally, if ξ and many ν_j for $j \in \mathcal{S}$ are large, one may expect the reduction in volume of $\mathbb{C}_n(\mathcal{S}, \nu, c)$ relative to $\mathbb{D}_n(\mathcal{S}, c)$ to lead to a larger restricted eigenvalue and in turn, an improved error bound. In addition, since the restricted eigenvalue condition **A4** depends on the user-specified matrix $\mathbf{B}$, one may select $\mathbf{B}$ such that this condition is more plausible than the analogous condition under the group Lasso penalization. The matrix $\mathbf{B}$ also affects the error bound through the ν_j, both through the modification of $\mathbb{C}_n$ and the ratio $\bar{\nu}/\xi$.

To get a sense of how the ν_j's depend on the choice of $\mathbf{B}$, we focus on a special case.

Proposition 4.1. *Suppose $n > p$ and $\mathbf{X}^\mathsf{T}\mathbf{X} \succ \mathbf{O}_p$. Consider the choice of $\mathbf{B} = \sqrt{\eta/\lambda}\mathbf{I}_p$ where $\eta > 0$ is the smallest eigenvalue of $\mathbf{X}^\mathsf{T}\mathbf{X}$ so that **A2** holds.*

In this situation,

$$\mathbf{v}_j^\star = \left(1 - \frac{nK_j\lambda}{\eta\|\boldsymbol{\beta}_j^\star\|_2}\right)_+ \boldsymbol{\beta}_j^\star, \qquad j \in \mathcal{S},$$

and $\mathbf{v}_j^\star = 0$ *for* $j \in \mathcal{S}^c$. *It thus follows that for* $j \in [J]$,

$$\nu_j = K_j + K_j \mathbb{1}\{\|\boldsymbol{\beta}_j^\star\|_2 > nK_j\eta/\lambda\} + \frac{\eta}{n\lambda}\|\boldsymbol{\beta}_j^\star\|_2 \mathbb{1}\{\|\boldsymbol{\beta}_j^\star\|_2 \le nK_j\eta/\lambda\},$$

where $\mathbb{1}(\cdot)$ *is the indicator function.*

In addition to providing insight regarding the ν_j, in light of Proposition 2.2, this result provides a new lens through which the group MCP can be viewed in the settings in which group MCP and group GMC are equivalent. Existing theory for group MCP is largely concerned with the oracle property rather than, say, error bounds.

Crucially, Proposition 4.1 implies $\nu_j = K_j$ for all $j \in \mathcal{S}^c$ in the considered scenario. This choice of $\mathbf{B}$ is useful because $\nu_j = K_j$ for $j \in \mathcal{S}^c$ (whereas alternative choices of $\mathbf{B}$ will yield $\nu_j < K_j$ for some $j \in \mathcal{S}^c$), and because it enables us to express $\mathbf{v}^\star$ explicitly in terms of $\boldsymbol{\beta}^\star$. In this setting of Proposition 4.1, if each $K_j = 1$ and the $\|\boldsymbol{\beta}_j^\star\|_2$ are sufficiently large for $j \in \mathcal{S}$, then $(\nu_j + \xi/c) = (2 + 1/c)$ for $j \in \mathcal{S}$ and $(\nu_k - \xi/c) = (c-1)/c$ for $k \in \mathcal{S}^c$ so that $\mathbb{C}_n(\mathcal{S}, \nu, c)$ could be written in the same form as $\mathbb{D}_n(\mathcal{S}, c)$ with $(c+1)/(c-1)$ replaced with $(2c+1)/(c-1)$ (i.e., $\mathbb{C}_n(\mathcal{S}, \nu, c)$ has less volume than $\mathbb{D}_n(\mathcal{S}, c)$).

In practice, of course, the ν_j cannot be computed since they depend on $\boldsymbol{\beta}^\star$. Likewise, the $\mathbf{B}$ which minimizes the bounds in Theorem 4.1 also depends on $\boldsymbol{\beta}^\star$ and the set $\mathcal{S}$, so this cannot be used in practice.

5. Simulation studies

We investigate the practical performance of the proposed group GMC method with experiments that build upon the simulation scenarios in [39]. We also compare the group GMC with the group Lasso, group MCP, and group SCAD. The computation of the three existing methods is done using the R package `grpreg` developed by [6], and the hyper-parameters in the group MCP and group SCAD are set as the default values given in their R package. The implementation of the proposed group GMC method can be found in the R package `GMC`, which is available at https://github.com/Xiaoqian-Liu/GMC.

There are four linear regression models considered in the simulation study in [39]. In this work, we consider the two most complicated ones, an ANOVA model with all two-way interactions and an additive model with both categorical and continuous variables. More importantly, we study different cases for each model to explore the effects of interesting factors, including the signal-to-noise ratio (SNR), the correlation among groups, the problem dimension, and the convexity-preserving parameter (only for the group GMC). In each case, we run the experiment for 100 replications and evaluate different methods with

respect to: (i) mean squared error (MSE) of the estimated coefficients; (ii) prediction error defined as $\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^\star\|_2^2/n$ where n is the sample size, $\mathbf{X}$ is the design matrix, and $\boldsymbol{\beta}^\star$ and $\hat{\boldsymbol{\beta}}$ are the vectors of true and estimated coefficients respectively; (iii) support recovery with respect to F1 score, number of true positives (TP), and number of false positives (FP). The F1 score is a metric of support recovery performance that accounts for both TP and FP; it is defined as $\mathrm{F1} = 2\mathrm{TP}/(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$ where FN denotes the number of false negatives. The F1 score takes on values between 0 and 1, and a higher value indicates better support recovery. Regarding the selection of the tuning parameter λ, we use five-fold cross-validation for all methods. The matrix parameter $\mathbf{B}$ for the group GMC method is set according to (2.9). Namely, given a convexity-preserving parameter α, $\mathbf{B}$ varies with λ, but we always have $\lambda\mathbf{B}^\mathsf{T}\mathbf{B} = \alpha\mathbf{X}^\mathsf{T}\mathbf{X}$ so that the convexity degree of the optimization problem is fixed. We include the study of the ANOVA model in this section and relegate the investigation of the additive model to Appendix C. We also report run times of different methods for all the simulation experiments in Appendix C.

The data generation process of the ANOVA model is as follows. Four categorical variables Z_1, Z_2, Z_3 and Z_4 are generated from a centered multivariate normal distribution with covariance between Z_i and Z_j being $\rho^{|i-j|}$ for $i, j = 1, \ldots, 4$. Then each Z_i is trichotomized to $0, 1$ or 2 if it is smaller than $\Phi^{-1}(\frac{1}{3})$, larger than $\Phi^{-1}(\frac{2}{3})$ or in between, where Φ is the cumulative density function (CDF) of the standard normal distribution. The true regression model is

$$\begin{aligned} y = {} & 3\mathbb{1}(Z_1 = 1) + 2\mathbb{1}(Z_1 = 0) + 3\mathbb{1}(Z_2 = 1) + 2\mathbb{1}(Z_2 = 0) + \mathbb{1}(Z_1 = 1, Z_2 = 1) \\ & + 1.5\mathbb{1}(Z_1 = 1, Z_2 = 0) + 2\mathbb{1}(Z_1 = 0, Z_2 = 1) + 2.5\mathbb{1}(Z_1 = 0, Z_2 = 0) + \epsilon, \end{aligned} \tag{5.1}$$

where ϵ is normally distributed with mean zero and variance σ^2. Therefore, we have 32 covariate variables from ten groups, where four of them with a group size of two represent the main effects and the other six groups with a group size of four indicate the two-way interactions. The response variable, however, is only related to three groups of covariates as shown in model (5.1). We next consider three different cases to explore the possible effects of interesting factors.

Case C1: The first factor we are interested in is the SNR, which is defined as $\|\mathbf{X}\boldsymbol{\beta}^\star\|_2/(\sqrt{n}\sigma)$. We consider uncorrelated groups, namely $\rho = 0$ in the data generation process. We set a sequence of σ so that the SNR ranges from 1 to 5. The sample size n is fixed as 100 for each setting. To better understand how the convexity-preserving parameter α affects the performance of the proposed group GMC method, we report the results of the group GMC with $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1\}$. These results also provide guidance on how to set α for the group GMC in practice.

Figure 1 presents the impact of SNR on the performance of the four methods for model (5.1). As expected, the MSE of the estimated coefficients and the prediction error decrease as the SNR increases for all methods. The group Lasso

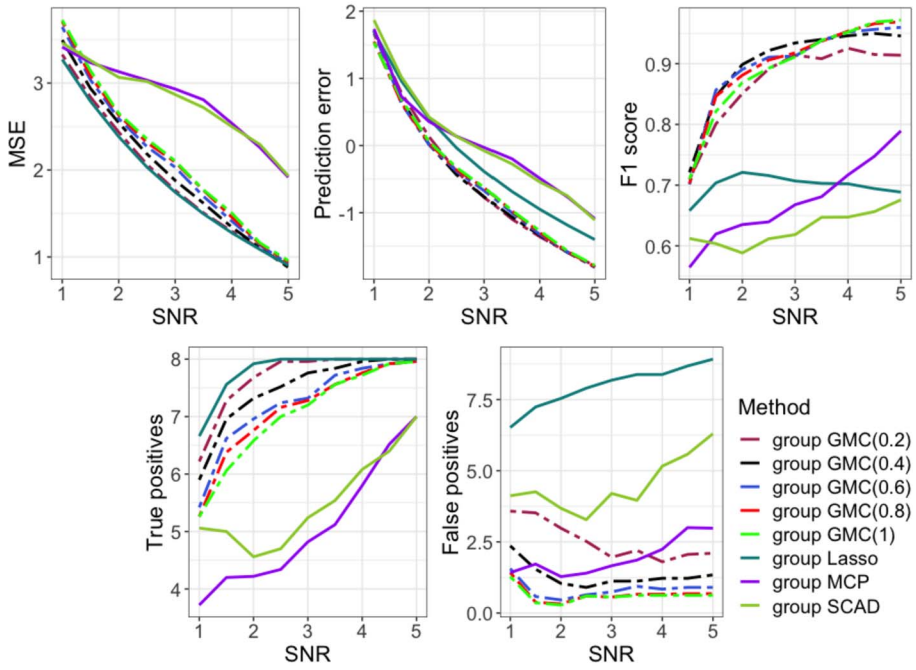

FIG 1. *Results for Case C1: Impact of SNR. Group GMC(·) stands for the group GMC with a specific value of* α. *Average performance based on* 100 *simulation replicates for each method. MSE and Prediction error are on a log scale.*

achieves the lowest MSE, while the group GMC gives the lowest prediction error among the four methods. The convexity-preserving parameter α does not show a significant effect on the prediction performance of the group GMC, but it has a mild effect on the coefficient estimation. As indicated in the top left panel of Figure 1, a smaller value of α leads to a lower MSE. When it comes to support recovery, the group GMC shows a distinct advantage over the other three methods. It achieves a higher F1 score than existing methods in all SNR settings. The two plots on the bottom panel of Figure 1 display the variable selection results of different methods in detail. The group Lasso obtains the most true positives but also the most false positives. Both group SCAD and group MCP miss some true positives and also include some irrelevant variables into the ANOVA model. The group GMC, however, can achieve a number of true positives comparable with the group Lasso while maintaining its number of false positives at a very low level. The convexity-preserving parameter α indeed affects the variable selection performance of the group GMC. Both numbers of true and false positives decrease as the value of α increases. In other words, a large value of α in the group GMC results in a sparse model. In general, a range of $0.4 < \alpha < 1$ works well for this ANOVA example.

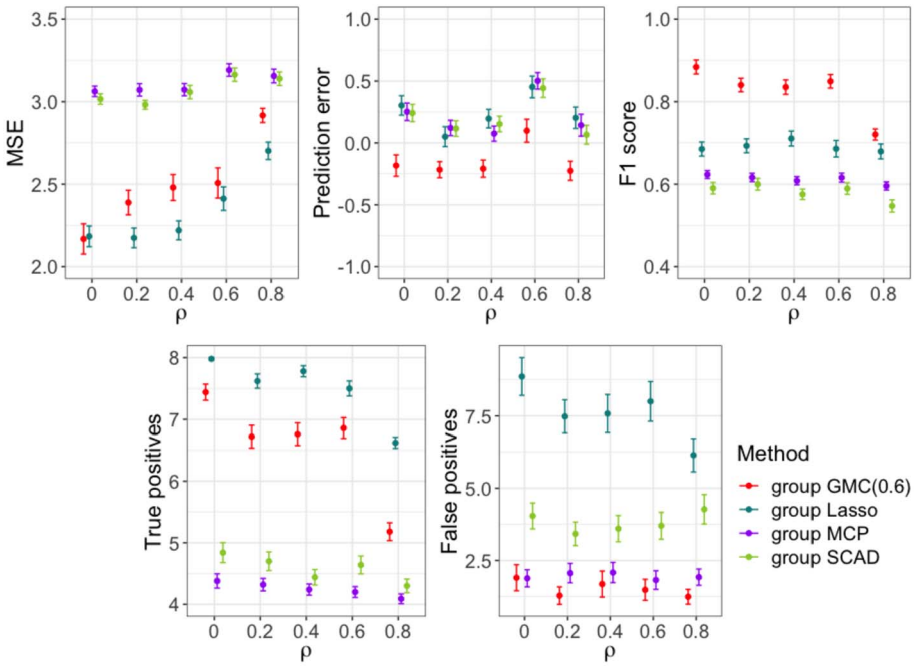

FIG 2. *Results for Case C2: Impact of group correlation. Average performance plus/minus one standard error based on* 100 *simulation replicates for each method. MSE and Prediction error are on a log scale.*

Case C2: When comparing different grouped variable selection methods, one factor of interest is to what extent the correlation among groups impacts their performance. For that purpose, we set a grid of values for ρ, $\rho = \{0, 0.2, 0.4, 0.6, 0.8\}$, so that the correlation between Z_i and Z_j is $\rho^{|i-j|}$ for $i \neq j$. We fix the SNR of the regression model to be 2 and the sample size to be 100 for each run. We set the convexity-preserving parameter $\alpha = 0.6$ for the group GMC method.

Figure 2 shows the performance of the four methods under different group correlations. Both the group GMC and group Lasso produce worse estimation as the correlation ρ increases, while the group MCP and group SCAD fail to achieve comparable estimation even in the uncorrelated setting. For the model prediction, all four methods are relatively stable across different correlation settings, and the group GMC compares favorably with the other three. Regarding the variable selection with respect to the F1 score, group GMC visibly outperforms the existing three methods. All methods see a drop in F1 score when the correlation ρ reaches up to 0.8. The plots of true and false positives provide detailed insight into the variable selection performance of different methods. The group Lasso includes the most false positives into the model, although it leads others in the inclusion of true positives. In contrast, the group MCP and group

SCAD build much sparser models and miss some true positives. The group GMC is capable of obtaining true positives comparable with the group Lasso and excluding those irrelevant variables from the regression model. When $\rho = 0.8$, all methods suffer a drop in their true positives, resulting in the drop in their F1 scores as seen in the corresponding plot.

Case C3: In this third case, our goal is to explore the impact of the problem dimension on the performance of different methods. To that end, we set three different dimension settings where four, ten, and sixteen independent categorical variables Z_i are generated accordingly in each setting and then trichotomized in the same way as described above. As a result, the problem dimension p is 32, 200, and 512, respectively. But the response variable y remains generated according to model (5.1) with SNR = 2, namely the number of true positives is eight in all dimension settings. The sample size is 100 for each setting. We again fix the convexity-preserving parameter α of the group GMC as 0.6.

Figure 3 summarizes the simulation results. In terms of the coefficient estimation, the group MCP and group SCAD behave quite similarly and worse than the group Lasso and group GMC. Regarding the model prediction, the group

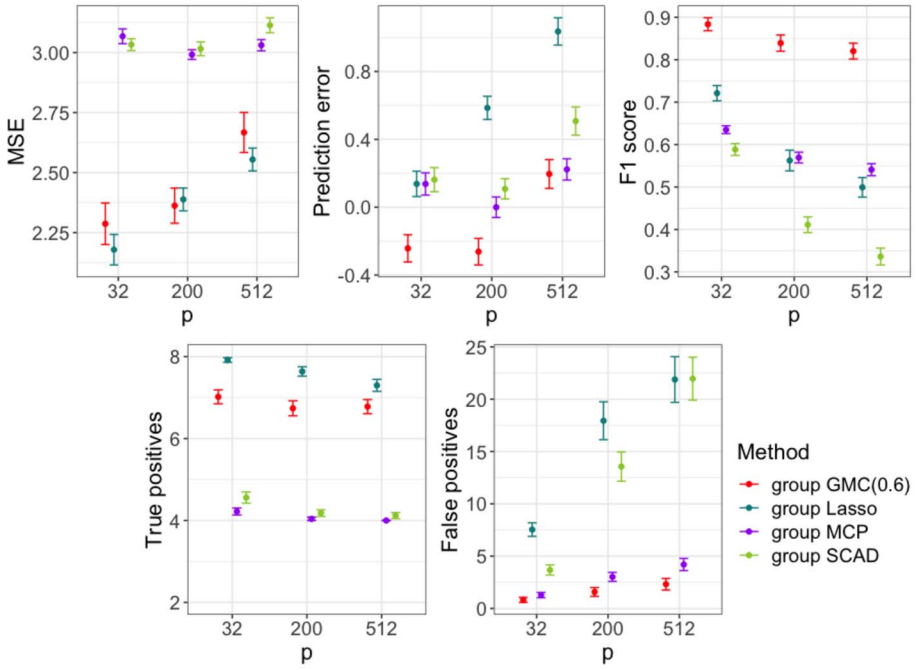

FIG 3. *Results for Case C3: Impact of problem dimension. Average performance plus/minus one standard error based on* 100 *simulation replicates for each method. MSE and Prediction error are on a log scale.*

TABLE 1
Description of the birth weight data set.

Name	Type	Variable description
Birth weight	Continuous	Infant birth weight in kilograms
Mother's age	Continuous	Mother's age in years
Mother's weight	Continuous	Mother's weight in pounds at last menstrual period
Race	Categorical	Mother's race (white, black or other)
Smoking	Categorical	Smoking status during pregnancy (yes or no)
# Premature	Categorical	Previous premature labors (0, 1, or more)
Hypertension	Categorical	History of hypertension (yes or no)
Uterine irritability	Categorical	Presence of uterine irritability (yes or no)
# Phys. visits	Categorical	Number of physician visits during the first trimester $(0, 1, 2,$ or more)

GMC fares well in all dimension settings. With respect to variable selection, the group GMC exhibits a distinct advantage over the existing three methods across different problem dimensions thanks to its robust behavior with respect to both true and false positives. While group MCP performs well at excluding false positives, it errs on the side of being too conservative and misses some true positives. The group Lasso and group SCAD, however, select too many irrelevant variables into the regression model, especially for the high-dimensional scenarios.

6. Real data application

We apply our group GMC method on the birth weight data set from [17], which studies risk factors associated with low infant birth weight. The data set is publicly available in the R package `MASS` and contains 189 observations of one response variable (infant birth weight) and eight explanatory variables from the mother, including both continuous and categorical factors. We include detailed description of the data set in Table 1. As with [39], we take into account the preliminary analysis that both mother's age and weight have non-linear effects on the birth weight. Therefore, we model these two effects by third-order polynomials. Finally, we get sixteen predictors from eight groups to fit a linear regression model.

Following our simulation studies, we analyze the data using the proposed group GMC as well as the group Lasso, group MCP, and group SCAD. For group GMC, we again set the matrix parameter $\mathbf{B}$ according to (2.9) and choose $\alpha = 0.8$ based on our experience from the simulation studies. For evaluation, we first randomly sample three-quarters of the observations (142 cases) as a training set for selecting the tuning parameter λ by ten-fold cross-validation. Then we use the obtained tuning parameter to fit the full data to get the estimated coefficients. Finally, we compute the prediction error based on the testing set of the remaining one-quarter records.

TABLE 2
Summarized results for the birth weight data.

Method	Prediction error	# nonzero groups	Excluded groups
Group Lasso	0.36	8	none
Group SCAD	0.35	8	none
Group MCP	0.35	7	# Phys. visits
Group GMC	0.35	7	# Phys. visits

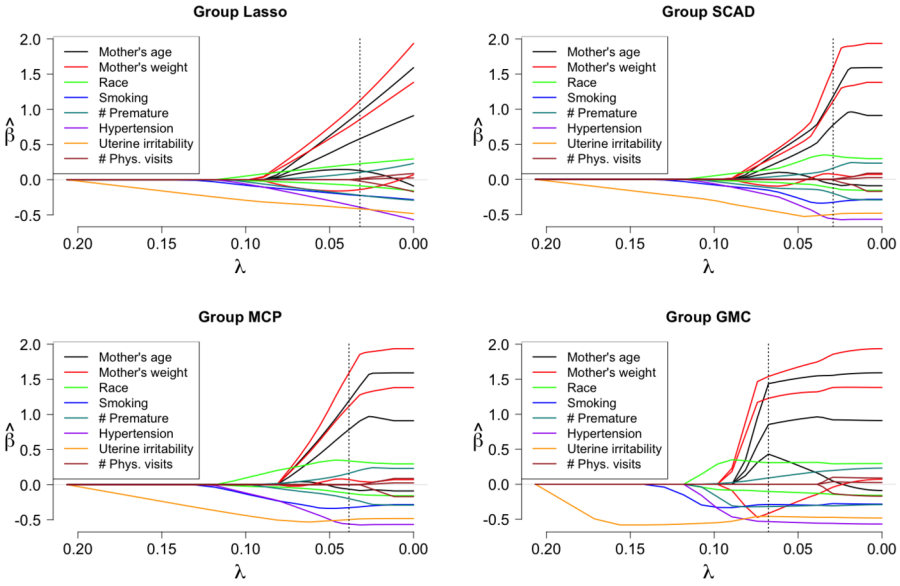

FIG 4. *Solution paths of four different methods on the birth weight data. The dotted vertical line in each subfigure indicates the selected λ via ten-fold cross-validation.*

Table 2 summarizes the prediction errors, numbers of nonzero groups, and excluded groups from the four methods. The group Lasso and group SCAD fail to exclude any group from the model. Both group MCP and group GMC, however, regard the number of physician visits during the first trimester as an unimportant factor to the infant birth weight. The prediction errors obtained from the four different methods are comparable.

It seems that, for this birth weight data analysis, the group GMC does not exhibit any advantage over existing methods. Nevertheless, the solution paths from the four methods, as shown in Figure 4, tell a different story. The estimated coefficients from the group GMC, as indicated by the vertical dotted line, undergo noticeably less shrinkage than those of the group Lasso and are similar to the estimates from the group MCP and group SCAD. This confirms the unbiased (or at least less biased) estimation of the group GMC as a nonconvex penalization method. What is more, the group GMC method is more robust

against the tuning parameter selection compared to the other three methods. It is anticipated that the estimated coefficients are increasingly shrunk as λ increases, and thus fewer variables are selected into the regression model. But as shown in Figure 4, the estimated coefficients and selected variables are stable over $\lambda \in [0.04, 0.07]$ for the group GMC, while the other three methods do not have as comparably a wide range of λ. This insensitivity furnishes some evidence that the group GMC method can potentially blunt estimation bias successfully while still simultaneously achieving satisfactory variable selection.

7. Discussion

In this paper, we used the convex-nonconvex strategy to propose a novel concave penalty, called the group GMC, for grouped variable selection and coefficient estimation in linear regression. The group GMC penalty is a variant of the GMC penalty and thus inherits its characteristic that it is able to maintain the convexity of the corresponding optimization problem. Therefore, the group GMC eliminates the possibility of suboptimal local minima while maintaining unbiased estimation as a nonconvex penalization approach. We formulated linear regression with the group GMC penalization as a convex optimization problem, or more specifically a saddle-point problem when a certain condition is satisfied. The resulting group GMC estimator enjoys desirable properties which help accelerate numerical computation and tuning parameter selection. We presented efficient algorithms for computing the group GMC estimator. Additionally, we analyzed statistical properties of the group GMC estimator as well as the original GMC estimator. Our results are the first to establish the l_2-norm error bounds for the GMC least squares estimators and as such, provide novel insights about the performance of convex nonconvex penalization. In our simulation study, we compared the practical performance of the group GMC with the group Lasso, group MCP, and group SCAD via a comprehensive evaluation, including variable selection, coefficient estimation, and model prediction. Through a battery of simulation experiments, we found that the group GMC can achieve better or at least competitive performance in comparison with the existing three methods under different scenarios such as different SNRs, correlated or uncorrelated groups, and different dimension settings. A real data application displays the advantage of the group GMC method in its robustness in unbiased coefficient estimation and grouped variable selection.

While this paper was under review, we became aware of the work by [8]. The key contribution of their work was a convexity-preserving algebraic design for the parameter matrix $\mathbf{B}$ for a wide class of convex-nonconvex regularizers composed with linear mappings lacking full row rank. As an illustration, the authors applied their model to a group-sparse least squares estimation problem, which is essentially the same as the group GMC problem. Nonetheless, there are notable differences between their investigation of convex-nonconvex group sparsity and the one in this paper. Their illustrative application considered a special case of group structure, specifically consecutive groupings. In contrast,

we put no restriction on the group structure and conducted a wider study of the empirical behavior of group GMC under various simulation scenarios. Moreover, we presented novel analysis of the statistical properties of the group GMC and original GMC estimators, which as far as we are aware have been unexplored to date in the convex-nonconvex literature.

Several related studies can be done in the future. First of all, how to set the matrix parameter $\mathbf{B}$ warrants more exploration and investigation. We discussed how $\mathbf{B}$ could affect the error bound of the group GMC estimator in Section 4 but anticipate that other approaches to set $\mathbf{B}$ could further improve the performance of the group GMC, both theoretically and practically. Second, the group GMC method could be extended to generalized linear models to deal with grouped variable selection problems in other high-dimensional cases. More generally, the convex-nonconvex strategy could be applied to other sparse learning scenarios so that one can enjoy the advantages of convex optimization and nonconvex penalization simultaneously.

Appendix A: Proofs

A.1. Proof of Proposition 2.1

We rewrite $F(\boldsymbol{\beta})$ as

$$\begin{aligned}
F(\boldsymbol{\beta}) &= \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\Phi_{\mathbf{B}}(\boldsymbol{\beta}) \\
&= \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}K_j\|\boldsymbol{\beta}_j\|_2-\lambda\min_{\mathbf{v}\in\mathbb{R}^p}g(\boldsymbol{\beta},\mathbf{v}) \\
&= \max_{\mathbf{v}\in\mathbb{R}^p}\left\{\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\sum_{j=1}^{J}K_j\|\boldsymbol{\beta}_j\|_2-\lambda g(\boldsymbol{\beta},\mathbf{v})\right\} \\
&= \max_{\mathbf{v}\in\mathbb{R}^p}\left\{\frac{1}{2n}\boldsymbol{\beta}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X}-\lambda\mathbf{B}^{\mathsf{T}}\mathbf{B})\boldsymbol{\beta}+\lambda\sum_{j=1}^{J}K_j\|\boldsymbol{\beta}_j\|_2+G(\boldsymbol{\beta},\mathbf{v})\right\} \\
&= \frac{1}{2n}\boldsymbol{\beta}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X}-\lambda\mathbf{B}^{\mathsf{T}}\mathbf{B})\boldsymbol{\beta}+\lambda\sum_{j=1}^{J}K_j\|\boldsymbol{\beta}_j\|_2+\max_{\mathbf{v}\in\mathbb{R}^p}G(\boldsymbol{\beta},\mathbf{v}),
\end{aligned}$$

where $g(\boldsymbol{\beta},\mathbf{v})=\sum_{j=1}^{J}K_j\|\mathbf{v}_j\|_2+\frac{1}{2n}\|\mathbf{B}(\boldsymbol{\beta}-\mathbf{v})\|_2^2$, and

$$G(\boldsymbol{\beta},\mathbf{v})=\frac{\lambda}{n}\mathbf{v}^{\mathsf{T}}(\mathbf{B}^{\mathsf{T}}\mathbf{B})\boldsymbol{\beta}-\frac{1}{n}\mathbf{y}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}-\frac{\lambda}{2n}\|\mathbf{B}\mathbf{v}\|_2^2-\lambda\sum_{j=1}^{J}K_j\|\mathbf{v}_j\|_2+\frac{1}{2n}\|\mathbf{y}\|_2^2$$

is affine in $\boldsymbol{\beta}$. Then $\max_{\mathbf{v}\in\mathbb{R}^p}G(\boldsymbol{\beta},\mathbf{v})$ is convex since it is the pointwise maximum of a set of convex functions. Therefore, if $\mathbf{X}^{\mathsf{T}}\mathbf{X}\succeq\lambda\mathbf{B}^{\mathsf{T}}\mathbf{B}$, F is convex. □

A.2. Proof of Proposition 2.2

When $\lambda = 0$, both the group GMC and group MCP reduce to the ordinary least squares problem. In the following proof, we assume that $\lambda > 0$.

Recall that the group MCP penalty is expressed as

$$P_{\lambda,\gamma}(\boldsymbol{\beta}) = \sum_{j=1}^{J} \rho_{K_j\lambda,\gamma}(\|\boldsymbol{\beta}_j\|_2),$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \ldots, \boldsymbol{\beta}_J^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^p$ has the group structure and $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$ is the vector of components in the j-th group. By the definition of the univariate MCP function (2.4),

$$\begin{aligned}\rho_{K_j\lambda,\gamma}(\|\boldsymbol{\beta}_j\|_2) &= \begin{cases} K_j\lambda\|\boldsymbol{\beta}_j\|_2 - \frac{\|\boldsymbol{\beta}_j\|_2^2}{2\gamma}, & \text{if } \|\boldsymbol{\beta}_j\|_2 \le K_j\gamma\lambda, \\ \frac{1}{2}\gamma K_j^2\lambda^2, & \text{if } \|\boldsymbol{\beta}_j\|_2 > K_j\gamma\lambda, \end{cases} \\ &= \begin{cases} K_j\lambda\|\boldsymbol{\beta}_j\|_2 - \frac{K_j b_j^2\lambda}{2}\|\boldsymbol{\beta}_j\|_2^2, & \text{if } \|\boldsymbol{\beta}_j\|_2 \le \frac{1}{b_j^2}, \\ \frac{1}{2b_j^2}K_j\lambda, & \text{if } \|\boldsymbol{\beta}_j\|_2 > \frac{1}{b_j^2}, \end{cases}\end{aligned}$$

with $b_j^2 = 1/(K_j\lambda\gamma)$.

Suppose that $\mathbf{B}^{\mathsf{T}}\mathbf{B} = (n/\lambda\gamma)\mathbf{I}_p$, then we have

$$\begin{aligned}\lambda\Phi_{\mathbf{B}}(\boldsymbol{\beta}) &= \lambda\sum_{j=1}^{J} K_j\|\boldsymbol{\beta}_j\|_2 - \min_{\mathbf{v}\in\mathbb{R}^p}\left\{\sum_{j=1}^{J} K_j\|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}(\boldsymbol{\beta}-\mathbf{v})\|_2^2\right\} \\ &= \lambda\sum_{j=1}^{J} K_j\|\boldsymbol{\beta}_j\|_2 - \min_{\mathbf{v}\in\mathbb{R}^p}\left\{\sum_{j=1}^{J} K_j\|\mathbf{v}_j\|_2 + \frac{b_j^2K_j}{2}\sum_{j=1}^{J}\|\boldsymbol{\beta}_j-\mathbf{v}_j\|_2^2\right\} \\ &= \lambda\sum_{j=1}^{J}\left(K_j\|\boldsymbol{\beta}_j\|_2 - \min_{\mathbf{v}_j\in\mathbb{R}^{p_j}}\left\{K_j\|\mathbf{v}_j\|_2 + \frac{b_j^2K_j}{2}\|\boldsymbol{\beta}_j-\mathbf{v}_j\|_2^2\right\}\right) \\ &= \lambda\sum_{j=1}^{J} K_j\left(\|\boldsymbol{\beta}_j\|_2 - \min_{\mathbf{v}_j\in\mathbb{R}^{p_j}}\left\{\|\mathbf{v}_j\|_2 + \frac{b_j^2}{2}\|\boldsymbol{\beta}_j-\mathbf{v}_j\|_2^2\right\}\right).\end{aligned}$$

We next show that under this special design of $\mathbf{B}^{\mathsf{T}}\mathbf{B}$, the group GMC method is equivalent to the group MCP method, namely,

$$\lambda\Phi_{\mathbf{B}}(\boldsymbol{\beta}) = P_{\lambda,\gamma}(\boldsymbol{\beta}).$$

We only need to prove that

$$\|\boldsymbol{\beta}_j\|_2 - \min_{\mathbf{v}_j\in\mathbb{R}^{p_j}}\left\{\|\mathbf{v}_j\|_2 + \frac{b_j^2}{2}\|\boldsymbol{\beta}_j-\mathbf{v}_j\|_2^2\right\} = \begin{cases} \|\boldsymbol{\beta}_j\|_2 - \frac{b_j^2}{2}\|\boldsymbol{\beta}_j\|_2^2, & \text{if } \|\boldsymbol{\beta}_j\|_2 \le \frac{1}{b_j^2}, \\ \frac{1}{2b_j^2}, & \text{if } \|\boldsymbol{\beta}_j\|_2 > \frac{1}{b_j^2}, \end{cases}$$

or equivalently,

$$\min_{\mathbf{v}_j\in\mathbb{R}^{p_j}}\left\{\|\mathbf{v}_j\|_2+\frac{b_j^2}{2}\|\boldsymbol{\beta}_j-\mathbf{v}_j\|_2^2\right\}=\begin{cases}\frac{b_j^2}{2}\|\boldsymbol{\beta}_j\|_2^2, & \text{if } \|\boldsymbol{\beta}_j\|_2\leq\frac{1}{b_j^2},\\ \|\boldsymbol{\beta}_j\|_2-\frac{1}{2b_j^2}, & \text{if } \|\boldsymbol{\beta}_j\|_2>\frac{1}{b_j^2}.\end{cases}\tag{A.1}$$

It is straightforward to see that the solution of the optimization problem on the left hand side of (A.1) is the proximal operator of the l_2 norm. Denote $\mathbf{v}_j^\star$ as the solution, then

$$\mathbf{v}_j^\star=\text{prox}_{\|\cdot\|_2/b_j^2}(\boldsymbol{\beta}_j)=\begin{cases}\mathbf{0}, & \text{if } \|\boldsymbol{\beta}_j\|_2\leq\frac{1}{b_j^2},\\ (1-\frac{1}{b_j^2\|\boldsymbol{\beta}_j\|_2})\boldsymbol{\beta}_j, & \text{if } \|\boldsymbol{\beta}_j\|_2>\frac{1}{b_j^2}.\end{cases}$$

Plugging $\mathbf{v}_j^\star$ into the left hand side of (A.1), we proved that the equation (A.1) holds.

It is trivial to see that the equivalence also holds for the GMC and MCP by taking $\boldsymbol{\beta}_j$ as each single component of $\boldsymbol{\beta}$ and $K_j=1$. □

A.3. Proof of Theorem 2.1

Recall that a coercive and strictly convex function has a unique minimizer in its domain. Therefore, to show the existence and uniqueness of the solution path to (2.6), we only need to show $F(\boldsymbol{\beta})$ in (2.7) is strictly convex and coercive.

According to the proof of proposition 2.1, it is easy to verify that $F(\boldsymbol{\beta})$ is strictly convex if $\mathbf{X}^\mathsf{T}\mathbf{X}\succ\lambda\mathbf{B}^\mathsf{T}\mathbf{B}$. Besides, we can see that $\lim_{\|\boldsymbol{\beta}\|_2\to\infty}F(\boldsymbol{\beta})=+\infty$, thus $F(\boldsymbol{\beta})$ is coercive.

We next show that the solution path $\hat{\boldsymbol{\beta}}(\lambda)$ is continuous in λ. Let us first rewrite F as

$$F(\boldsymbol{\beta},\lambda)=\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2+\lambda\left\{\sum_{j=1}^J K_j\|\boldsymbol{\beta}_j\|_2-f_1(\boldsymbol{\beta})\right\},$$

where $f_1(\boldsymbol{\beta})=\min_{\mathbf{v}\in\mathbb{R}^p}\left(\sum_{j=1}^J K_j\|\mathbf{v}_j\|_2+\frac{1}{2n}\|\mathbf{B}(\boldsymbol{\beta}-\mathbf{v})\|_2^2\right)$. Then f_1 is a convex function of $\boldsymbol{\beta}$ and its domain is $\mathbb{R}^p$. Therefore, f_1 is continuous in $\boldsymbol{\beta}$, thus jointly continuous in $(\boldsymbol{\beta},\lambda)$. So for all $(\boldsymbol{\beta},\lambda)\in(\mathbb{R}^p,\mathbb{R}_+)$, F is a jointly continuous function of $(\boldsymbol{\beta},\lambda)$. Besides,

$$\begin{aligned}&\sum_{j=1}^J K_j\|\boldsymbol{\beta}_j\|_2-\min_{\mathbf{v}\in\mathbb{R}^p}\left(\sum_{j=1}^J K_j\|\mathbf{v}_j\|_2+\frac{1}{2n}\|\mathbf{B}(\boldsymbol{\beta}-\mathbf{v})\|_2^2\right)\\ \geq&\sum_{j=1}^J K_j\|\boldsymbol{\beta}_j\|_2-\left(\sum_{j=1}^J K_j\|\boldsymbol{\beta}_j\|_2+\frac{1}{2n}\|\mathbf{B}(\boldsymbol{\beta}-\boldsymbol{\beta})\|_2^2\right)\\ =&\,0.\end{aligned}$$

So F is non-decreasing in λ. Therefore, for an arbitrary subinterval $[a,b] \subseteq (0,+\infty)$ and for all $\tilde{\lambda} \in [a,b]$,

$$F(\hat{\boldsymbol{\beta}}(\tilde{\lambda}),a) \leq F(\hat{\boldsymbol{\beta}}(\tilde{\lambda}),\tilde{\lambda}) \leq F(\mathbf{0}_p,\tilde{\lambda}) \leq F(\mathbf{0}_p,b),$$

where $\mathbf{0}_p$ is the zero vector in $\mathbb{R}^p$ and $\hat{\boldsymbol{\beta}}(\tilde{\lambda})$ is the unique minimizer of $F(\cdot,\tilde{\lambda})$. Since $F(\boldsymbol{\beta},a)$ is coercive in $\boldsymbol{\beta}$, we have

$$S = \{\boldsymbol{\beta} : F(\boldsymbol{\beta},a) \leq F(\mathbf{0}_p,b)\} \text{ is a compact set.}$$

Hence, $\hat{\boldsymbol{\beta}}(\tilde{\lambda}) \in S$ for all $\tilde{\lambda} \in [a,b]$. Suppose that $\hat{\boldsymbol{\beta}}(\lambda)$ is not continuous at some point $\tilde{\lambda}$ and choose a,b such that $\tilde{\lambda} \in [a,b]$, then there exists some $\epsilon_0 > 0$ and a sequence $\{\lambda_n\}_{n\in\mathbb{N}} \in [a,b]$ such that

$$\lambda_n \to \tilde{\lambda} \text{ but } \|\hat{\boldsymbol{\beta}}(\lambda_n) - \hat{\boldsymbol{\beta}}(\tilde{\lambda})\|_2 \geq \epsilon_0 \text{ for all } n \in \mathbb{N}.$$

We have seen that for each n, $\hat{\boldsymbol{\beta}}(\lambda_n) \in S$ and S is compact, thus $\hat{\boldsymbol{\beta}}(\lambda_n)$ is a bounded sequence. Therefore, there exists a subsequence $\hat{\boldsymbol{\beta}}(\lambda_{n_k})$ such that

$$\hat{\boldsymbol{\beta}}(\lambda_{n_k}) \to \boldsymbol{\beta}^* \in S,$$

and

$$F(\hat{\boldsymbol{\beta}}(\lambda_{n_k}),\lambda_{n_k}) \leq F(\hat{\boldsymbol{\beta}}(\tilde{\lambda}),\lambda_{n_k}). \tag{A.2}$$

By the joint continuity of F and taking limit on both sides of (A.2), we get

$$F(\boldsymbol{\beta}^*,\tilde{\lambda}) \leq F(\hat{\boldsymbol{\beta}}(\tilde{\lambda}),\tilde{\lambda}).$$

By the uniqueness of the globalm minimizer, $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}(\tilde{\lambda})$, which contradicts the fact that $\|\hat{\boldsymbol{\beta}}(\lambda_n) - \hat{\boldsymbol{\beta}}(\tilde{\lambda})\|_2 \geq \epsilon_0$ for all $n \in \mathbb{N}$. Therefore, the solution path $\hat{\boldsymbol{\beta}}(\lambda)$ is continuous in λ. □

A.4. Proof of Theorem 2.2

A point $\hat{\boldsymbol{\beta}}$ furnishes a global minimum of the convex function $F(\boldsymbol{\beta})$ if and only if all forward directional derivatives $d_\theta F(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}$ are nonnegative. Let $a(\boldsymbol{\beta}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ and $b(\boldsymbol{\beta}) = \min_{\mathbf{v}\in\mathbb{R}^p} \sum_{j=1}^J K_j\|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}(\boldsymbol{\beta}-\mathbf{v})\|_2^2$, then

$$\begin{aligned} d_{\boldsymbol{\theta}} F(\hat{\boldsymbol{\beta}}) &= \lim_{h\to 0} \frac{F(\hat{\boldsymbol{\beta}} + h\boldsymbol{\theta}) - F(\hat{\boldsymbol{\beta}})}{h} \\ &= \left\langle \nabla a(\hat{\boldsymbol{\beta}}), \boldsymbol{\theta} \right\rangle + \lambda \lim_{h\to 0} \frac{\sum_{j=1}^J K_j(\|\hat{\boldsymbol{\beta}}_j + h\boldsymbol{\theta}_j\|_2 - \|\hat{\boldsymbol{\beta}}_j\|_2)}{h} - \lambda d_{\boldsymbol{\theta}} b(\hat{\boldsymbol{\beta}}). \end{aligned}$$

To prove Theorem 2.2, we need to show that $d_{\boldsymbol{\theta}} F(\mathbf{0}_p) \geq 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$ when $\lambda \geq \lambda_0$.

First, we show that

$$d_{\boldsymbol{\theta}} b(\mathbf{0}_p) = 0.$$

Note that

$$d_{\boldsymbol{\theta}} b(\mathbf{0}_p) = \lim_{h\to 0} \frac{b(h\boldsymbol{\theta}) - b(\mathbf{0}_p)}{h} = \lim_{h\to 0} \frac{b(h\boldsymbol{\theta})}{h}.$$

We define $l(\boldsymbol{\beta}, \mathbf{v}) = \sum_{j=1}^{J} K_j \|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}(\boldsymbol{\beta} - \mathbf{v})\|_2^2$, namely $b(\boldsymbol{\beta}) = \min_{\mathbf{v}\in\mathbb{R}^p} l(\boldsymbol{\beta}, \mathbf{v})$. We next show that for all h sufficiently small, $b(h\boldsymbol{\theta}) = l(h\boldsymbol{\theta}, \mathbf{0}_p) = \frac{1}{2n}\|\mathbf{B}h\boldsymbol{\theta}\|_2^2$. Thus

$$d_{\boldsymbol{\theta}} b(\mathbf{0}_p) = \lim_{h\to 0} \frac{b(h\boldsymbol{\theta})}{h} = \lim_{h\to 0} \frac{h}{2n}\|\mathbf{B}\boldsymbol{\theta}\|_2^2 = 0.$$

Since $b(h\boldsymbol{\theta}) \le l(h\boldsymbol{\theta}, \mathbf{0}_p)$, we only need to show that $b(h\boldsymbol{\theta}) \ge l(h\boldsymbol{\theta}, \mathbf{0}_p)$.

$$\begin{aligned}
b(h\boldsymbol{\theta}) - l(h\boldsymbol{\theta}, \mathbf{0}_p) &= \min_{\mathbf{v}\in\mathbb{R}^p} \sum_{j=1}^{J} K_j\|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}(h\boldsymbol{\theta} - \mathbf{v})\|_2^2 - \frac{1}{2n}\|\mathbf{B}h\boldsymbol{\theta}\|_2^2 \\
&= \min_{\mathbf{v}\in\mathbb{R}^p} \sum_{j=1}^{J} K_j\|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}\mathbf{v}\|_2^2 - \frac{h}{n}\boldsymbol{\theta}^{\mathsf{T}}\mathbf{B}^{\mathsf{T}}\mathbf{B}\mathbf{v} \\
&= \min_{\mathbf{v}\in\mathbb{R}^p} \sum_{j=1}^{J} K_j\|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}\mathbf{v}\|_2^2 - \sum_{j=1}^{J} \left\langle \frac{h}{n}(\mathbf{B}^{\mathsf{T}}\mathbf{B})_{j,\cdot}\boldsymbol{\theta}, \mathbf{v}_j \right\rangle \\
&\ge \min_{\mathbf{v}\in\mathbb{R}^p} \sum_{j=1}^{J} K_j\|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}\mathbf{v}\|_2^2 - \sum_{j=1}^{J} \left\| \frac{h}{n}(\mathbf{B}^{\mathsf{T}}\mathbf{B})_{j,\cdot}\boldsymbol{\theta} \right\|_2 \|\mathbf{v}_j\|_2 \\
&= \min_{\mathbf{v}\in\mathbb{R}^p} \sum_{j=1}^{J} \left(K_j - \frac{h}{n}\|(\mathbf{B}^{\mathsf{T}}\mathbf{B})_{j,\cdot}\boldsymbol{\theta}\|_2 \right) \|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}\mathbf{v}\|_2^2,
\end{aligned}$$

where we denote $(\mathbf{B}^{\mathsf{T}}\mathbf{B})_{j,\cdot} \in \mathbb{R}^{p_j\times p}$ as the submatrix of $\mathbf{B}^{\mathsf{T}}\mathbf{B}$ with rows from the j-th group. Then for all h sufficiently small, $\frac{h}{n}\|(\mathbf{B}^{\mathsf{T}}\mathbf{B})_{j,\cdot}\boldsymbol{\theta}\|_2 \le K_j$ for all $j \in [J]$. Therefore,

$$b(h\boldsymbol{\theta}) - l(h\boldsymbol{\theta}, \mathbf{0}_p) \ge \min_{\mathbf{v}\in\mathbb{R}^p} \sum_{j=1}^{J} \left(K_j - \frac{h}{n}\|(\mathbf{B}^{\mathsf{T}}\mathbf{B})_{j,\cdot}\boldsymbol{\theta}\|_2 \right) \|\mathbf{v}_j\|_2 + \frac{1}{2n}\|\mathbf{B}\mathbf{v}\|_2^2 = 0.$$

Having proven that $d_{\boldsymbol{\theta}} b(\mathbf{0}_p) = 0$, we now have

$$\begin{aligned}
d_{\boldsymbol{\theta}} F(\mathbf{0}_p) &= \langle \nabla a(\mathbf{0}_p), \boldsymbol{\theta} \rangle + \lambda \lim_{h\to 0} \frac{\sum_{j=1}^{J} K_j \|h\boldsymbol{\theta}_j\|_2}{h} \\
&= \left\langle -\frac{1}{n}\mathbf{X}^{\mathsf{T}}\mathbf{y}, \boldsymbol{\theta} \right\rangle + \lambda \sum_{j=1}^{J} K_j \|\boldsymbol{\theta}_j\|_2 \\
&= \sum_{j=1}^{J} \left\langle -\frac{1}{n}(\mathbf{X}_{\cdot,j})^{\mathsf{T}}\mathbf{y}, \boldsymbol{\theta}_j \right\rangle + \lambda \sum_{j=1}^{J} K_j \|\boldsymbol{\theta}_j\|_2
\end{aligned}$$

$$\geq \lambda \sum_{j=1}^{J} K_j \|\boldsymbol{\theta}_j\|_2 - \sum_{j=1}^{J} \frac{1}{n} \left\| (\mathbf{X}_{\cdot,j})^{\mathsf{T}} \mathbf{y} \right\|_2 \cdot \left\| \boldsymbol{\theta}_j \right\|_2$$

$$= \sum_{j=1}^{J} \left(\lambda K_j - \frac{1}{n} \|(\mathbf{X}_{\cdot,j})^{\mathsf{T}} \mathbf{y}\|_2 \right) \|\boldsymbol{\theta}_j\|_2,$$

where the inequality is obtained from the Cauchy-Schwarz inequality. So we have $d_{\boldsymbol{\theta}} F(\mathbf{0}_p) \geq 0$ for all $\boldsymbol{\theta} \in \mathbb{R}^p$ when $\lambda \geq \lambda_0 = \max_{j \in \{1,\ldots,J\}} \{ \frac{\|(\mathbf{X}_{\cdot,j})^{\mathsf{T}} \mathbf{y}\|_2}{n K_j} \}$. □

A.5. Proof of Theorem 4.1 and Theorem 4.2

To prove the result of Theorem 4.1, we first establish the following lemma. A proof of the lemma can be found in a subsequent section.

Lemma A.1. *Assume that **A1**–**A4** hold. Define the event*

$$\mathcal{A}_\lambda(c, \xi) = \left\{ \lambda \geq \max_{j \in [J]} \left[c \|(\mathbf{X}_{\cdot,j})^{\mathsf{T}} \boldsymbol{\epsilon}\|_2 / n\xi \right] \right\}.$$

Then,

$$\|\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^\star\|_2 \leq \frac{\lambda \sqrt{|\mathcal{S}|}}{\kappa_{\mathbf{B}}(\mathcal{S}, c)} \left(\bar{\nu} + \frac{\xi}{c} \right)$$

with probability at least $\Pr\{\mathcal{A}_\lambda(c, \xi)\}$.

With this lemma in hand, we only need a concentration inequality for the random variable $\max_{j \in [J]} \left[c \|\epsilon^{\mathsf{T}} \mathbf{X}_{\cdot,j}\|_2 / n\xi \right]$ in order to establish the main result. For that, we have the following.

Lemma A.2. *If* $\mathbf{X}$ *satisfies the block-normalization condition and assumption **A1** holds, then for* $k_1 > 1$

$$\Pr \left\{ \max_{j \in [J]} n^{-1} \|(\mathbf{X}_{\cdot,j})^{\mathsf{T}} \boldsymbol{\epsilon}\|_2 \geq 4\sigma \left(\max_{j \in [J]} \sqrt{\frac{p_j}{n}} + \sqrt{\frac{k_1 \log J}{n}} \right) \right\}$$

$$\leq 2 \exp\{-(k_1 - 1) \log J\}.$$

This further implies that if

$$\lambda = \frac{4c\sigma}{\xi} \left(\max_{j \in [J]} \sqrt{\frac{p_j}{n}} + \sqrt{\frac{k_1 \log J}{n}} \right),$$

then it follows that

$$\Pr\{\mathcal{A}_\lambda(c, \xi)\} \geq 1 - 2 \exp\{(k_1 - 1) \log J\}.$$

Proof of Theorem 4.1. Assume that **A1**–**A4** hold and that $\mathbf{X}$ satisfies the block-normalization condition. If

$$\lambda = \frac{4c\sigma}{\xi} \left(\max_{j \in [J]} \sqrt{\frac{p_j}{n}} + \sqrt{\frac{k_1 \log J}{n}} \right),$$

then applying Lemmas A.1 and A.2,

$$\begin{aligned}
&\Pr\left\{\|\hat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}_*\|_2 \le \frac{\lambda\sqrt{|\mathcal{S}|}}{\kappa_{\mathbf{B}}(\mathcal{S},c)}\left(\bar{\nu}+\frac{\xi}{c}\right)\right\}\\
&=\Pr\left\{\|\hat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}^\star\|_2 \le \frac{4c\sigma\sqrt{|\mathcal{S}|}}{\kappa_{\mathbf{B}}(\mathcal{S},c)}\left(\frac{\bar{\nu}}{\xi}+\frac{1}{c}\right)\left(\max_{j\in[J]}\sqrt{\frac{p_j}{n}}+\sqrt{\frac{k_1\log J}{n}}\right)\right\}\\
&\ge \Pr\{\mathcal{A}_\lambda(c,\xi)\}\\
&\ge 1-2\exp\{-(k_1-1)\log J\}.
\end{aligned}$$

□

We can also use Lemma A.1 to establish an error bound for the case that each $p_j = 1$. By identical arguments used to prove Lemma A.1, we have the following.

Lemma A.3. *Assume that **A1**–**A4** hold and that $p_j = 1$ for $j = 1,\dots,p$. Define the event*

$$\tilde{\mathcal{A}}_\lambda(c,\xi) = \left\{\lambda \ge \max_{j\in[p]} c|(\mathbf{X}_{\cdot,j})^\mathsf{T}\boldsymbol{\epsilon}|/n\xi\right\}.$$

Then,

$$\|\hat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}^\star\|_2 \le \frac{\lambda\sqrt{|\mathcal{S}|}}{\kappa_{\mathbf{B}}(\mathcal{S},c)}\left(\bar{\nu}+\frac{\xi}{c}\right)$$

with probability at least $\Pr\{\tilde{\mathcal{A}}_\lambda(c,\xi)\}$.

Thus, to apply Lemma A.3, we only require another concentration inequality for $\tilde{\mathcal{A}}_\lambda(c,\xi)$. The following lemma provides such a result.

Lemma A.4. *For fixed matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ satisfying the column-wise normalization condition and $\boldsymbol{\epsilon} \in \mathbb{R}^n$, a random vector with independent σ-subgaussian entries, then for $\alpha_0 \in (0,1/2)$,*

$$P\left(\frac{1}{n}\|\mathbf{X}^\mathsf{T}\boldsymbol{\epsilon}\|_\infty \le \sigma\sqrt{\frac{2\log(p/\alpha_0)}{n}}\right) \ge 1-2\alpha_0,$$

where the norm $\|\cdot\|_\infty$ is the maximum absolute value of its argument.

For a proof of Lemma A.4, see the proof of Corollary 2 of [25]. Finally, combining the previous two lemmas leads to the proof of Theorem 4.2.

Proof of Theorem 4.2. Assume that **A1**–**A4** hold and that $\mathbf{X}$ satisfies the column-wise normalization condition. If

$$\lambda = \frac{c\sigma}{\xi}\sqrt{\frac{2\log(p/k_2)}{n}},$$

then applying Lemmas A.3 and A.4, it follows that

$$\Pr\left\{\|\hat{\boldsymbol{\beta}}(\lambda)-\boldsymbol{\beta}^\star\|_2 \le \frac{\lambda\sqrt{|\mathcal{S}|}}{\kappa_{\mathbf{B}}(\mathcal{S},c)}\left(\bar{\nu}+\frac{\xi}{c}\right)\right\}$$

$$
\begin{aligned}
&= \Pr\left\{ \|\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}^\star\|_2 \leq \frac{c\sigma}{\kappa_{\mathbf{B}}(\mathcal{S}, c)} \left(\frac{\bar{\nu}}{\xi} + \frac{1}{c} \right) \sqrt{\frac{2|\mathcal{S}| \log(p/k_2)}{n}} \right\} \\
&\geq \Pr\{\tilde{\mathcal{A}}_\lambda(c, \xi)\} \\
&\geq 1 - 2k_2.
\end{aligned}
$$

□

A.6. Proofs of Lemmas

Throughout this section, we refer to $\hat{\boldsymbol{\beta}}(\lambda)$ as simply $\hat{\boldsymbol{\beta}}$ for ease of display. In order to prove Lemma A.1, we will need an auxiliary lemma which we will prove later.

Lemma A.5. *Suppose assumptions **A1**–**A4** hold. Then, on the event $\mathcal{A}_\lambda(c, \xi)$, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \in \mathbb{C}_n(\mathcal{S}, \nu, c)$.*

Proof of Lemma A.1. In order to prove the main result, we will use the convexity of objective function under assumption A2. Let $F_{\lambda,\mathbf{B}}$ denote the objective function from (2.6). Since $\hat{\boldsymbol{\beta}}$ is the global minimizer (because $F_{\lambda,\mathbf{B}}$ is convex by assumption), we know that

$$
F_{\lambda,\mathbf{B}}(\hat{\boldsymbol{\beta}}) - F_{\lambda,\mathbf{B}}(\boldsymbol{\beta}^\star) \leq 0. \tag{A.3}
$$

We will utilize this inequality to establish the bound. Recall that we define

$$
\Phi_{\mathbf{B}}(\boldsymbol{\beta}) = \sum_{j=1}^{J} K_j \|\boldsymbol{\beta}_j\|_2 - \min_{\mathbf{v} \in \mathbb{R}^p} \left\{ \sum_{j=1}^{J} K_j \|\mathbf{v}_j\|_2 + \frac{1}{2n} \|\mathbf{B}(\boldsymbol{\beta} - \mathbf{v})\|_2^2 \right\},
$$

and recall that $\boldsymbol{\beta}_j^\star \in \mathbb{R}^{p_j}$ is the vector of regression coefficients corresponding to the j-th group. Similarly, recall $\mathbf{v}_j^\star \in \mathbb{R}^{p_j}$ is the subvector of $\mathbf{v}^\star$ analogous to $\boldsymbol{\beta}_j^\star$ for $j \in [J]$. Hence, (A.3) implies

$$
\begin{aligned}
&\frac{1}{2n} \left\{ \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 - \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^\star\|_2^2 \right\} \leq \lambda \left\{ \Phi_{\mathbf{B}}(\boldsymbol{\beta}^\star) - \Phi_{\mathbf{B}}(\hat{\boldsymbol{\beta}}) \right\} \\
\implies &\frac{1}{2n} \left\{ \|\boldsymbol{\epsilon} - \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2 \right\} \leq \lambda \left\{ \Phi_{\mathbf{B}}(\boldsymbol{\beta}^\star) - \Phi_{\mathbf{B}}(\hat{\boldsymbol{\beta}}) \right\} \\
\implies &\frac{1}{2n} \operatorname{tr} \left\{ (\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}})^\mathsf{T} \mathbf{X}^\mathsf{T} \mathbf{X} (\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) - 2\epsilon^\mathsf{T} \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \right\} \leq \lambda \left\{ \Phi_B(\boldsymbol{\beta}^\star) - \Phi_B(\hat{\boldsymbol{\beta}}) \right\}.
\end{aligned} \tag{A.4}
$$

Next, we focus on the difference in the penalties on the right hand side of (A.4).

$$
\begin{aligned}
\Phi_{\mathbf{B}}(\boldsymbol{\beta}^\star) - \Phi_{\mathbf{B}}(\hat{\boldsymbol{\beta}}) = &\sum_{j=1}^{J} K_j \|\boldsymbol{\beta}_j^\star\|_2 - \min_{\mathbf{v} \in \mathbb{R}^p} \left\{ \sum_{j=1}^{J} K_j \|\mathbf{v}_j\|_2 + \frac{1}{2n} \|\mathbf{B}(\boldsymbol{\beta}^\star - \mathbf{v})\|_2^2 \right\} \\
&- \sum_{j=1}^{J} K_j \|\hat{\boldsymbol{\beta}}_j\|_2 + \min_{\mathbf{v} \in \mathbb{R}^p} \left\{ \sum_{j=1}^{J} K_j \|\mathbf{v}_j\|_2 + \frac{1}{2n} \|\mathbf{B}(\hat{\boldsymbol{\beta}} - \mathbf{v})\|_2^2 \right\},
\end{aligned}
$$

and thus, using $\mathbf{v}^\star$ as defined in (4.1), we have

$$\begin{aligned}\Phi_{\mathbf{B}}(\boldsymbol{\beta}^\star) - \Phi_{\mathbf{B}}(\hat{\boldsymbol{\beta}}) \leq & \sum_{j=1}^{J} K_j \|\boldsymbol{\beta}_j^\star\|_2 - \left\{ \sum_{j=1}^{J} K_j \|\mathbf{v}_j^\star\|_2 + \frac{1}{2n} \|\mathbf{B}(\boldsymbol{\beta}^\star - \mathbf{v}^\star)\|_2^2 \right\} \\ & - \sum_{j=1}^{J} K_j \|\hat{\boldsymbol{\beta}}_j\|_2 + \left\{ \sum_{j=1}^{J} K_j \|\mathbf{v}_j^\star\|_2 + \frac{1}{2n} \|\mathbf{B}(\hat{\boldsymbol{\beta}} - \mathbf{v}^\star)\|_2^2 \right\} \\ = & - \frac{1}{2n} \|\mathbf{B}(\boldsymbol{\beta}^\star - \mathbf{v}^\star)\|_2^2 - \sum_{j=1}^{J} K_j (\|\hat{\boldsymbol{\beta}}_j\|_2 - \|\boldsymbol{\beta}_j^\star\|_2) \\ & + \frac{1}{2n} \|\mathbf{B}(\hat{\boldsymbol{\beta}} - \mathbf{v}^\star)\|_2^2 \\ = & - \sum_{j=1}^{J} K_j (\|\hat{\boldsymbol{\beta}}_j\|_2 - \|\boldsymbol{\beta}_j^\star\|_2) + \frac{1}{2n} \|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 \\ & + \frac{1}{n} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)^\mathsf{T} \mathbf{B}^\mathsf{T} \mathbf{B} (\boldsymbol{\beta}^\star - \mathbf{v}^\star).\end{aligned}$$

Let $(\mathbf{B}^\mathsf{T}\mathbf{B})_{j,\cdot} \in \mathbb{R}^{p_j \times p}$ be the submatrix of $\mathbf{B}^\mathsf{T}\mathbf{B}$ consisting of only the rows corresponding to the j-th group of variables. To proceed, we apply Hölder's inequality to the final term.

$$\begin{aligned}\Phi_{\mathbf{B}}(\boldsymbol{\beta}^\star) - \Phi_{\mathbf{B}}(\hat{\boldsymbol{\beta}}) \leq & \sum_{j=1}^{J} K_j (\|\boldsymbol{\beta}_j^\star\|_2 - \|\hat{\boldsymbol{\beta}}_j\|_2) + \frac{1}{2n} \|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 \\ & + \frac{1}{n} \sum_{j=1}^{J} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 \|(\mathbf{B}^\mathsf{T}\mathbf{B})_{j,\cdot} (\boldsymbol{\beta}^\star - \mathbf{v}^\star)\|_2,\end{aligned}$$

and by the reverse triangle inequality,

$$\begin{aligned} = & \sum_{j \in \mathcal{S}} K_j (\|\boldsymbol{\beta}_j^\star + \hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_j\|_2 - \|\hat{\boldsymbol{\beta}}_j\|_2) - \sum_{k \in \mathcal{S}^c} K_j \|\hat{\boldsymbol{\beta}}_k\|_2 \\ & + \frac{1}{2n} \|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 + \frac{1}{n} \sum_{j=1}^{J} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 \|(\mathbf{B}^\mathsf{T}\mathbf{B})_{j,\cdot} (\boldsymbol{\beta}^\star - \mathbf{v}^\star)\|_2 \\ \leq & \sum_{j \in \mathcal{S}} K_j \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 - \sum_{k \in \mathcal{S}^c} K_k \|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^\star\|_2 + \frac{1}{2n} \|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 \\ & + \frac{1}{n} \sum_{j=1}^{J} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 \|(\mathbf{B}^\mathsf{T}\mathbf{B})_{j,\cdot} (\boldsymbol{\beta}^\star - \mathbf{v}^\star)\|_2, \end{aligned}$$

so that finally, we have

$$\begin{aligned}\Phi_{\mathbf{B}}(\boldsymbol{\beta}^\star) - \Phi_{\mathbf{B}}(\hat{\boldsymbol{\beta}}) \leq \sum_{j\in\mathcal{S}} \underbrace{\left\{K_j + \frac{1}{n}\|(\mathbf{B}^\mathsf{T}\mathbf{B})_{j,\cdot}(\boldsymbol{\beta}^\star - \mathbf{v}^\star)\|_2\right\}}_{=:\nu_j \quad (j\in\mathcal{S})} (\|\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j\|_2) \\ - \sum_{k\in\mathcal{S}^c} \underbrace{\left\{K_k - \frac{1}{n}\|(\mathbf{B}^\mathsf{T}\mathbf{B})_{k,\cdot}(\boldsymbol{\beta}^\star - \mathbf{v}^\star)\|_2\right\}}_{=:\nu_k \quad (k\in\mathcal{S}^c)} \|\boldsymbol{\beta}_k^\star - \hat{\boldsymbol{\beta}}_k\|_2 \\ + \frac{1}{2n}\|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2.\end{aligned}$$

Along with (A.4), this thus implies

$$\begin{aligned}&\frac{1}{2n}\operatorname{tr}\left\{(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) - 2\boldsymbol{\epsilon}^\mathsf{T}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\right\} \\ &\leq \lambda\sum_{j\in\mathcal{S}}\nu_j\|\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j\|_2 - \lambda\sum_{k\in\mathcal{S}^c}\nu_k\|\boldsymbol{\beta}_k^\star - \hat{\boldsymbol{\beta}}_k\|_2 + \frac{\lambda}{2n}\|\mathbf{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\|_2^2 \\ &\implies \frac{1}{2n}\operatorname{tr}\left\{(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}})^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X} - \lambda\mathbf{B}^\mathsf{T}\mathbf{B})(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) - 2\boldsymbol{\epsilon}^\mathsf{T}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\right\} \\ &\leq \lambda\sum_{j\in\mathcal{S}}\nu_j\|\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j\|_2 - \lambda\sum_{k\in\mathcal{S}^c}\nu_k\|\boldsymbol{\beta}_k^\star - \hat{\boldsymbol{\beta}}_k\|_2, \qquad \text{(A.5)}\end{aligned}$$

and under **A3** and **A4**,

$$\frac{1}{2n}(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}})^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X} - \lambda\mathbf{B}^\mathsf{T}\mathbf{B})(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) \geq \kappa_B(\mathcal{S}, c)\|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}\|_2^2.$$

Thus, letting $\bar{\nu} = \max_{j\in\mathcal{S}}\nu_j$ and $\underline{\nu} = \min_{k\in\mathcal{S}^c}\nu_k$ and assuming there exists a constant ξ such that $\underline{\nu} \geq \xi > 0$, we have that on $\mathcal{A}_\lambda(c,\xi)$ by Lemma A.5, the inequality (A.5) implies

$$\begin{aligned}&\kappa_B(\mathcal{S}, c)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2 - \frac{1}{n}\boldsymbol{\epsilon}^\mathsf{T}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \\ &\leq \lambda\bar{\nu}\sum_{j\in\mathcal{S}}\|\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j\|_2 - \lambda\underline{\nu}\sum_{k\in\mathcal{S}^c}\|\boldsymbol{\beta}_k^\star - \hat{\boldsymbol{\beta}}_k\|_2. \qquad \text{(A.6)}\end{aligned}$$

Then, because

$$\begin{aligned}\boldsymbol{\epsilon}^\mathsf{T}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \leq |\boldsymbol{\epsilon}^\mathsf{T}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)| &\leq \sum_{j=1}^J \|\boldsymbol{\epsilon}^\mathsf{T}\mathbf{X}_{\cdot,j}\|_2\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 \\ &\leq \max_{j\in[J]}\|\boldsymbol{\epsilon}^\mathsf{T}\mathbf{X}_{\cdot,j}\|_2\sum_{j=1}^J\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2,\end{aligned}$$

(A.6) implies

$$\kappa_{\mathbf{B}}(\mathcal{S}, c)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2 - \max_{j\in[J]}\frac{\|\boldsymbol{\epsilon}^\mathsf{T}\mathbf{X}_{\cdot,j}\|_2}{n}\sum_{j=1}^J\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2$$

$$\leq \lambda\bar{\nu}\sum_{j\in\mathcal{S}}\|\boldsymbol{\beta}_j^\star-\hat{\boldsymbol{\beta}}_j\|_2-\lambda\underline{\nu}\sum_{k\in\mathcal{S}^c}\|\boldsymbol{\beta}_k^\star-\hat{\boldsymbol{\beta}}_k\|_2. \tag{A.7}$$

On $\mathcal{A}_\lambda(c,\xi)$, $\lambda\geq\max_{j\in[J]}c\|\mathbf{X}_{\cdot,j}^\mathsf{T}\boldsymbol{\epsilon}\|_2/n\xi$, so that (A.7) implies

$$\begin{aligned}&\kappa_{\mathbf{B}}(\mathcal{S},c)\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^\star\|_2^2-\frac{\lambda\xi}{c}\sum_{j=1}^J\|\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j^\star\|_2\\&\leq\lambda\bar{\nu}\sum_{j\in\mathcal{S}}\|\boldsymbol{\beta}_j^\star-\hat{\boldsymbol{\beta}}_j\|_2-\lambda\underline{\nu}\sum_{k\in\mathcal{S}^c}\|\boldsymbol{\beta}_k^\star-\hat{\boldsymbol{\beta}}_k\|_2,\end{aligned}$$

which implies

$$\begin{aligned}\kappa_{\mathbf{B}}(\mathcal{S},c)\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^\star\|_2^2&\leq\lambda\left(\bar{\nu}+\frac{\xi}{c}\right)\sum_{j\in\mathcal{S}}\|\boldsymbol{\beta}_j^\star-\hat{\boldsymbol{\beta}}_j\|_2-\lambda\left(\underline{\nu}-\frac{\xi}{c}\right)\sum_{k\in\mathcal{S}^c}\|\boldsymbol{\beta}_k^\star-\hat{\boldsymbol{\beta}}_k\|_2\\&\leq\lambda\left(\bar{\nu}+\frac{\xi}{c}\right)\sum_{j\in\mathcal{S}}\|\boldsymbol{\beta}_j^\star-\hat{\boldsymbol{\beta}}_j\|_2\\&\leq\lambda\left(\bar{\nu}+\frac{\xi}{c}\right)\sqrt{|\mathcal{S}|}\|\boldsymbol{\beta}^\star-\hat{\boldsymbol{\beta}}\|_2,\end{aligned}$$

from which we can finally conclude that

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^\star\|_2\leq\frac{\lambda(\bar{\nu}+\frac{\xi}{c})\sqrt{|\mathcal{S}|}}{\kappa_{\mathbf{B}}(\mathcal{S},c)}.$$

□

Proof of Lemma A.2. We follow the same arguments as in the proof of Corollary 9.28 of [34]. Throughout, let $\mathbf{x}_{i(j)}\in\mathbb{R}^{p_j}$ be the i-th row of $\mathbf{X}_{\cdot,j}$ for $i\in[n]$ and $j\in[J]$, and let $\mathbb{S}^{p_j-1}=\{\mathbf{u}\in\mathbb{R}^{p_j}:\|\mathbf{u}\|_2=1\}$. First, note that for each $j\in[J]$, we can write

$$n^{-1}(\mathbf{X}_{\cdot,j})^\mathsf{T}\boldsymbol{\epsilon}=n^{-1}\sum_{i=1}^n\mathbf{x}_{(i)j}\epsilon_i=\sum_{i=1}^n\mathbf{w}_{i(j)},$$

where $\mathbf{w}_{i(j)}=n^{-1}\mathbf{x}_{(i)j}\epsilon_i$, so that

$$\left\|n^{-1}(\mathbf{X}_{\cdot,j})^\mathsf{T}\boldsymbol{\epsilon}\right\|_2=\left\|\sum_{i=1}^n\mathbf{w}_{i(j)}\right\|_2=\sup_{\mathbf{u}\in\mathbb{S}^{p_j-1}}\left\langle\mathbf{u},\sum_{i=1}^n\mathbf{w}_{i(j)}\right\rangle.$$

Then, applying the fact that we can construct a 1/2-covering of $\mathbb{S}^{p_j-1}$ in the Euclidean norm, say $\{\mathbf{u}_1,\ldots,\mathbf{u}_M\}$, with cardinality $M\leq5^{p_j}$, a straightforward discretization argument [34, Chapter 5] yields

$$\left\|\sum_{i=1}^n\mathbf{w}_{i(j)}\right\|_2\leq2\max_{k\in[M]}\left\langle\mathbf{u}_k,\sum_{i=1}^n\mathbf{w}_{i(j)}\right\rangle.$$

Next, by definition of $\mathbf{w}_{i(j)}$, it can be verified that $\langle \mathbf{u}_k, \sum_{i=1}^n \mathbf{w}_{i(j)} \rangle$ is χ-subgaussian where

$$\chi = \frac{\sigma}{n} \|(\mathbf{X}_{\cdot,j})^\mathsf{T} \mathbf{u}_k\|_2 \leq \frac{\sigma}{n} \|\mathbf{X}_{\cdot,j}\| \leq \frac{\sigma}{\sqrt{n}}.$$

Thus, by applying a standard subgaussian tail bound to $\langle \mathbf{u}_k, \sum_{i=1}^n \mathbf{w}_{i(j)} \rangle$, then applying the union bound over the $k \in [M]$, we have that

$$\Pr\left(n^{-1} \|(\mathbf{X}_{\cdot,j})^\mathsf{T} \boldsymbol{\epsilon}\|_2 \geq 2t\right) \leq 2 \exp\left(-\frac{nt^2}{2\sigma^2} + p_j \log 5\right).$$

Then, applying the union bound over $j \in [J]$, we have

$$\Pr\left(n^{-1} \max_{j \in [J]} \|(\mathbf{X}_{\cdot,j})^\mathsf{T} \epsilon\|_2 \geq 2t\right) \leq 2 \exp\left(-\frac{nt^2}{2\sigma^2} + \max_{j \in [J]} p_j \log 5 + \log J\right). \tag{A.8}$$

Thus, if for constant $k_1 > 1$, we take

$$\begin{aligned} t &= 2\sigma \left(\sqrt{\frac{\max_{j \in [J]} p_j}{n}} + \sqrt{\frac{k_1 \log J}{n}}\right) \\ &\geq \sqrt{2\sigma^2} \left(\sqrt{\frac{\max_{j \in [J]} p_j \log 5 + k_1 \log J}{n}}\right) =: t^*, \end{aligned}$$

then (A.8) implies

$$\begin{aligned} \Pr\left(\max_{j \in [J]} n^{-1} \|(\mathbf{X}_{\cdot,j})^\mathsf{T} \epsilon\|_2 \geq 2t\right) &\leq 2 \exp\left(-\frac{nt^2}{2\sigma^2} + \log 5 \max_{j \in [J]} p_j + \log J\right) \\ &\leq 2 \exp\left(-\frac{nt^{*2}}{2\sigma^2} + \log 5 \max_{j \in [J]} p_j + \log J\right) \\ &= 2 \exp\left\{-(k_1 - 1) \log J\right\}. \end{aligned}$$

□

It remains only to prove Lemma A.5.

Proof of Lemma A.5. We want to show that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \in \mathbb{C}_n(\mathcal{S}, \nu, c)$ on $\mathcal{A}_\lambda(c, \xi)$. Starting from (A.5), we have

$$\begin{aligned} &\frac{1}{2n} \operatorname{tr}\left\{(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}})^\mathsf{T} (\mathbf{X}^\mathsf{T}\mathbf{X} - \lambda \mathbf{B}^\mathsf{T}\mathbf{B})(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}) - 2\boldsymbol{\epsilon}^\mathsf{T}\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)\right\} \\ &\leq \lambda \sum_{j \in \mathcal{S}} \nu_j \|\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j\|_2 - \lambda \sum_{k \in \mathcal{S}^c} \nu_k \|\boldsymbol{\beta}_k^\star - \hat{\boldsymbol{\beta}}_k\|_2. \end{aligned}$$

Then, because $\mathbf{X}^\mathsf{T}\mathbf{X} \succeq \lambda \mathbf{B}^\mathsf{T}\mathbf{B}$ by assumption **A2**, the previous inequality, along with the same argument used to obtain (A.7) yields

$$-\frac{1}{n} \sum_{j=1}^J \|\epsilon^\mathsf{T} \mathbf{X}_{\cdot,j}\|_2 \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 \leq \lambda \sum_{j \in \mathcal{S}} \nu_j \|\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j\|_2 - \lambda \sum_{k \in \mathcal{S}^c} \nu_k \|\boldsymbol{\beta}_k^\star - \hat{\boldsymbol{\beta}}_k\|_2,$$

which implies

$$
\begin{aligned}
&- \frac{\max_{j\in[J]} \|\boldsymbol{\epsilon}^{\mathsf{T}}\mathbf{X}_{\cdot,j}\|_2}{n} \sum_{j=1}^{J} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 \\
&\leq \lambda \sum_{j\in\mathcal{S}} \nu_j \|\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j\|_2 - \lambda \sum_{k\in\mathcal{S}^c} \nu_k \|\boldsymbol{\beta}_k^\star - \hat{\boldsymbol{\beta}}_k\|_2. \qquad \text{(A.9)}
\end{aligned}
$$

Hence, on $\mathcal{A}_\lambda(c,\xi)$, $\lambda \geq \max_{j\in[J]} c\|\mathbf{X}_{\cdot,j}^{\mathsf{T}}\boldsymbol{\epsilon}\|_2/n\xi$ so that (A.9) implies

$$
\begin{aligned}
&- \frac{\lambda\xi}{c} \sum_{j=1}^{J} \|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j^\star\|_2 \leq \lambda \sum_{j\in\mathcal{S}} \nu_j \|\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j\|_2 - \lambda \sum_{k\in\mathcal{S}^c} \nu_k \|\boldsymbol{\beta}_k^\star - \hat{\boldsymbol{\beta}}_k\|_2 \\
&\Longrightarrow 0 \leq \lambda \sum_{j\in\mathcal{S}} \left(\nu_j + \frac{\xi}{c}\right) \|\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j\|_2 - \lambda \sum_{k\in\mathcal{S}^c} \left(\nu_k - \frac{\xi}{c}\right) \|\boldsymbol{\beta}_k^\star - \hat{\boldsymbol{\beta}}_k\|_2 \\
&\Longrightarrow \sum_{k\in\mathcal{S}^c} \left(\nu_k - \frac{\xi}{c}\right) \|\boldsymbol{\beta}_k^\star - \hat{\boldsymbol{\beta}}_k\|_2 \leq \sum_{j\in\mathcal{S}} \left(\nu_j + \frac{\xi}{c}\right) \|\boldsymbol{\beta}_j^\star - \hat{\boldsymbol{\beta}}_j\|_2
\end{aligned}
$$

from which the conclusion follows. □

A.7. Proof of Proposition 4.1

In the case that $\mathbf{B} = \sqrt{\eta/\lambda}\mathbf{I}_p$, by definition

$$
\mathbf{v}^\star = \underset{\mathbf{v}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{j=1}^{J} K_j \|\mathbf{v}_j\|_2 + \frac{\eta}{2\lambda n} \|\boldsymbol{\beta}^\star - \mathbf{v}\|_2^2 \right\}.
$$

One can see then that this reduces to a version of the proximal operator of the l_2-norm, so that for $\boldsymbol{\beta}_j^\star \neq \mathbf{0}$, it follows that

$$
\mathbf{v}_j^\star = \max\left(1 - \frac{\lambda n K_j}{\eta\|\boldsymbol{\beta}_j^\star\|_2}, 0\right) \boldsymbol{\beta}_j^\star, \quad j \in \mathcal{S},
$$

and for $j \in \mathcal{S}^c$, $\mathbf{v}_j^\star = \mathbf{0}$. Hence, for $j \in \mathcal{S}$,

$$
\begin{aligned}
\nu_j &= K_j + \frac{1}{n}\|[\mathbf{B}^{\mathsf{T}}\mathbf{B}]_{j,\cdot}(\boldsymbol{\beta}^\star - \mathbf{v}^\star)\|_2 \\
&= K_j + \frac{\eta}{\lambda n}\|\boldsymbol{\beta}_j^\star - \mathbf{v}_j^\star\|_2 \\
&= K_j + \frac{\eta}{\lambda n}\left\{1 - \max\left(1 - \frac{\lambda n K_j}{\eta\|\boldsymbol{\beta}_j^\star\|_2}, 0\right)\right\}\|\boldsymbol{\beta}_j^\star\|_2.
\end{aligned}
$$

Thus, if $\|\boldsymbol{\beta}_j^\star\|_2 > \lambda n K_j/\eta$,

$$
\nu_j = K_j + \frac{\eta}{\lambda n}\left(\frac{\lambda n K_j}{\eta}\right) = 2K_j.
$$

Algorithm 3 Adaptive PDHG for the group GMC problem (3.2)

Input: Set $\boldsymbol{\beta}_0 \in \mathbb{R}^p, \mathbf{v}_0 \in \mathbb{R}^p$, and a small sufficient value `tol`.
Initialize $\tau_0\sigma_0 < \|\mathbf{Z}^\mathsf{T}\mathbf{Z}\|^{-1}$ and $(\alpha_0, \zeta) \in (0,1)^2$.

while $\|\mathbf{r}_k\|_2, \|\mathbf{d}_k\|_2 >$ `tol` **do**

Compute PDHG updates with $f(\boldsymbol{\beta})$ and $g(\mathbf{v})$ defined in (3.3) and (3.4):

$$\begin{cases} \hat{\boldsymbol{\beta}}_{k+1} = \boldsymbol{\beta}_k - \tau_k \mathbf{Z}^\mathsf{T}\mathbf{v}_k \\ \boldsymbol{\beta}_{k+1} = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} f(\boldsymbol{\beta}) + \frac{1}{2\tau_k}\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k+1}\|_2^2 \\ \hat{\mathbf{v}}_{k+1} = \mathbf{v}_k + \sigma_k \mathbf{Z}(2\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k) \\ \mathbf{v}_{k+1} = \operatorname{argmin}_{\mathbf{v}\in\mathbb{R}^p} g(\mathbf{v}) + \frac{1}{2\sigma_k}\|\mathbf{v} - \hat{\mathbf{v}}_{k+1}\|_2^2 \end{cases}$$

Compute residuals:

$$\begin{cases} \mathbf{r}_{k+1} = \frac{1}{\tau_k}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k+1}) - \mathbf{Z}^T(\mathbf{v}_k - \mathbf{v}_{k+1}) \\ \mathbf{d}_{k+1} = \frac{1}{\sigma_k}(\mathbf{v}_k - \mathbf{v}_{k+1}) + \mathbf{Z}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k+1}) \end{cases}$$

if $2\|\mathbf{r}_{k+1}\|_2 < \|\mathbf{d}_{k+1}\|_2$ **then**

$$\begin{cases} \tau_{k+1} = \tau_k(1-\alpha_k) \\ \sigma_{k+1} = \sigma_k/(1-\alpha_k) \\ \alpha_{k+1} = \zeta\alpha_k \end{cases}$$

else if $\|\mathbf{r}_{k+1}\|_2 > 2\|\mathbf{d}_{k+1}\|_2$ **then**

$$\begin{cases} \tau_{k+1} = \tau_k/(1-\alpha_k) \\ \sigma_{k+1} = \sigma_k(1-\alpha_k) \\ \alpha_{k+1} = \zeta\alpha_k \end{cases}$$

else

$$\begin{cases} \tau_{k+1} = \tau_k \\ \sigma_{k+1} = \sigma_k \\ \alpha_{k+1} = \alpha_k \end{cases}$$

end if

end while

Similarly, if $\|\boldsymbol{\beta}_j^\star\|_2 \le \lambda n K_j/\eta$,

$$\nu_j = K_j + \frac{\eta}{\lambda n}\|\boldsymbol{\beta}_j^\star\|_2.$$

Finally, since $\mathbf{v}_j^\star = \mathbf{0}$ for $j \in \mathcal{S}^c$, it is trivial to see that $\nu_j = K_j$ for $j \in \mathcal{S}^c$. □

Appendix B: Adaptive PDHG for group GMC

Algorithm 3 provides pseudocode of the adaptive PDHG algorithm for solving the group GMC problem (3.2). Algorithm 3 is an instance of the adaptive PDHG algorithm proposed by [15] to a saddle point formulation of the group GMC problem.

The following convergence guarantee for Algorithm 3 is adapted from Theorem 1 in [15].

Proposition B.1. *Under the convexity preserving condition, the iterate sequence* $\{(\boldsymbol{\beta}_k, \mathbf{v}_k)\}$ *generated by Algorithm 3 converges in the sense that the pri-*

mal and dual residuals $\mathbf{r}_k$ *and* $\mathbf{d}_k$ *vanish, i.e.,*

$$\lim_{k\to\infty} \|\mathbf{r}_k\|_2^2 + \|\mathbf{d}_k\|_2^2 = 0.$$

Before proving Proposition B.1, we unpack the motivation for characterizing convergence in terms vanishing residuals. The pair $(\boldsymbol{\beta}^\star, \mathbf{v}^\star)$ is a solution to the saddle-point problem (3.2) if and only if

$$\mathbf{0} \in \partial f(\boldsymbol{\beta}) + \mathbf{Z}^\mathsf{T}\mathbf{v} \tag{B.1}$$

$$\mathbf{0} \in \partial g(\mathbf{v}) - \mathbf{Z}\boldsymbol{\beta} \tag{B.2}$$

where $\partial h(\mathbf{x})$ denotes the subdifferential of a function h at $\mathbf{x}$ [7]. The updates in line 3 of Algorithm 3 satisfy the following conditions for all k

$$\mathbf{r}_{k+1} \in \partial f(\boldsymbol{\beta}_{k+1}) + \mathbf{Z}^\mathsf{T}\mathbf{v}_{k+1}$$
$$\mathbf{d}_{k+1} \in \partial g(\mathbf{v}_{k+1}) - \mathbf{Z}\boldsymbol{\beta}_{k+1}.$$

These conditions follow from the convexity of f and g and Fermat's rule.

The subdifferential $\partial h(\mathbf{x})$ of a function h is a closed point-to-set mapping if h is a lower-semicontinuous, proper, and convex function [28, Theorem 24.4]. Under the convexity-preserving condition (2.8), both $f(\boldsymbol{\beta})$ and $g(\mathbf{v})$ are lower-semicontinuous, proper, convex functions, so the subdifferentials $\partial f(\boldsymbol{\beta})$ and $\partial g(\mathbf{v})$ are closed point-to-set mappings. Recall a closed mapping is the analog of a continuous function for a point-to-set mapping. Specifically, ϕ is closed point-to-set mapping if

i. $\mathbf{x}_k \to \mathbf{x}$
ii. $\mathbf{y}_k \to \mathbf{y}$ and $\mathbf{y}_k \in \phi(\mathbf{x}_k)$

together imply that $\mathbf{x} \in \phi(\mathbf{y})$. Thus, if $(\boldsymbol{\beta}_k, \mathbf{v}_k)$ converges to $(\boldsymbol{\beta}', \mathbf{v}')$, then $\mathbf{0} \in \partial f(\boldsymbol{\beta}') + \mathbf{Z}^\mathsf{T}\mathbf{v}'$ and $\mathbf{0} \in \partial g(\mathbf{v}') - \mathbf{Z}\boldsymbol{\beta}'$. In other words, convergence in residuals implies that if the primal dual sequence converges, the limit is a solution to the saddle point problem.

Proof. Goldstein et al. [15] establish three conditions that ensure the residuals of Algorithm 3 vanish provided that f and g are lower-semicontinuous, proper, convex functions. The convexity-preserving condition (2.8) ensures that f and g are lower-semicontinous, proper, convex functions. Note that these three conditions are standard assumptions that ensure the existence and uniqueness of proximal mappings. The three additional conditions are given below.

(1) *The sequences* $\{\tau_k\}$ *and* $\{\sigma_k\}$ *are positive and bounded.*
The initializations and multiplicative updates in Algorithm 3 imply the positiveness of $\{\tau_k\}$ and $\{\sigma_k\}$.
Since the step-sizes are bounded from below, we just need to show that they are bounded from above. We prove that the sequence $\{\sigma_k\}$ is bounded from above. The proof of the boundedness of $\{\tau_k\}$ is identical. Construct a surrogate sequence $\{s_k\}$ where

$$s_0 = \sigma_0 \quad \text{and} \quad s_k = \frac{s_{k-1}}{1 - \alpha_0 \zeta^{k-1}} \quad \text{for all } k \geq 1.$$

We use an induction argument to establish that $\sigma_k \leq s_k$ for all k. The inequality holds for $k = 0$. Select an arbitrary k such that the inequality holds. Then

$$\begin{aligned}\sigma_{k+1} &\leq \max\{\sigma_k/(1-\alpha_k), \sigma_k(1-\alpha_k), \sigma_{k-1}\} \\ &= \sigma_k/(1-\alpha_k) \leq \sigma_k/(1-\alpha_0\zeta^k) \leq s_k/(1-\alpha_0\zeta^k) = s_{k+1}.\end{aligned}$$

We next show that $\{s_k\}$ is bounded from above, which will imply $\{\sigma_k\}$ is bounded since $\sigma_k \leq s_k$ for all k. Note that

$$s_k = \left[\prod_{j=0}^{k-1}(1-\alpha_0\zeta^j)\right]^{-1}\sigma_0 \tag{B.3}$$

and

$$\begin{aligned}-\log(1-\alpha_0\zeta^k) &= -\log(1-\zeta^k+\zeta^k-\alpha_0\zeta^k) \\ &\leq -(1-\zeta^k)\log(1) - \zeta^k\log(1-\alpha_0) \\ &= -\zeta^k\log(1-\alpha_0)\end{aligned} \tag{B.4}$$

where the inequality follows from the convexity of $-\log x$.
Taking the logarithm of both sides of (B.3) and applying (B.4) gives

$$\begin{aligned}\log s_k &= \log\sigma_0 - \sum_{j=0}^{k-1}\log(1-\alpha_0\zeta^j) \\ &\leq \log\sigma_0 - \log(1-\alpha_0)\sum_{j=0}^{k-1}\zeta^j \\ &= \log\sigma_0 - \log(1-\alpha_0)\frac{1-\zeta^k}{1-\zeta}.\end{aligned}$$

Therefore,

$$\log s_k \leq \log\sigma_0 + \frac{-\log(1-\alpha_0)}{1-\zeta},$$

and consequently $\sigma_k \leq \sigma_0(\frac{1}{1-\alpha_0})^{\frac{1}{1-\zeta}}$ for all k.

(2) *The sequence* $\{\phi_k\}$ *is summable, where* $\phi_k = \max\{\frac{\tau_k-\tau_{k+1}}{\tau_k}, \frac{\sigma_k-\sigma_{k+1}}{\sigma_k}, 0\}$.
According to the updates of τ_k and σ_k, we have

$$\frac{\tau_k-\tau_{k+1}}{\tau_k} = \begin{cases}\alpha_k, & 2\|\mathbf{r}_{k+1}\|_2 < \|\mathbf{d}_{k+1}\|_2 \\ -\frac{\alpha_k}{1-\alpha_k}, & \|\mathbf{r}_{k+1}\|_2 > 2\|\mathbf{d}_{k+1}\|_2 \\ 0, & \text{otherwise.}\end{cases}$$

and

$$\frac{\sigma_k-\sigma_{k+1}}{\sigma_k} = \begin{cases}-\frac{\alpha_k}{1-\alpha_k}, & 2\|\mathbf{r}_{k+1}\|_2 < \|\mathbf{d}_{k+1}\|_2 \\ \alpha_k, & \|\mathbf{r}_{k+1}\|_2 > 2\|\mathbf{d}_{k+1}\|_2 \\ 0, & \text{otherwise.}\end{cases}$$

Therefore, we have

$$\phi_k = \max\left\{\frac{\tau_k - \tau_{k+1}}{\tau_k}, \frac{\sigma_k - \sigma_{k+1}}{\sigma_k}, 0\right\} = \alpha_k.$$

The geometric decay of α_k implies that condition (2) holds.

(3) *There is a constant L such that for all $k > 0$,*

$$\tau_k \sigma_k < L < \|\mathbf{Z}^\mathsf{T}\mathbf{Z}\|^{-1}. \tag{B.5}$$

This condition is easily met by setting $\tau_0\sigma_0 < \|\mathbf{Z}^\mathsf{T}\mathbf{Z}\|^{-1}$ since the product $\tau_k\sigma_k$ is constant after the updates in lines 4–10 in Algorithm 3. □

Condition (B.5) imposes a conservative bound on the step-size parameters. Consequently, Goldstein et al. [15] also introduced an adaptive PDHG algorithm with backtracking line search that frees the algorithm to take larger step-sizes which can lead to fewer iterations and in turn faster convergence. Algorithm 4 provides pseudocode of the adaptive PDHG algorithm with backtracking. Convergence guarantees that are identical to Proposition B.1 hold for Algorithm 4 under the additional assumption that either $\boldsymbol{\beta}$ or $\mathbf{v}$ is bounded [15, 12]. In practice, however, Goldstein et al. [15] observed that adaptive PDHG with backtracking converged even for unbounded problems, e.g., linear programs.

In our R package `GMC`, users have the flexibility to use basic PDHG, adaptive PDHG, or adaptive PDHG with backtracking to solve the group GMC problem. We used adaptive PDHG with backtracking for all our experiments.

While convergence in the sense given in Proposition B.1 provides some assurances about the limiting behavior of the PDHG iterate sequence, we do not have global convergence guarantees, i.e., the iterate sequence converges to a saddle point of the group GMC problem (3.2). We can, however, certify whether a pair $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$ is saddle point of the group GMC problem (3.2) using (B.1) and (B.2).

Substituting (3.3) and (3.4) for f and g respectively into (B.1) and (B.2) gives us

$$-\frac{1}{\lambda n}\mathbf{X}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) + \frac{1}{n}\mathbf{B}^\mathsf{T}\mathbf{B}(\boldsymbol{\beta} - \mathbf{v}) \in \partial\mathcal{R}(\boldsymbol{\beta}), \tag{B.6}$$

$$\frac{1}{n}\mathbf{B}^\mathsf{T}\mathbf{B}(\boldsymbol{\beta} - \mathbf{v}) \in \partial\mathcal{R}(\mathbf{v}), \tag{B.7}$$

where $\mathcal{R}(\boldsymbol{\beta}) = \sum_{j=1}^{J} K_j\|\boldsymbol{\beta}_j\|_2$ is the group Lasso penalty. The subgradient $\partial\mathcal{R}(\boldsymbol{\beta})$ is given by

$$\partial\mathcal{R}(\boldsymbol{\beta}) = \left\{\mathbf{u} \in \mathbb{R}^p : \langle \mathbf{u}_j, \boldsymbol{\beta}_j\rangle = K_j\|\boldsymbol{\beta}_j\|_2,\ \|\mathbf{u}_j\|_2 \le K_j,\ j \in [J]\right\},$$

where $\mathbf{u}_j \in \mathbb{R}^{p_j}$ is the subvector of $\mathbf{u}$ with components in the j-th group. Let $\mathbf{b} = \frac{1}{n}\mathbf{B}^\mathsf{T}\mathbf{B}(\boldsymbol{\beta} - \mathbf{v})$ and $\mathbf{a} = -\frac{1}{\lambda n}\mathbf{X}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) + \mathbf{b}$. Then checking the optimality conditions (B.6)–(B.7) for problem (3.2) are equivalent to checking the following hold for all $j \in [J]$:

$$\langle \mathbf{a}_j, \boldsymbol{\beta}_j\rangle = K_j\|\boldsymbol{\beta}_j\|_2, \tag{B.8}$$

Algorithm 4 Backtracking PDHG for the group GMC problem (3.2)

Input: Set $\boldsymbol{\beta}_0 \in \mathbb{R}^p, \mathbf{v}_0 \in \mathbb{R}^p$, and a small sufficient value `tol`.
Initialize $(\tau_0, \sigma_0) \in (0, +\infty)^2$, and $(\alpha_0, \zeta, b) \in (0,1)^3$.

while $\|\mathbf{r}_k\|_2, \|\mathbf{d}_k\|_2 >$ `tol` **do**

Compute PDHG updates with $f(\boldsymbol{\beta})$ and $g(\mathbf{v})$ defined in (3.3) and (3.4):

$$\begin{cases} \hat{\boldsymbol{\beta}}_{k+1} = \boldsymbol{\beta}_k - \tau_k \mathbf{Z}^{\mathsf{T}} \mathbf{v}_k \\ \boldsymbol{\beta}_{k+1} = \operatorname{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} f(\boldsymbol{\beta}) + \frac{1}{2\tau_k} \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_{k+1}\|_2^2 \\ \hat{\mathbf{v}}_{k+1} = \mathbf{v}_k + \sigma_k \mathbf{Z}(2\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k) \\ \mathbf{v}_{k+1} = \operatorname{argmin}_{\mathbf{v} \in \mathbb{R}^p} g(\mathbf{v}) + \frac{1}{2\sigma_k} \|\mathbf{v} - \hat{\mathbf{v}}_{k+1}\|_2^2 \end{cases}$$

Check the following backtracking condition:

$$\frac{b}{2\tau_k} \|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k\|_2^2 - 2(\mathbf{v}_{k+1} - \mathbf{v}_k)^T \mathbf{Z}(\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k) + \frac{b}{2\tau_k} \|\mathbf{v}_{k+1} - \mathbf{v}_k\|_2^2 > 0$$

and if it fails, set

$$\begin{cases} \tau_k = \tau_k / 2 \\ \sigma_k = \sigma_k / 2 \end{cases}$$

Compute residuals:

$$\begin{cases} \mathbf{r}_{k+1} = \frac{1}{\tau_k}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k+1}) - \mathbf{Z}^T(\mathbf{v}_k - \mathbf{v}_{k+1}) \\ \mathbf{d}_{k+1} = \frac{1}{\sigma_k}(\mathbf{v}_k - \mathbf{v}_{k+1}) + \mathbf{Z}(\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k+1}) \end{cases}$$

if $2\|\mathbf{r}_{k+1}\|_2 < \|\mathbf{d}_{k+1}\|_2$ **then**

$$\begin{cases} \tau_{k+1} = \tau_k (1 - \alpha_k) \\ \sigma_{k+1} = \sigma_k / (1 - \alpha_k) \\ \alpha_{k+1} = \zeta \alpha_k \end{cases}$$

else if $\|\mathbf{r}_{k+1}\|_2 > 2\|\mathbf{d}_{k+1}\|_2$ **then**

$$\begin{cases} \tau_{k+1} = \tau_k / (1 - \alpha_k) \\ \sigma_{k+1} = \sigma_k (1 - \alpha_k) \\ \alpha_{k+1} = \zeta \alpha_k \end{cases}$$

else

$$\begin{cases} \tau_{k+1} = \tau_k \\ \sigma_{k+1} = \sigma_k \\ \alpha_{k+1} = \alpha_k \end{cases}$$

end if

end while

$$\|\mathbf{a}_j\|_2 \le K_j, \tag{B.9}$$

$$\langle \mathbf{b}_j, \mathbf{v}_j \rangle = K_j \|\mathbf{v}_j\|_2, \tag{B.10}$$

$$\|\mathbf{b}_j\|_2 \le K_j. \tag{B.11}$$

If Algorithm 4 outputs a pair $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{v}})$ that satisfies conditions (B.8)–(B.11), then the pair is a saddle point of (3.2) and thererfore $\hat{\boldsymbol{\beta}}$ is an optimal solution of the group GMC problem. We have used this check to certify our group GMC estimates are globally optimal in our experiments.

We provide an example to illustrate how we verified the optimality of our computed group GMC estimates. We quantify violations of conditions (B.8)–(B.11) using vectors $\mathbf{e}_1 \in \mathbb{R}^J, \mathbf{e}_2 \in \mathbb{R}^J, \mathbf{e}_3 \in \mathbb{R}^J$, and $\mathbf{e}_4 \in \mathbb{R}^J$, where the j-th compo-

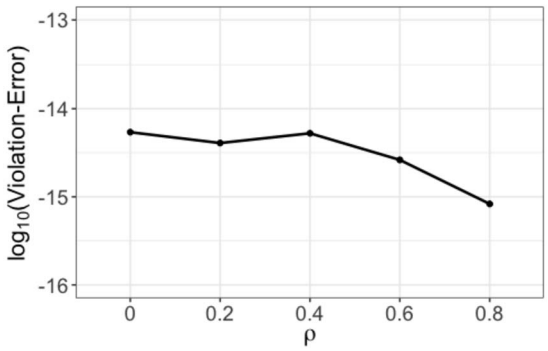

FIG 5. *Maximum* `Violation-Error` *over* 20 *replicates over different values of* ρ *for experiments in Case C2.*

nent of each vector is given by

$$\begin{aligned}
\mathbf{e}_{1,j} &= \langle \mathbf{a}_j, \boldsymbol{\beta}_j \rangle - K_j \|\boldsymbol{\beta}_j\|_2, \\
\mathbf{e}_{2,j} &= \max\left(\|\mathbf{a}_j\|_2 - K_j, 0\right), \\
\mathbf{e}_{3,j} &= \langle \mathbf{b}_j, \mathbf{v}_j \rangle - K_j \|\mathbf{v}_j\|_2, \\
\mathbf{e}_{4,j} &= \max\left(\|\mathbf{b}_j\|_2 - K_j, 0\right).
\end{aligned}$$

We encode the overall violation of the optimality conditions for problem (3.2) with the mean squared error

$$\texttt{Violation-Error} = \frac{1}{4J} \sum_{i=1}^{4} \|\mathbf{e}_i\|_2^2.$$

Figure 5 displays the maximum `Violation-Error` over 20 replicates over different values of ρ for experiments in *Case C2* in Section 5, where the value of λ was selected by cross-validation. The `Violation-Error` is smaller than 10^{-14} for all replicates for all values of ρ, certifying that our group GMC estimates are optimal up to machine precision.

Appendix C: Additional simulation experiments

We consider an additive model with both categorical and continuous variables. Similar to what we did in Section 5, we explore the effects of four factors of interest including the SNR, the correlation among groups, the problem dimension, and the convexity-preserving parameter α for the group GMC method.

The data generation process of the additive model is as follows. Twenty continuous covariates $X_1, \ldots, X_{20}$ are defined as $X_i = Z_i + tW$, where Z_i and W are independently sampled from a standard normal distribution, and t is a constant controlling the correlation between X_i and X_j. Then $X_{11}, \ldots, X_{20}$ are

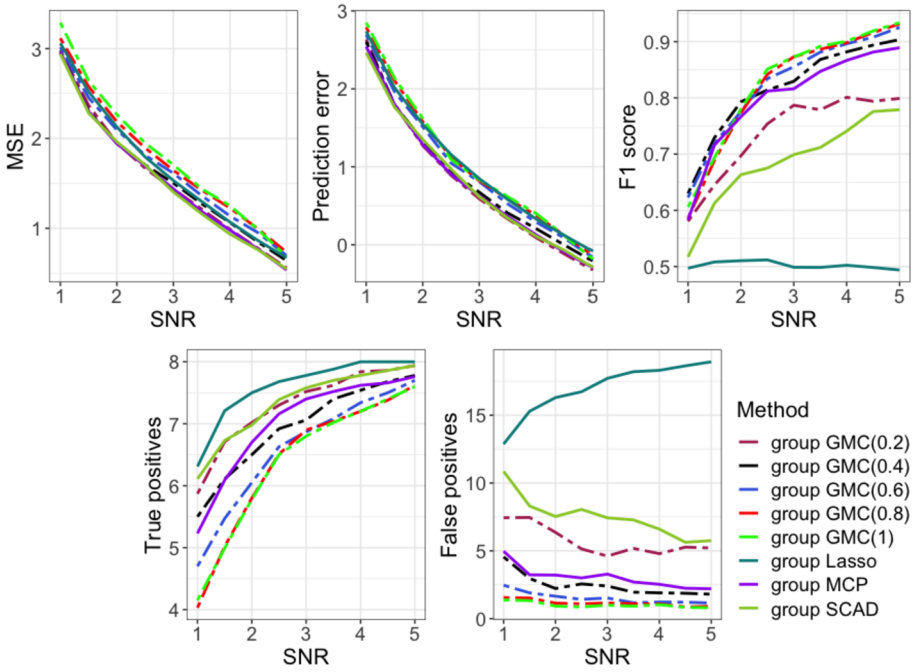

FIG 6. *Results for Case I: Impact of SNR. Group GMC(·) stands for the group GMC with a specific value of* α. *Average performance based on* 100 *simulation replicates for each method. MSE and Prediction error are on a log scale.*

trichotomized in the same way as in the simulation study in Section 5. Then the response y is simulated from

$$y = X_3^3 + X_3^2 + X_3 + \frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6 + 2\mathbb{1}(X_{11} = 0) + \mathbb{1}(X_{11} = 1) + \epsilon, \quad \text{(C.1)}$$

where the noise $\epsilon \sim N(0, \sigma^2)$. That is, we have fifty covariate variables from twenty groups, where ten of them are continuous with a group size of three and the other ten are categorical with a group size of two. But the response variable only depends on three groups of covariates. We next consider three different cases to explore the possible effects of interesting factors.

Case I: In this case, we aim to see how the SNR affects the performance of different methods. We fix $t = 0$ and set a sequence of σ so that the SNR takes on values from 1 to 5. We sample 100 data sets for each setting. We also report the results of the group GMC with different values of the convexity-preserving parameter α to show how α impacts the performance of the group GMC.

Figure 6 displays how the performance of different methods varies with the SNR in the considered aspects. All methods achieve better coefficient estimation

and model prediction as the SNR increases. The four different methods perform comparably in coefficient estimation and model prediction across different SNR settings. When it comes to support recovery, we can see significant differences among different methods. The group Lasso suffers from a much larger number of false positives than the others, thus resulting in a much lower F1 score under all SNR settings. The numbers of true positives of both group GMC and group MCP increase as the SNR becomes higher, but their numbers of false positives are less responsive to the change in SNR. The group SCAD, however, not only sees an increase in its number of true positives but also a noticeable decrease in the number of false positives. Therefore, we can see an improvement in the F1 score for the nonconvex penalization methods as the SNR increases. The group MCP obtains a higher F1 score than the group SCAD thanks to its fewer false positives in all SNR settings. For the group GMC method, we observe that the value of α clearly affects its variable selection performance. That is, a larger value of α gives fewer true and false positives but a larger F1 score for the group GMC. An α in $(0.4, 1)$ is recommended for practical use, which makes the group GMC perform competitively with or even better than existing methods in all SNR settings.

Case II: In this case, we investigate how the group GMC as well as the existing three methods perform for correlated and uncorrelated groups. Note that the correlation among X_i and X_j is $\rho = t^2/(1 + t^2)$ for $i \neq j$. We set five different values for $t \in \{0, 0.5, 1, 2, 3\}$, thus leading to five different correlations $\rho \in \{0, 0.2, 0.5, 0.8, 0.9\}$ among each pair of groups. We fix the SNR of the regression model to be 2 and collect 100 observations for each run. We set the convexity-preserving parameter $\alpha = 0.8$ for the group GMC method.

Figure 7 summarizes the simulation results. Regarding the coefficient estimation and model prediction, there is a clear increasing trend in MSE and prediction error for all methods. The group GMC outperforms existing methods in these two aspects in high group correlation settings. In terms of variable selection, the group GMC, group MCP, and group SCAD perform worse as the correlation gets higher, while the group Lasso always produces a stable but low F1 score over different correlation settings. The group MCP achieves a slightly lower F1 score than the group GMC and leads the other two methods in low correlation settings ($\rho \leq 0.5$). Interestingly, in the two high correlation settings, the group MCP is no longer competitive with the group GMC and behaves comparably with the group Lasso and group SCAD. This suggests that the group GMC method is more robust against the effect of correlation compared to the group MCP. The bottom panel of Figure 7 shows that the group GMC and group MCP are able to keep their false positives low at different correlation settings but miss more true positives as the correlation gets higher. While for the group Lasso and group SCAD, both their true positives and false positives decrease as ρ increases.

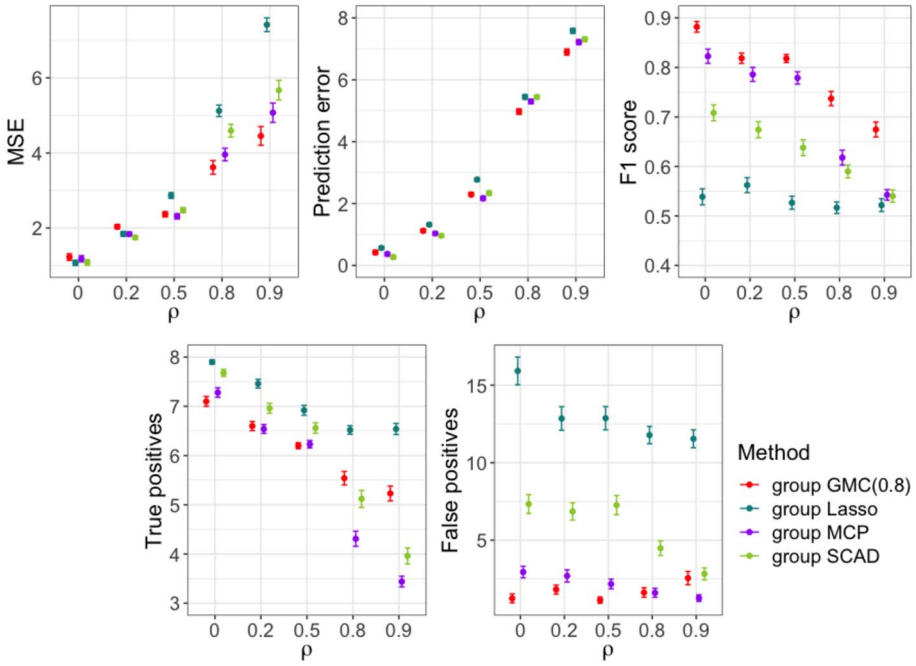

FIG 7. *Results for Case II: Impact of group correlation. Average performance plus/minus one standard error based on* 100 *simulation replicates for each method. MSE and Prediction error are on a log scale.*

Case III: Another factor of interest is how the variable selection methods perform as the problem dimension increases. To investigate this, we consider three different data generation processes. Twenty, eighty, and two hundred uncorrelated continuous covariate variables are first simulated in respective settings and then the last half of the covariates are trichotomized in the same way as described above. Therefore, the problem dimensions p are $50, 200$ and 500, respectively. The response variable y remains generated from (C.1) with an SNR of 2. We fix the sample size $n = 100$ and the convexity-preserving parameter $\alpha = 0.8$ for the group GMC method.

Figure 8 displays the behaviors of four methods in different dimension settings as described in *Case III*. For coefficient estimation and model prediction, the performance of all methods degrades as the problem dimension p increases. In high-dimensional settings where $p > n$, the group GMC achieves the best coefficient estimation. In regards to variable selection, all the existing methods tend to select more irrelevant variables into the model as p increases, thereby causing a significant deterioration in their variable selection performance, especially for the group Lasso and group SCAD. The group GMC, however, can keep the number of false positives small but have to lose some true positives as the dimension increases. Overall, the group GMC achieves the best variable selection under different dimension settings.

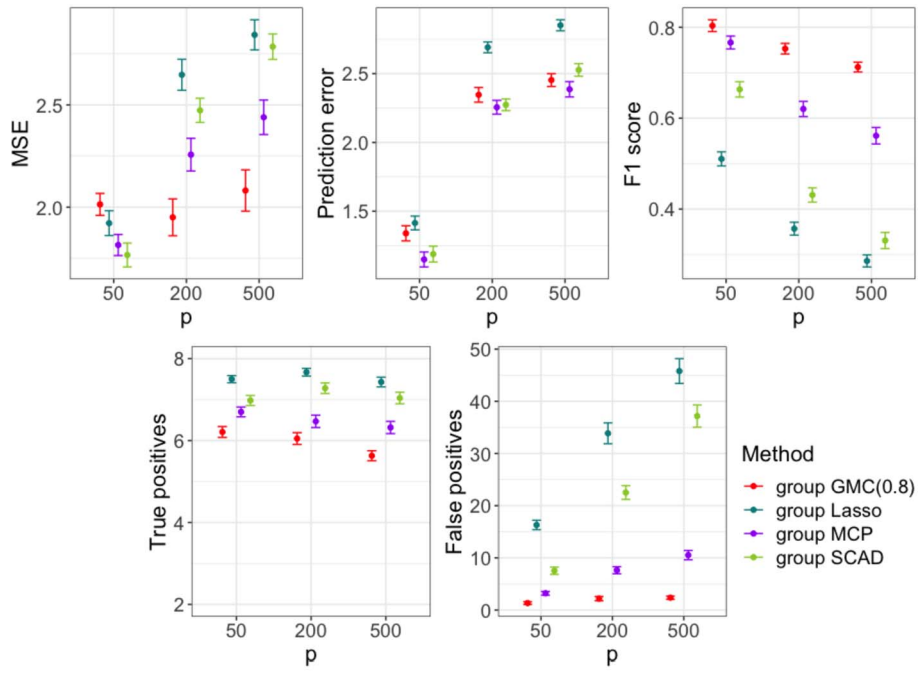

FIG 8. *Results for Case III: Impact of problem dimension. Average performance plus/minus one standard error based on* 100 *simulation replicates for each method. MSE and Prediction error are on a log scale.*

Finally, we report and compare run times of different methods for all the experiments involved in the two simulation examples. Table 3 summarizes median run times (with mean and variance in parentheses) of different methods under different SNRs. Due to space limitation, we only report run times at SNR $= \{1, 3, 5\}$. Run times of existing methods are immune to the SNR. Run times of the group GMC increase as the SNR grows, particularly for the ANOVA example. Overall, the group GMC method is less computationally efficient than the existing three methods. Table 4 displays run times of different methods under different group correlations ρ. Existing methods behave stably with the group correlation in their run times. For the ANOVA example, run times of the group GMC are less responsive to the change of ρ, while for the additive model, we see an increasing trend in the run times of the group GMC as ρ increases. Table 5 reports run times of different methods in different problem dimension settings. As anticipated, all methods require longer run times as p increases. The group GMC, unfortunately, is more sensitive to the problem dimension.

It is not surprising that the group GMC method is computationally expensive compared to the existing methods. First, the group GMC has to deal with a saddle-point problem (3.2), which is more complicated than the penalized least

TABLE 3
Median run times (with mean and variance in parentheses) of different methods under different SNRs.

Example	Method	SNR=1	SNR=3	SNR=5
ANOVA	Group Lasso	0.005 (0.005, 2.5e-6)	0.005 (0.005, 2.5e-7)	0.004 (0.005, 6.7e-7)
	Group SCAD	0.004 (0.004, 2.7e-7)	0.004 (0.004, 3.6e-7)	0.004 (0.004, 3.1e-7)
	Group MCP	0.004 (0.005, 2.9e-7)	0.004 (0.004, 5.2e-7)	0.004 (0.004, 1.8e-7)
	Group GMC (0.2)	0.006 (0.006, 2.4e-6)	0.010 (0.010, 1.3e-5)	0.015 (0.015, 1.9e-5)
	Group GMC (0.4)	0.007 (0.008, 4.9e-6)	0.019 (0.018, 4.7e-5)	0.027 (0.025, 3.4e-5)
	Group GMC (0.6)	0.010 (0.011, 2.6e-5)	0.029 (0.025, 9.7e-5)	0.034 (0.033, 3.1e-5)
	Group GMC (0.8)	0.012 (0.015, 6.0e-5)	0.033 (0.029, 1.1e-4)	0.038 (0.037, 1.9e-5)
	Group GMC (1)	0.011 (0.015, 7.1e-5)	0.029 (0.025, 7.1e-5)	0.029 (0.029, 6.4e-6)
Additive	Group Lasso	0.008 (0.008, 6.5e-7)	0.008 (0.008, 4.9e-7)	0.008 (0.008, 4.6e-7)
	Group SCAD	0.008 (0.008, 1.2e-6)	0.008 (0.008, 6.4e-7)	0.008 (0.008, 1.5e-6)
	Group MCP	0.008 (0.008, 1.7e-6)	0.008 (0.008, 9.5e-7)	0.008 (0.008, 1.3e-6)
	Group GMC (0.2)	0.010 (0.012, 8.4e-6)	0.014 (0.015, 3.9e-6)	0.016 (0.016, 2.6e-6)
	Group GMC (0.4)	0.013 (0.013, 2.0e-6)	0.014 (0.014, 2.4e-6)	0.015 (0.015, 2.1e-6)
	Group GMC (0.6)	0.014 (0.014, 5.2e-6)	0.015 (0.015, 2.5e-6)	0.016 (0.016, 2.8e-6)
	Group GMC (0.8)	0.015 (0.017, 1.8e-5)	0.016 (0.017, 9.3e-6)	0.017 (0.017, 6.3e-6)
	Group GMC (1)	0.014 (0.016, 8.3e-5)	0.014 (0.016, 4.5e-5)	0.015 (0.016, 2.6e-5)

squares problems involved in the existing methods. Second, a feature screening strategy, called SSR-BEDPP [40], was implemented in the R package `grpreg` to reduce the computational burden for the three existing methods by utilizing their KKT conditions. However, we did not adopt any screening rule for the group GMC due to the complexity of its KKT conditions. Third, there are other options for solving the group GMC problem, such as the Forward-Backward algorithm and the Douglas–Rachford algorithm, which could be more efficient than PDHG. We leave the computation of the group GMC for future work.

Acknowledgments

The authors are grateful to the editor, the associate editor, and the two referees for their helpful comments and suggestions.

TABLE 4
Median run times (with mean and variance in parentheses) of different methods under different group correlations.

Example	ρ	Group Lasso	Group SCAD	Group MCP	Group GMC
ANOVA	0	0.005 (0.007, 1.9e-4)	0.005 (0.005, 8.2e-7)	0.005 (0.005, 2.9e-7)	0.028 (0.027, 1.5e-4)
	0.2	0.005 (0.005, 3.0e-7)	0.005 (0.005, 5.7e-7)	0.005 (0.005, 5.3e-6)	0.019 (0.022, 1.6e-4)
	0.4	0.005 (0.005, 5.2e-7)	0.005 (0.005, 6.9e-7)	0.005 (0.005, 5.3e-7)	0.017 (0.022, 1.6e-4)
	0.6	0.005 (0.005, 1.3e-7)	0.005 (0.005, 7.8e-7)	0.005 (0.005, 2.5e-7)	0.013 (0.017, 1.1e-4)
	0.8	0.005 (0.005, 3.4e-7)	0.005 (0.005, 2.2e-6)	0.005 (0.005, 5.2e-7)	0.015 (0.020, 1.9e-4)
Additive	0	0.008 (0.008, 5.3e-7)	0.007 (0.007, 6.9e-7)	0.007 (0.007, 4.9e-7)	0.014 (0.015, 1.1e-5)
	0.2	0.007 (0.008, 1.1e-6)	0.007 (0.007, 2.6e-7)	0.007 (0.007, 4.8e-7)	0.022 (0.022, 2.9e-5)
	0.5	0.008 (0.008, 1.6e-6)	0.007 (0.007, 2.3e-7)	0.007 (0.007, 1.7e-6)	0.071 (0.070, 1.1e-4)
	0.8	0.007 (0.007, 2.7e-7)	0.007 (0.007, 7.2e-7)	0.007 (0.007, 1.7e-7)	0.086 (0.077, 5.4e-4)
	0.9	0.008 (0.008, 5.5e-7)	0.007 (0.008, 2.3e-6)	0.007 (0.007, 9.3e-7)	0.093 (0.085, 4.6e-4)

TABLE 5
Median run times (with mean and variance in parentheses) of different methods under different problem dimensions.

Example	p	Group Lasso	Group SCAD	Group MCP	Group GMC
ANOVA	32	0.005 (0.005, 8.2e-7)	0.004 (0.004, 3.7e-7)	0.004 (0.005, 7.0e-7)	0.024 (0.021, 1.0e-4)
	200	0.019 (0.019, 2.7e-6)	0.018 (0.019, 2.4e-6)	0.018 (0.019, 2.7e-6)	0.238 (0.337, 0.044)
	512	0.040 (0.041, 9.4e-6)	0.042 (0.044, 1.7e-4)	0.041 (0.041, 7.8e-6)	2.307 (3.113, 4.869)
Additive	50	0.007 (0.008, 1.0e-6)	0.007 (0.007, 2.3e-7)	0.007 (0.007, 2.4e-7)	0.014 (0.015, 1.0e-5)
	200	0.025 (0.026, 3.1e-6)	0.025 (0.026, 3.9e-6)	0.024 (0.025, 4.2e-6)	0.093 (0.121, 0.004)
	500	0.060 (0.062, 2.8e-5)	0.059 (0.060, 1.1e-5)	0.063 (0.064, 1.2e-4)	0.630 (0.986, 0.367)

Funding

Aaron J. Molstad's research was supported by NSF DMS-2113589. Eric C. Chi's research was supported by NSF DMS-2201136.

References

[1] ABE, J., YAMAGISHI, M. and YAMADA, I. (2019). Convexity-edge-preserving signal recovery with linearly involved generalized minimax con-

cave penalty function. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 4918–4922. IEEE.

[2] Abe, J., Yamagishi, M. and Yamada, I. (2020). Linearly involved generalized Moreau enhanced models and their proximal splitting algorithm under overall convexity condition. *Inverse Problems* **36** 035012. MR4068241

[3] Bauschke, H. H. and Combettes, P. L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* **408**. Springer. MR2798533

[4] Bayram, I. (2015). On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty. *IEEE Transactions on Signal Processing* **64** 1597–1608. MR3548876

[5] Blake, A. and Zisserman, A. (1987). *Visual Reconstruction*. MIT Press. MR0919733

[6] Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing* **25** 173–187. MR3306699

[7] Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision* **40** 120–145. MR2782122

[8] Chen, Y., Yamagishi, M. and Yamada, I. (2023). A unified design of generalized Moreau enhancement matrix for sparsity aware LiGME models. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 2022EAP1118.

[9] Esser, E., Zhang, X. and Chan, T. (2009). A general framework for a class of first order primal-dual algorithms for TV minimization. *UCLA CAM Report* 09–67.

[10] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

[11] Fan, J., Xue, L. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics* **42** 819. MR3210988

[12] Goldstein, T., Li, M. and Yuan, X. (2015). Adaptive primal-dual splitting methods for statistical learning and image processing. *Advances in Neural Information Processing Systems* **28** 2089–2097.

[13] Goldstein, T., Studer, C. and Baraniuk, R. (2014). A field guide to forward-backward splitting with a FASTA implementation. arXiv eprint arXiv:1411.3406.

[14] Goldstein, T., Studer, C. and Baraniuk, R. (2015). FASTA: A feneralized implementation of forward-backward splitting. arXiv preprint arXiv:1501.04979.

[15] Goldstein, T., Li, M., Yuan, X., Esser, E. and Baraniuk, R. (2013). Adaptive primal-dual hybrid gradient methods for saddle-point problems. arXiv preprint arXiv:1305.0546.

[16] Hastie, T., Tibshirani, R. and Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on Statistics and Applied Probability* **143** 143. MR3616141

[17] Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley & Sons.
[18] Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science* **27** 481–499. MR3025130
[19] Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in Statistics* 492–518. Springer.
[20] Kim, Y., Kim, J. and Kim, Y. (2006). Blockwise sparse regression. *Statistica Sinica* **16** 375–390. MR2267240
[21] Lanza, A., Morigi, S., Selesnick, I. W. and Sgallari, F. (2019). Sparsity-inducing nonconvex nonseparable regularization for convex image processing. *SIAM Journal on Imaging Sciences* **12** 1099–1134. MR3968243
[22] Liu, X. and Chi, C. E. (2022). Revisiting convexity-preserving signal recovery with the linearly involved GMC penalty. *Pattern Recognition Letters* **156** 60–66.
[23] Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research* **16** 559–616. MR3335800
[24] Meier, L., Van De Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70** 53–71. MR2412631
[25] Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* **27** 538–557. https://doi.org/10.1214/12-STS400. MR3025133
[26] Nikolova, M. (1998). Estimation of binary images by minimizing convex criteria. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)* **2** 108–112. IEEE.
[27] Nikolova, M., Ng, M. K. and Tam, C.-P. (2010). Fast nonconvex nonsmooth minimization methods for image restoration and reconstruction. *IEEE Transactions on Image Processing* **19** 3073–3088. MR2789705
[28] Rockafellar, R. T. (1970). *Convex Analysis. Princeton Mathematical Series*. Princeton University Press, Princeton, N.J. MR0274683
[29] Selesnick, I. (2017a). Total variation denoising via the Moreau envelope. *IEEE Signal Processing Letters* **24** 216–220.
[30] Selesnick, I. (2017b). Sparse regularization via convex analysis. *IEEE Transactions on Signal Processing* **65** 4481–4494. MR3684078
[31] Selesnick, I., Lanza, A., Morigi, S. and Sgallari, F. (2020). Nonconvex total variation regularization for convex denoising of signals. *Journal of Mathematical Imaging and Vision* **62** 825–841. MR4126452
[32] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. MR1379242
[33] Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics* **7** 1456–1490. https://doi.org/10.1214/13-EJS815. MR3066375

[34] Wainwright, M. J. (2019). *High-dimensional Statistics: A Non-asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press. https://doi.org/10.1017/9781108627771. MR3967104

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

[36] Wang, L., Li, H. and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association* **103** 1556–1569. MR2504204

[37] Wei, F. and Huang, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli* **16** 1369–1384. MR2759183

[38] Yata, W., Yamagishi, M. and Yamada, I. (2022). A constrained LIGME model and its proximal splitting algorithm under overall convexity condition. *Journal of Applied & Numerical Optimization* **4**.

[39] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67. MR2212574

[40] Zeng, Y., Yang, T. and Breheny, P. (2021). Hybrid safe–strong rules for efficient optimization in lasso-type problems. *Computational Statistics & Data Analysis* **153** 107063. MR4146815

[41] Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942. MR2604701

[42] Zhang, C.-H. and Zhang, T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science* **27** 576–593. MR3025135

[43] Zhao, P., Rocha, G., Yu, B. et al. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* **37** 3468–3497. MR2549566

[44] Zou, J., Shen, M., Zhang, Y., Li, H., Liu, G. and Ding, S. (2018). Total variation denoising with non-convex regularizers. *IEEE Access* **7** 4422–4431.