Statistical Methods & Applications (2024) 33:1251–1313
https://doi.org/10.1007/s10260-024-00768-8

ORIGINAL PAPER

Group penalized expectile regression

Mohamed Ouhourane[1] · **Karim Oualkacha**[1] · **Archer Yi Yang**[2]

Accepted: 27 October 2024 / Published online: 20 November 2024

Abstract

The asymmetric least squares regression (or expectile regression) allows estimating unknown expectiles of the conditional distribution of a response variable as a function of a set of predictors and can handle heteroscedasticity issues. High dimensional data, such as omics data, are error prone and usually display heterogeneity. Such heterogeneity is often of scientific interest. In this work, we propose the Group Penalized Expectile Regression (GPER) approach, under high dimensional settings. GPER considers implementation of sparse expectile regression with group Lasso penalty and the group non-convex penalties. However, GPER may fail to tell which groups variables are important for the conditional mean and which groups of variables are important for the conditional scale/variance. To that end, we further propose a COupled Group Penalized Expectile Regression (COGPER) regression which can be efficiently solved by an algorithm similar to that for solving GPER. We establish theoretical properties of the proposed approaches. In particular, GPER and COGPER using the SCAD penalty or MCP is shown to consistently identify the two important subsets for the mean and scale simultaneously. We demonstrate the empirical performance of GPER and COGPER by simulated and real data.

Keywords Expectile · Regression · Lasso · Heterogeneity

Karim Oualkacha and Archer Yi Yang have contributed equally to this work.

✉ Mohamed Ouhourane
Mohamed.ouhourane@gmail.com

Karim Oualkacha
oualkacha.karim@uqam.ca

Archer Yi Yang
archer.yang@mcgill.ca

[1] Department of Mathematics, Université du Québec à Montréal, 201, Ave Président-Kennedy, Montreal, QC H2X 3Y7, Canada

[2] Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, QC H3A 0B9, Canada

1 Introduction

Sparse regression methods, which use penalization techniques for both estimation and variable selection, have been introduced as a mainstream approach for analyzing high-dimensional data. Popular penalized estimators are the l_1-type selectors such as the Lasso (Tibshirani 1996) and Dantzig estimators (Candes and Tao 2007), and the non-convex penalized estimators such as the Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li 2001) and the Minimax Concave Penalty (MCP) (Zhang 2010) estimators. L1-type selectors are useful due to their computational efficiency and the non-convex selectors are known to enjoy the oracle property. Several computationally efficient algorithms have also been proposed for computing the non-convex estimators. Zou and Li (2008) worked out the Local Linear Approximation (LLA) algorithm, which approximates the non-convex penalties using a series of reweighted l_1 penalization.

In many situations, it is suitable to perform selection of a group/set of predictors sharing a common function (e.g., genes participate in a common biological function or pathway; methylation levels in nearby positions along the genome present high spatial correlation). Capturing group-variable effects can improve the outcome prediction. Another attractive motivation of the group-variable selection methods is the additive model with polynomial or non-parametric components, thereby each component/group may be expressed as a linear combination of basis functions of the original variables. In this context, the selection of important variables corresponds to the selection of groups of basis functions. Yuan and Lin (2006) have proposed group-Lasso as an extension of the Lasso for achieving group-wise variable selection. Although theoretical consistency can be achieved by the Lasso and group-Lasso estimators if one assumes some regularity assumptions on the design matrix [e.g. the restricted eigenvalue or compatibility conditions (Bickel et al. 2009; Meier et al. 2008)], in general, both estimators introduce bias for the model parameter estimation in high dimensions. To reduce bias and achieve oracle properties, Wei and Zhu (2012) and Ogutu and Piepho (2014) have introduced extensions of non-convex penalties (SCAD, MCP) for group-variable selection.

Recent advances in data collection from multi-sources in many areas of research such as genomics, economics, and finance, generate an error accumulation in data pre-processing, and the assumption of homoscedasticity does not hold. To remedy this problem, flexible methods, which incorporate heteroscedasticity in modelling such data, are necessary to take into account the specificity of the collected datasets (Wang et al. 2012). In the standard regression, the conditional mean function is explained by a linear combination of the predictors and its estimation results from minimizing a squared error loss function, which assigns equal weights to the residuals. On the opposite, when different weights are assigned to residuals, an exhaustive description of the outcome conditional distribution can be explored. Newey and Powell (1987) have introduced the Expectile Regression (ER) in which a squared error loss function puts different weights on the residuals, depending on their signs. Like the quantile regression (Koenker and

Bassett Jr 1978), ER is appropriate to detect heteroscedasticity since both methods use an asymmetric loss function to estimate the regression function linking the outcome to the predictors.

Inspired by the success of sparse quantile regression (Mkhadri et al. 2017), many advances have been made on variable selection in ER under high dimensional settings. For instance, Zhao and Zhang (2018) studied penalized ER with the SCAD penalty. Gu and Zou (2016) developed a unified and efficient algorithm, which fits ER with the Lasso penalty and uses the LLA approximation to handle the non-convex penalties SCAD and MCP. Gu and Zou (2016) have also established the estimation consistency of the Lasso selector under the restricted eigenvalue condition and the generalized invertability factor (GIF) condition (Ye and Zhang 2010; Huang and Zhang 2012), and proved the convergence of LLA algorithm to the oracle estimator in two steps. Moreover, Gu and Zou (2016) have developed the oracles properties for their proposed estimators under the assumption that the model errors follow a sub-Gaussian distribution. Liao et al. (2019) provided asymptotic distributions of penalized expectile regression with SCAD and adaptive Lasso penalties for both independent and identically distributed (i.i.d.) and non-i.i.d. random errors. Furthermore, penalized ER approaches have been introduced in the context of semi- and non-parametric methods where the penalty is used to impose smoothness for non-parametric estimators (Jiang et al. 2017; Sobotka et al. 2013; Yang et al. 2018; Yang and Zou 2015).

When dealing with heteroscedastic high dimensional data, it is of interest to distinguish which variables are significant for the conditional mean and which ones are important for the conditional scale/variance of the outcome, particularly when some variables are important for both the mean and the scale. For instance, most of genomics data display heterogeneity due to either heteroscedastic variance or other forms of non-location-scale covariate effects. Body mass index (BMI) is a classical illustration in this context. Its distribution varies with age and obesity genetic risk factors (i.e., high-dimensional genetic variants) at different quantiles, and several genetic variants have shown more strong association with BMI at the upper tail compared to the lower tail of the conditional distribution (Bottai et al. 2014; Mitchell et al. 2013). In low dimension, Efron (1991) have proposed a method, which combines both symmetric and asymmetric least squares loss functions, to differentiate the effects of the important variables for both the mean and the scale simultaneously. Such a resulting combined loss function has also been studied in high dimensional settings by Gu and Zou (2016) as an extension of the ER approach. The authors have established theoretical proprieties and efficient algorithms for the proposed estimators, and termed the approach the COupled Sparse Asymmetric LEast Squares regression, COSALES for short.

In this paper, we address the challenge of selecting grouped variables (factors) in presence of heteroscedastic high dimensional data, in situations where the groups of predictors influence either the conditional mean, the variance of the outcome, or both. To this end, we develop the methodology and theory for accurate prediction in group penalized ER and Coupled ER. We extend the computational algorithms and the consistency results of Gu and Zou (2016) from Lasso and non-convex penalties to group Lasso and group non-convex penalties. First, we propose a bloc coordinate

descent (BCD) algorithm for group Lasso and group non-convex penalties, which uses an efficient approach to minimize each sub-problem exactly. Moreover, we propose the group local linear approximation (GLLA) algorithm as an alternative approach for solving ER with the non-convex penalties SCAD and MCP. Yet, we demonstrate that if the GLLA algorithm starts with a reasonable initial estimator, we obtain the oracle estimator in a one-step iteration. Finally, we derive necessary conditions for the consistency of our ER and Coupled ER estimators by adapting both the generalized invertability factor (GIF) and the compatibility condition (Bühlmann and Van De Geer 2011) to the group variable selection context.

The plan of the paper is as follows: in Sect. 2, we briefly review ER and we present our approach termed the Group Penalized Expectile Regression (GPER). In Sect. 3, we present our Coupled GPER framework. Evaluation of the performance of our methods through exhaustive simulation studies is considered in Sect. 4. The use of the proposed methodology is illustrated by analysing real datasets, in Sect. 5. Discussion is given in Sect. 6. All the proofs are postponed to the Appendix.

2 Expectile regression and group penalizations

2.1 Overview of the unconditional expectile

The τ-mean (or τ-expectile) of a continuous random variable Y is defined as the solution of the following problem

$$\mathscr{E}^{\tau}(Y) = \arg\min_{\mathscr{E}\in\mathbb{R}} \mathbb{E}\{\rho_{\tau}(Y-\mathscr{E})\}, \quad \tau \in (0,1), \tag{1}$$

where

$$\rho_{\tau}(u) := |\tau - \mathbb{1}_{(u\leq 0)}|u^2 \tag{2}$$

is known as the asymmetric square loss function, which assigns weights τ and $1-\tau$ to positive and negative deviations, respectively.

By equating the first derivative of (1) to zero, one has

$$\mathbb{E}\{\psi_{\tau}(Y-\mathscr{E}^{\tau})(Y-\mathscr{E}^{\tau})\} = 0, \tag{3}$$

where $\psi_{\tau}(u) := |\tau - \mathbb{1}_{(u\leq 0)}|$ is the check function. The solution of (3) leads to a more meaningful definition of the τ-mean, which is given as follows

$$\mathscr{E}^{\tau}(Y) = \mathbb{E}\left[\frac{\psi_{\tau}(Y-\mathscr{E}^{\tau})}{\mathbb{E}\left[\psi_{\tau}(Y-\mathscr{E}^{\tau})\right]}Y\right].$$

When $\tau = 0.5$, $\psi_{0.5}(u) = 0.5$ and $\mathscr{E}^{\tau}$ reduces to the mean of Y, (i.e. $\mathscr{E}^{0.5}(Y) = \mathbb{E}[Y]$). Thus, the τ-expectile can be viewed as a generalization of the mean, and like the mean, $\mathscr{E}^{\tau}$ is a weighted average with random weights. By varying τ, the τ-expectile provides insight at different "locations" of the distribution of Y and thus it is an alternative measure of "locations" of the distribution.

Given a random sample, $\{(y_i)\}_{i=1}^{n}$, the τ-th empirical expectile

$$\widehat{\mathscr{E}}_\tau = \sum_{i=1}^{n} \frac{\psi_\tau(y_i - \widehat{\mathscr{E}}_\tau)}{\sum_{i=1}^{n} \psi_\tau(y_i - \widehat{\mathscr{E}}_\tau)} y_i$$

is the solution that minimizes the empirical loss function

$$\frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \mathscr{E}).$$

The extension of the expectile concept to regression has been investigated by Newey and Powell (1987). Let $\{(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)\}$ be an observed data, where y_i is the observed response and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is a p-dimensional observed vector of predictors for subject $i = 1, \ldots, n$. Note $\mathbf{X}$ the design matrix with n rows and p columns. If an intercept is used in the model, we let the first column of $\mathbf{X}$ be a vector of $\mathbf{1}$. The ER model uses the weighted least squares loss $\rho_\tau(u)$ given in (2) to assign different weights to negative and positive residuals, and assumes that conditional τ-expectile given the predictors, denoted as $\mathscr{E}^\tau(\mathbf{x}_i)$, is a linear function of $\mathbf{x}_i$ (i.e. $\mathscr{E}^\tau(\mathbf{x}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}_\tau$). This leads to the following estimator of the regression coefficients

$$\hat{\boldsymbol{\beta}}_\tau = \arg\min_{\beta_\tau} \left(\Psi_\tau(\boldsymbol{\beta}_\tau) := \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_\tau) \right). \tag{4}$$

Again, when $\tau = 0.5$, the ER model reduces to the ordinary least squares regression.

Several theoretical properties of the ER model have been established under some assumptions about the random error term of the regression model (Newey and Powell 1987)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_\tau + \boldsymbol{\epsilon}_\tau, \tag{5}$$

where $\boldsymbol{\epsilon}_\tau$ is the vector of n independent errors, which satisfies $\mathscr{E}^\tau(\boldsymbol{\epsilon}_\tau | \mathbf{X}) = \mathbf{0}$ for some $\tau \in (0, 1)$. Therefore $\mathscr{E}^\tau(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}_\tau$, which is to say that the conditional τ-mean of $\mathbf{y}$ is a linear combination of the columns of $\mathbf{X}$. In the ER model, the estimated coefficients $\boldsymbol{\beta}_\tau$ vary as a function of τ, which makes modeling of different "locations" of the conditional distribution possible, and as a consequence heteroscedasticity when it exists, can be investigated by this model. For ease of notation, the subscript in $\boldsymbol{\beta}_\tau$ and $\boldsymbol{\epsilon}_\tau$ is dropped hereafter.

In this work, we focus on the ER model (5) with a pre-defined group structure, i.e. we assume that there is a natural grouping of the regression predictors. We assume that the predictors $x_1, x_2 \ldots x_p$ are put into K groups $(\{1, 2, 3 \ldots p\} = \cup_{k=1}^{K} I_k)$, such that the size of each group is p_k (the cardinality of index set I_k is p_k) and the groups are non-overlapping ($I_k \cap I_{k'} =$ for $k \neq k'$). This leads to the block representation of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top)^\top$.

In general, the GPER model in high dimensions can be formulated as a minimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(R_\tau(\boldsymbol{\beta}) := \Psi_\tau(\boldsymbol{\beta}) + \sum_{k=1}^{K} w_k P_\lambda(\|\boldsymbol{\beta}_k\|_2) \right), \tag{6}$$

with $(\hat{\boldsymbol{\beta}}_k)_{k=1,\ldots,K}$ the sub-vector of $\hat{\boldsymbol{\beta}}$ corresponding to the effects of the predictors belonging to group k for the τ-expectile of the response. $P_\lambda(\cdot)$ is the penalty function with a regularization parameter λ, and w_k is used to adjust for the group sizes in the penalty. A reasonable choice is $w_k = \sqrt{p_k}$. This choice is crucial because it balances the contribution of different groups to the penalty term. Without these weights, groups with a larger number of variables might be unfairly advantaged and more likely to be selected, simply due to their size. This could lead to biased model selection, where larger groups are favored regardless of their true importance (Yuan and Lin 2006). If the intercept is included in (6), then $w_1 = 0$ is taken, which means that the first group is not penalized.

In this work, we consider the group Lasso (GLasso), group MCP (GMCP) and group SCAD (GSCAD) penalties which are defined respectively by the penalty function, $P_\lambda(t)$, as follows

$$\lambda t, \tag{7}$$

$$\begin{cases} (\lambda t - \dfrac{t^2}{2\theta}) & \text{if} \quad 0 \le t \le \theta\lambda, \\ \frac{1}{2}\lambda^2\theta & \text{if} \quad t \ge \theta\lambda, \end{cases} \tag{8}$$

$$\begin{cases} \lambda t & \text{if} \quad 0 \le t \le \lambda, \\ \dfrac{\theta\lambda t - (t^2+\lambda^2)/2}{\theta - 1} & \text{if} \quad \lambda \le t \le \theta\lambda, \\ \dfrac{\lambda^2(\theta^2-1)}{2(\theta-1)} & \text{if} \quad t \ge \theta\lambda, \end{cases} \tag{9}$$

where θ is a second tuning parameter of GMCP and GSCAD penalties, with $\theta > 1$ for GMCP and $\theta > 2$ for GSCAD. In this work, we set $\theta = 4$ for GSCAD and $\theta = 3$ for GMCP, which are suggested values for this tuning parameter. In fact, θ serves as a secondary tuning parameter for both the GMCP and GSCAD penalties, with the constraint that $\theta > 1$ for GMCP and $\theta > 2$ for GSCAD. Optimal values for θ have been examined in the literature, and fixed values like $\theta = 4$ for GSCAD and $\theta = 3$ for GMCP are often recommended for a wide range of problems. However, performance improvements using data-driven methods to select θ are typically minimal. For more details about optimal values of θ see Fan and Li (2001) and Ogutu and Piepho (2014)). Consequently, we adopt these recommended values for θ in all our simulations and real data analyses.

The non-convex penalties (8) and (9) enjoy the oracle property (Fan and Li 2001; Fan and Peng 2004), which means that they achieve the asymptotic equivalence to

the ideal non-penalized estimator (oracle estimator) whose coefficients of irrelevant groups of variables equal to zero in advance. That is, the GSCAD and GMCP estimators can perform as well as the oracle estimator if the penalization parameter is appropriately chosen. For the GPER regression, the oracle estimator is defined by

$$\hat{\boldsymbol{\beta}}^{oracle} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p:\boldsymbol{\beta}_{\mathcal{A}^c}=\mathbf{0}} \boldsymbol{\Psi}_\tau(\boldsymbol{\beta}), \tag{10}$$

where $\mathcal{A}$ is the true support set.

The main difficulty of solving the optimization problem (6) is that the loss function $\rho_\tau(.)$ in (4) does not have the second derivative everywhere. To overcome this problem, we adopt the Majorization-Minimization principle (MM) and the block coordinate descent (BCD) algorithm to find the optimal solution by iteratively minimizing a surrogate function that majorizes the objective function in (4), for each group (i.e. block-/group-wise minimization) (Yang et al. 2018; Ouhourane et al. 2021). In fact, the penalty $\sum_{k=1}^{K} w_k P_\lambda(\|\boldsymbol{\beta}_k\|_2)$ in (6) is group-wise separable. This property is used to make group-wise update in each iteration over one group of variables k ($k = 1, \ldots, K$). This technical resolution is detailed next.

2.2 GPER algorithm

This section gives details about the group-wise descent algorithm for the expectile regression with GLasso, GMCP and GSCAD penalties.

Let $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_{k-1}, \tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{k+1}, \ldots, \tilde{\boldsymbol{\beta}}_K)$ be the current iteration and $\tilde{\boldsymbol{\beta}}_{-k} = (\tilde{\boldsymbol{\beta}}_1, \ldots, \tilde{\boldsymbol{\beta}}_{k-1}, \tilde{\boldsymbol{\beta}}_{k+1}, \ldots, \tilde{\boldsymbol{\beta}}_K)$ be the current iterate with k^{th} group excluded. Assume we are about to update the effects of the k^{th} group $\boldsymbol{\beta}_k = (\beta_1, \ldots, \beta_{p_k})^\top$ for some $k \in \{1, \ldots, K\}$. Also, consider both the objective function $R_\tau(\boldsymbol{\beta})$ in (6) and the ER loss function in (4) as functions of the k^{th} group, $\boldsymbol{\beta}_k$, while keeping all other groups fixed at $\tilde{\boldsymbol{\beta}}_{-k}$, i.e., $R_\tau(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) = R_\tau(\boldsymbol{\beta})_{\beta_{k'}=\tilde{\beta}_{k'}, 1\le k'\le K, k'\ne k}$ and $\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) = \boldsymbol{\Psi}_\tau(\boldsymbol{\beta})_{\beta_{k'}=\tilde{\beta}_{k'}, 1\le k'\le K, k'\ne k}$.

The following proposition summarizes the quadratic majorization property for $R_\tau(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$, which leads to solve our problem efficiently for each group k.

Proposition 1 *Let $\mathbf{X}_k$ be the sub-matrix of $\mathbf{X}$ corresponding to group k. The quadratic majorization condition is satisfied by the function $\Psi_\tau(.)$. That is, for all $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ we have*

$$\begin{aligned} R_\tau(\beta_k, \tilde{\beta}_{-k}) \le & Q(\beta_k, \tilde{\beta}_{-k}) := \Psi_\tau(\tilde{\beta}_k, \tilde{\beta}_{-k}) + (\beta_k - \tilde{\beta}_k)^T \nabla_k \Psi_\tau(\tilde{\beta}_k, \tilde{\beta}_{-k}) \\ & + \frac{\gamma_k}{2}(\beta_k - \tilde{\beta}_k)^T(\beta_k - \tilde{\beta}_k) + w_k P_\lambda(\|\beta_k\|_2), \end{aligned} \tag{11}$$

where γ_k is the largest eigenvalue of the matrix $\mathbf{H}_k = c\dfrac{\mathbf{X}_k^\top \mathbf{X}_k}{n}$, with $c = 2\max(1-\tau, \tau)$.

The proof of Proposition (1) is detailed in Appendix 1.

Replacing the penalty term $P_\lambda(\|\boldsymbol{\beta}_k\|_2)$ by (7), (8) or (9) in (11) leads to a closed form solution of the update, $\tilde{\boldsymbol{\beta}}_k^{\text{new}}$, for the three penalties. The following proposition summarizes these results.

Proposition 2 *Let $Q(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$ be the surrogate function given by (11) and let $P_\lambda(\|\boldsymbol{\beta}_k\|_2)$ be one of the tree penalties given in (7), (8) and (9). The closed form solution to (11) of $\tilde{\boldsymbol{\beta}}_k^{\text{new}}$ for GPER algorithm with GLasso, GMCP and GSCAD penalties is given respectively by*

$$\tilde{\boldsymbol{\beta}}_k^{\text{new}} = F(\mathbf{Z}_k) \longleftarrow \frac{1}{\gamma_k} \frac{S(\|\mathbf{Z}_k\|_2, \lambda w_k)}{\|\mathbf{Z}_k\|_2} \mathbf{Z}_k, \tag{12}$$

$$\tilde{\boldsymbol{\beta}}_k^{\text{new}} = F(\mathbf{Z}_k) \longleftarrow \begin{cases} \frac{1}{\gamma_k - w_k/\theta} \frac{S(\|\mathbf{Z}_k\|_2, \lambda w_k)}{\|\mathbf{Z}_k\|_2} \mathbf{Z}_k, & \text{if } \|\mathbf{Z}_k\|_2 \le \gamma_k \theta \lambda \\ \frac{1}{\gamma_k} \mathbf{Z}_k, & \text{if } \|\mathbf{Z}_k\|_2 > \gamma_k \theta \lambda, \end{cases} \tag{13}$$

$$\tilde{\boldsymbol{\beta}}_k^{\text{new}} = F(\mathbf{Z}_k) \longleftarrow \begin{cases} \frac{1}{\gamma_k} \frac{S(\|\mathbf{Z}_k\|_2, \lambda w_k)}{\|\mathbf{Z}_k\|_2} \mathbf{Z}_k, & \text{if } \|\mathbf{Z}_k\|_2 \le (w_k + \gamma_k)\lambda \\ \frac{S(\|\mathbf{Z}_k\|_2, \frac{\lambda w_k \theta}{\theta - 1})}{(\gamma_k - \frac{w_k}{\theta - 1})\|\mathbf{Z}_k\|_2} \mathbf{Z}_k, & \text{if } (w_k + \gamma_k)\lambda < \|\mathbf{Z}_k\|_2 \le \gamma_k \theta \lambda \\ \frac{1}{\gamma_k} \mathbf{Z}_k, & \text{if } \|\mathbf{Z}_k\|_2 > \gamma_k \theta \lambda, \end{cases} \tag{14}$$

where $\mathbf{Z}_k = \mathbf{U}_k^\tau + \gamma_k \tilde{\boldsymbol{\beta}}_k$, $\mathbf{U}_k^\tau = -\nabla_k \Psi_\tau(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k})$ and $S(.)$ is the soft-thresholding operator given by

$$S(z, \lambda) = \begin{cases} z - \lambda \text{ if } z > \lambda \\ 0 \text{ if } |z| \le \lambda \\ z + \lambda \text{ if } z < -\lambda. \end{cases}$$

The proof of Proposition (2) is detailed in Appendix 2.

The following algorithm gives some details about the groupwise descent algorithm for GPER with GLasso, GMCP and GSCAD penalties:

Algorithm 1 The GPER algorithm for GLasso/GMCP/GSCAD penalties

1: Initialize $\tilde{\boldsymbol{\beta}}$
2: **repeat**
3: **for** $k = 1$ to K **do**
4: $\tilde{\boldsymbol{\beta}}_k^{\text{new}} \leftarrow F(\mathbf{Z}_k)$
5: **end for**
6: **until** Convergence of $\tilde{\boldsymbol{\beta}}$
7: **Return** $\tilde{\boldsymbol{\beta}}$

2.3 ER with group local linear approximation (GLLA) penalty

Proposition 2 allows us to provide an explicit solution for two important special non-convex penalty functions (GMCP and GSCAD). In this section, we propose to extend the local linear approximation trick to solve ER for a more general form of non-convex penalties. We restrict our theory development in Sect. 2.6 for a class of non-convex penalties that satisfy certain conditions. This class includes GMCP and GSCAD.

The GLLA approximation is based on first order Taylor expansion of the non-convex penalty functions around $\|\tilde{\boldsymbol{\beta}}_k\|_2$. Thus, one can write

$$P_\lambda(\|\boldsymbol{\beta}_k\|_2) \approx P_\lambda(\|\tilde{\boldsymbol{\beta}}_k\|_2) + P'_\lambda(\|\tilde{\boldsymbol{\beta}}_k\|_2)(\|\boldsymbol{\beta}_k\|_2 - \|\tilde{\boldsymbol{\beta}}_k\|_2). \tag{15}$$

Substituting (15) in (6) leads to the following GPER problem with the Group LLA (GLLA) penalty

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\Psi_\tau(\boldsymbol{\beta}) + \sum_{k=1}^{K} w'_k \|\boldsymbol{\beta}_k\|_2 \right), \tag{16}$$

where $w'_k = w_k P'_\lambda(\|\tilde{\boldsymbol{\beta}}_k\|_2)$ for $k = 1, \ldots, K$. The weight w'_k depends on the non-convex penalty function through the first derivative $P'_\lambda(\|\tilde{\boldsymbol{\beta}}_k\|_2)$. The problem (16) can be solved using a GPER-GLasso update similar to the algorithm described in Sect. 2.2.

The details of the GPER approach with GLLA penalty is described in the following algorithm.

Algorithm 2 The GPER algorithm with GLLA penalty

Initialize $i = 0$; $\tilde{\boldsymbol{\beta}}^i = \tilde{\boldsymbol{\beta}}^{initial}$;
Compute the weights $\tilde{w}_k^i = w_k P'_\lambda(\|\tilde{\boldsymbol{\beta}}_k^i\|_2)$ for $k = 1, \ldots, K$;
repeat
  $i \leftarrow i + 1$;
  Solve the optimization problem:

$$\tilde{\boldsymbol{\beta}}^i = \arg\min_{\boldsymbol{\beta}} \left(\Psi_\tau(\boldsymbol{\beta}) + \sum_{k=1}^{K} \tilde{w}_k^{i-1} \|\boldsymbol{\beta}_k\|_2 \right); \tag{17}$$

  $\tilde{w}_k^i = w_k P'_\lambda(\|\tilde{\boldsymbol{\beta}}_k^i\|_2)$ for $k = 1, \ldots, K$;
until Convergence of $\tilde{\boldsymbol{\beta}}^i$
Return $\tilde{\boldsymbol{\beta}}$.

To solve the problem (17), we use Algorithm 1 with GLasso (GPER-GLasso) with $w_k = \tilde{w}_k^{i-1}$ for $k = 1, \ldots, K$.

Note that the GLLA penalty is a convex approximation of non-convex penalties (e.g. GMCP, GSCAD). Thus, for each fixed value of λ, GLLA allows a search of the solution in a locally convex region, and consequently it may lead to stable and smooth path solutions.

2.4 Implementation

We discuss some techniques used in our implementation to further improve the computational speed of Algorithm 1 of the GPER approach. In sparse modeling, the solution is computed by using a descending sequence $(\lambda_m)_{m=1}^M$ of λ values. To generate such a sequence, we set $M-2$ points uniformly (in the log-scale) between the starting and ending points, $\lambda_{\max}$ and $\lambda_{\min}$, where $\lambda_{\max}$ is the smallest λ to let all groups $\boldsymbol{\beta}_k$ to be zero ($2 \le k \le K$), except the intercept. To determine $\lambda_{\max}$, firstly, we initially estimate the intercept β_0 by considering the null model:

$$\widehat{\beta}_0 = \arg\min_{\beta_0} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0). \tag{18}$$

Subsequently, according to the KKT conditions of (18), we can obtain the following formula

$$\lambda_{\max} = \max_{k=2,\ldots,K} \frac{\|\nabla_k \boldsymbol{\Psi}_\tau(\widehat{\beta}_0, \mathbf{0})\|_2}{\omega_k}.$$

We take $\lambda_{\min} = \upsilon \lambda_{\max}$ and we set the default value of υ to be 10^{-2} for data with $n > p$ and $\upsilon = 10^{-4}$ for data with $n \le p$. We also adopt the warm-start trick to implement the solution paths along λ values (i.e. assume that we have already computed the solution $\tilde{\boldsymbol{\beta}}_k^{(m)}$ $(k = 1, \ldots, K)$ at λ_m, then $\tilde{\boldsymbol{\beta}}_k^{(m)}$ will be used as the initial value for computing the solution at λ_{m+1} in Algorithm 1. We refer readers to Ouhourane et al. (2021) and Yang and Zou (2015) for more details about such computational techniques.

2.5 Theory for GPER-GLasso

We assume a fixed design for the covariates. Before presenting our principal theoretical results (Theorems), some notations must be defined and some necessary results must be shown. Let $\mathcal{A} \equiv \text{supp}(\boldsymbol{\beta}^*) = \{k = 1, \ldots, K : \boldsymbol{\beta}_k^* \neq 0\}$ and $\mathcal{B} \equiv \{j = 1, \ldots, p : \beta_j^* \neq 0\}$ be the active set of the true vector of parameters $\boldsymbol{\beta}^*$. For any sequence $\{a_i\}_{i \in \mathcal{A}}$, denote $\underline{a}_{\mathcal{A}} = \min_{i \in \mathcal{A}} a_i$ and $\overline{a}_{\mathcal{A}} = \max_{i \in \mathcal{A}} a_i$. For any vector $\mathbf{v} = (\mathbf{v}_1^\top, \ldots, \mathbf{v}_K^\top)^\top \in \mathbb{R}^p$ and an arbitrary index set $I \subset \{1, \ldots, K\}$, we write $\mathbf{v}_I = (\mathbf{v}_k^\top, k \in I)^\top$ and define $\mathbf{X}_I = (\mathbf{X}_k, k \in I)$ to be the sub-matrix consisting of the columns of $\mathbf{X}$ with indices in I. Sub-Gaussian norm (Rudelson et al. 2013) of a random variable Z is denoted by $\|Z\|_{SG} = \sup_{r \ge 1} r^{-1/2} (E(|Z|^r))^{1/r}$. Let $\overline{c} = \tau \vee (1-\tau) = \max(\tau, 1-\tau)$ and $\underline{c} = \tau \wedge (1-\tau) = \min(\tau, 1-\tau)$. We use $\nabla f(\mathbf{v}) = \partial f(\mathbf{v}) / \partial \mathbf{v}$ to represent the gradient of a differentiable function $f : \mathbb{R}^p \to \mathbb{R}$, and we denote $\nabla_I f(\mathbf{v}) = (\partial f(\mathbf{v}) / \partial v_k, k \in I)$. The $\ell_{2,1}$-norm and $\ell_{2,\infty}$-norm of $\mathbf{v}$ are defined by $\|\mathbf{v}\|_{2,1} = \sum_{k=1}^K \|\mathbf{v}_k\|_2$ and $\|\mathbf{v}\|_{2,\infty} = \max_{1 \le k \le K} \|\mathbf{v}_k\|_2$. Denote s be the number of no null groups for the true coefficients $\boldsymbol{\beta}^*$, $s_{\mathcal{A}} = \sum_{k \in \mathcal{A}} p_k$ the number of

variables in the set $\mathcal{A}$, $\overline{p}_m = \max_{1\le k\le K} p_k$ and $\overline{p}_{\mathcal{A}} = \max_{k\in\mathcal{A}} p_k$. Let $\lambda_{\min}(.)$ and $\lambda_{\max}(.)$ are two functions that return the smallest and largest eigenvalues of a symmetric matrix respectively, and define $\overline{\rho} = \max_{1\le k\le K} \rho_k$ and $\underline{\rho} = \min_{1\le k\le K} \rho_k$, where $\rho_k = \lambda_{\max}(n^{-1}\mathbf{X}_k^\top \mathbf{X}_k)$. Finally, let $\rho_{\min} = \lambda_{\min}(n^{-1}\mathbf{X}_{\mathcal{B}}^\top \mathbf{X}_{\mathcal{B}})$ and $\overline{\rho}_{\max} = \lambda_{\max}(n^{-1}\mathbf{X}_{\mathcal{B}}^\top \mathbf{X}_{\mathcal{B}})$. We assume $\rho_{\min} > 0$, thereby the important variables are not linearly dependent. Define $[a]^+ = \max(0, a)$ for any $a \in \mathbb{R}$.

Let $\mathcal{C}_3 = \{\boldsymbol{\delta} \in \mathbb{R}^p, \|\boldsymbol{\delta}_{\mathcal{A}^c}\|_{2,1} \le 3\|\boldsymbol{\delta}_{\mathcal{A}}\|_{2,1}\}$ be a cone in $\mathbb{R}^p$. To study the estimation accuracy of the GPER-Lasso, we impose the following conditions on the design matrix $\mathbf{X}$ and the random errors $\boldsymbol{\epsilon}$.

- (C1) The columns of $\mathbf{X}$ are normalizable, that is, $M_0 = \max_{1\le j\le p} \dfrac{\|X_j\|_2}{\sqrt{n}} \in (0, \infty)$;
- (C2) The random errors ϵ_i are i.i.d. sub-Gaussian random variables satisfying $\mathscr{E}^{\tau}(\epsilon_i) = 0$, for $i = 1, \ldots, n$;
- (C3) $\kappa = inf_{\boldsymbol{\delta}\in\mathcal{C}_3} \dfrac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}\|_{2,1}} \in (0, \infty)$;
- (C4) $\varrho = inf_{\boldsymbol{\delta}\in\mathcal{C}_3} \dfrac{\|\mathbf{X}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_{\mathcal{A}}\|_{2,1}\|\boldsymbol{\delta}\|_{2,\infty}} \in (0, \infty)$.

The consistency of the group Lasso estimator has been extensively studied in the literature under some conditions (Meier et al. 2008; Bühlmann and Van De Geer 2011). Condition (C3) is known as the restricted eignvalue condition and has been frequently assumed in the literature to study the group Lasso (Meier et al. 2009). Condition (C4) is an extension of the generalized invertibility factor (GIF) condition for group variables (Ye and Zhang 2010). Both conditions (*C*3) and (*C*4) are crucial assumptions to establish estimation consistency of the GPER-Lasso, for high-dimensional data.

Theorem 3 *Assume the true vector of coefficients $\boldsymbol{\beta}^*$ in (5) is s-sparse (s is the number of no null groups) and assume the conditions (C1)-(C2). Let $\hat{\boldsymbol{\beta}}^{\text{GLasso}}$ be any optimal solution to GPER-Lasso problem. Then with probability at least $1 - p^*$, we have $\|\hat{\boldsymbol{\beta}}^{\text{GLasso}} - \boldsymbol{\beta}^*\|_{2,1} \le 3\lambda^{\text{GLasso}}(4\kappa\underline{c})^{-1}$ if condition (C3) holds,*

and $\|\hat{\boldsymbol{\beta}}^{\text{GLasso}} - \boldsymbol{\beta}^*\|_{2,\infty} \le 3\lambda^{\text{GLasso}}(4\underline{c}\varrho)^{-1}$ if condition (C4) holds, where

$$p^* = 2p \ \exp\left(-\frac{Cn(\lambda^{\text{GLasso}})^2}{4K_0^2 M_0^2 \overline{p}_m}\right), \tag{19}$$

$\nu_0 = var(\Psi'_\tau(\epsilon_i))$, $K_0 = \|\Psi'_\tau(\epsilon_i)\|_{SG}$ and $C > 0$ is an absolute constant.

The proof of Theorem 3 is detailed in Appendix 4.

2.6 Theory for non-convex penalized GPER

To give a unified theoretical analysis of GMCP and GSCAD, we assume that the penalty $P_\lambda(t)$ is a general folded concave penalty function defined on $t \in (-\infty, \infty)$ satisfying (see Fan et al. 2014; Gu and Zou 2016):

1. (P1) $P_\lambda(t) = P_\lambda(-t)$;
2. (P2) $P_\lambda(t)$ is non-decreasing, concave in $t \in [0, \infty)$ and $P_\lambda(0) = 0$;
3. (P3) $P_\lambda(t)$ is differentiable in $t \in (0, \infty)$;
4. (P4) $P'_\lambda(t) \geq a_1\lambda$ for $t \in (0, a_2\lambda]$ and $P'_\lambda(0) := P'_\lambda(0+) \geq a_1\lambda$;
5. (P5) $P'_\lambda(t) = 0$ for $t \in [a\lambda, \infty)$ with some prespecified constant $a > a_2$.

The parameters a_1 and a_2 are fixed constants characterising the penalty function. One can verify that a corresponds to θ in Eqs. (8) and (9) for both penalties GMCP and GSCAD, respectively, and $a_1 = a_2 = 1$ for GSCAD, and $a_1 = 1 - \theta^{-1}$, $a_2 = 1$ for GMCP.

In the following theorem, we show that the solution given by GLLA Algorithm 2 with any non-convex penalty satisfying the above conditions (1)-(5), enjoys the oracle property. Assume we have a sufficient signal strength in the nonzero components of $\boldsymbol{\beta}^*$. That is, assume $(A1)$ $\min_{k\in\mathcal{A}}\|\boldsymbol{\beta}^*_k\|_2 > (a+1)\lambda$. Our result is outlined next.

Theorem 4 *Assume in model* (5) *the vector of the true coefficients* $\boldsymbol{\beta}^*$ *is s-sparse and satisfies assumption* $(A1)$. *Assume conditions* $(C1)$-$(C2)$ *hold and take* $\hat{\boldsymbol{\beta}}_{GLasso}$ *as the initial value in Algorithm 2. Let* $a_0 = 1 \wedge a_2$. *Take* $\lambda \geq 3\lambda^{\mathrm{GLasso}}(4\kappa\underline{c}a_0)^{-1}$ *when* $(C3)$ *holds, or* $\lambda \geq 3\lambda^{\mathrm{GLasso}}(4\underline{c}\varrho a_0)^{-1}$ *when* $(C4)$ *holds, or take* $\lambda \geq 3\lambda^{\mathrm{GLasso}}a_0^{-1}\big((4\underline{c}\varrho)^{-1} \wedge (4\kappa\underline{c})^{-1}\big)$ *when both* $(C3)$ *and* $(C4)$ *hold. The GPER-GLLA estimator converges to* $\hat{\boldsymbol{\beta}}^{oracle}$ *after two iterations with probability at least* $1 - p_1 - p_2 - p_3$, *where* $p_1 = p^*$ *is given by* (19),

$$p_2 = 2(p - s_{\mathcal{A}})\ \exp\Big(-\frac{Cn\lambda^2 a_1^2}{4K_0^2 M_0^2 \overline{p}_m}\Big) + \Gamma(Q_1\lambda, n, s_{\mathcal{A}}, K_0, M_0, \rho_{\max}, \nu_0),$$

and

$$p_3 = \Gamma(2\underline{c}\rho_{\min}R\overline{p}_{\mathcal{A}}^{-1}; n, s_{\mathcal{A}}, K_0, M_0, \rho_{\max}, \nu_0),$$

where $Q_1 = \dfrac{a_1\underline{c}\rho_{\min}}{2\overline{c}M_0\rho_{\max}^{1/2}\overline{p}_m^{1/2}}$, $\nu_0 = var(\Psi'_\tau(\epsilon_i))$, $R = \min_{k\in\mathcal{A}}\|\boldsymbol{\beta}^*_k\|_2 - a\lambda > 0$, K_0 *is defined in Theorem 3,* $\Gamma(x; n, s, K, M, \rho, \nu)$ *is given by*

$$\Gamma(x; n, s, K, M, \rho, \nu) = 2\exp\left(-\frac{C\nu^2[(n^{1/2}x - \nu\rho^{1/2}s^{1/2})]^+}{K^4\rho}\right) \wedge 2s\exp\left(-\frac{Cnx^2}{K^2M^2s}\right),$$

and $C > 0$ *is an absolute constant.*

The proof of Theorem 4 is detailed in Appendix 4.

2.7 Some solution paths of GPER methods

Our motivation for introducing the GPER approach is illustrated in Fig. 1 below.

We adapted the illustration example of Mkhadri and Ouhourane (2015) for illustration. We generated one dataset of $n = 50$ observations and five initial predictors, say $\tilde{X}_k$ ($k = 1, \ldots, K = 5$), from a multivariate standard normal distribution with the correlation among the predictors was set to be equal to 0.5. We computed a cubic B-spline basis (W_k^1, W_k^2, W_k^3) from each predictor $\tilde{X}_k$, $k = 1, \ldots, 5$. Then we set $X_k^j = W_k^j$, for $j = 1, 2, 3$ and $k = 1, \ldots, 5$, which leads to 15 predictors X_k^j that are clustered in $K = 5$ groups, i.e. $G_k = \{X_k^1, X_k^2, X_k^3\}$, for $k = 1, \ldots, 5$.

The response $\mathbf{y}$ is generated as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Phi}(\tilde{X}_1)\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, 1),$$

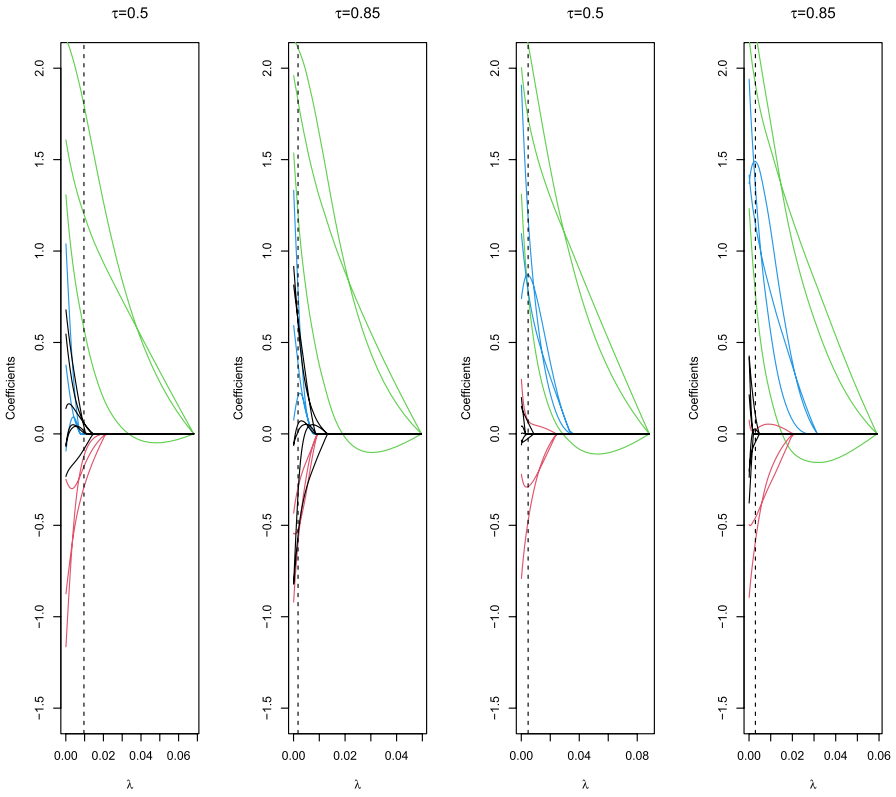

Fig. 1 The first two panels (from left to right) are for example 1, and the last two panels are for example 2. The results show the coefficients' profiles as a function of the tuning parameter λ corresponding to LS-GLasso (GPER-GLasso with $\tau = 0.5$) and GPER-GLasso with $\tau = 0.85$. The dashed vertical lines report selected optimal λ using 5-fold CV. The group coefficients of G_1, G_2 and G_3 are plotted in blue, green and red colors, respectively. The black color corresponds to the noisy groups of predictors G_4 and G_5

where $\Phi(\cdot)$ is the cumulative distribution function of the univariate standard normal distribution. Using $\Phi(\cdot)$ in the term of the variance in the simulations is considered by many authors to generate a model with heteroscedasticity (Wang et al. 2012; Gu and Zou 2016).

We considered two illustration examples. In the first example, we considered that G_2 and G_3 have an effect on the mean of the outcome and G_1 has an effect only on the scale. Thus, $\boldsymbol{\beta}$ is defined as

$$\boldsymbol{\beta} = (\underbrace{0,0,0}_{blueG_1}, \underbrace{2,2,2}_{greenG_2}, \underbrace{-1,-1,-1}_{redG_3}, \underbrace{0,0,0}_{blackG_4}, \underbrace{0,0,0}_{blackG_5}).$$

The second example is similar to the first one, except that G_1 has an effect on both the mean and scale (i.e. overlapping effect). That is, $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta} = (\underbrace{1,1,1}_{blueG_1}, \underbrace{2,2,2}_{greenG_2}, \underbrace{-1,-1,-1}_{redG_3}, \underbrace{0,0,0}_{blackG_4}, \underbrace{0,0,0}_{blackG_5}).$$

Figure 1 shows the results of the coefficient profiles as a function of λ values for GPER-GLasso, at different locations. In the first two panels (from the left to right), we show major advantages of using group penalized expectile regression approaches when τ is different than 0.5 ($\tau \neq 0.5$) for detecting heteroscedasticity when the groups of variables have an effect only on the scale. Indeed, GPER-GLasso selected the Group G_1 (blue color) for $\tau = 0.95$, but it does not for $\tau = 0.5$, whcih means that G_1 is detected as a heteroscedastic group. However, in the second scenario (two last panels of Fig. 1), the effect of G_1 overlaps for the mean and scale. In this case, GPER-GLasso selected G_1 for both values of $\tau = 0.5$ and 0.95, and thus, one cannot answer the question if G_1 is a heteroscedastic group or not. This is the main motivation to introduce the COupled (Group) Expectile Regression for analyzing the heteroscedasticity in high-dimensional settings.

3 Coupled group penalized expectile regression: COGPER

3.1 Methodology: COGPER general algorithm

We consider the following linear scale model for analyzing heteroscedasticity

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{x}_i^\top \boldsymbol{\gamma} \epsilon_i,$$

where ϵ_i are i.i.d. random errors, and we assume that $\mathbb{E}(\epsilon_i) = 0$. The unknown parameters to be estimated are the p-dimensional vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, corresponding to the effect of the covariates on the mean and the scale of the response variable, respectively. We suppose that $\mathbf{x}_i^\top \boldsymbol{\gamma} > 0$ for all i. This model has been studied by many authors in standard regression (Efron 1991; Koenker and Zhao 1994). It has been proposed by Gu and Zou (2016) in high dimension to select important variables that have an effect on both the mean and the scale functions. Let

$e_\tau = \mathscr{E}^\tau(\epsilon_1)$ be the τ-mean of the random error for $\tau \in (0,1)$, then the τ-mean of y_i given $\mathbf{x}_i$ is $\mathscr{E}^\tau(y_i|\mathbf{x}_i) = \mathbf{x}_i^\top(\boldsymbol{\beta} + \boldsymbol{\gamma} e_\tau)$. Let $\mathcal{A}_1 \equiv supp(\boldsymbol{\beta}^*) = \{k : \boldsymbol{\beta}_k^* \neq 0\}$ and $\mathcal{A}_2 \equiv supp(\boldsymbol{\gamma}^*) = \{k : \boldsymbol{\gamma}_k^* \neq 0\}$ be the active sets of $\boldsymbol{\beta}^*$ and of $\boldsymbol{\gamma}^*$, respectively. Then, when we take $\boldsymbol{\phi} = \boldsymbol{\gamma} e_\tau$, we will deal with $\boldsymbol{\phi}$ instead of $\boldsymbol{\gamma}$, and if $e_\tau \neq 0$ we have $supp(\boldsymbol{\gamma}^*) \equiv supp(\boldsymbol{\phi}^*)$.

We rely again on Gu and Zou (2016) and develop the COupled Group Expectile Regression (COGPER) method, which estimates and selects the relevant groups of variables that have effect on the mean and scale simultaneously. The COGPER model is defined as follows

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}) = \arg\min_{(\boldsymbol{\beta},\boldsymbol{\phi}) \in \mathbb{R}^{2p}} S_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}) + \sum_{k=1}^{K} w_k P_{\lambda_1}(\|\boldsymbol{\beta}_k\|_2) + \sum_{k=1}^{K} u_k P_{\lambda_2}(\|\boldsymbol{\phi}_k\|_2), \tag{20}$$

where

$$S_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}) = \Psi_{0.5}(\boldsymbol{\beta}) + \Psi_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}), \tag{21}$$

with $\Psi_{0.5}(\boldsymbol{\beta})$ is given by (4) and

$$\Psi_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \mathbf{x}_i^\top \boldsymbol{\phi}).$$

The penalties $P_{\lambda_1}(.)$ and $P_{\lambda_2}(.)$ could be one of the penalties GLasso, GMCP or GSCAD defined in (7), (8) and (9), respectively. The scalars w_k and u_k are known weights for each group, and can be defined in a similar way as in the GPER approach to control for the group size, for instance. In this work, we set $w_k = u_k = \sqrt{p_k}$.

Notice that the non-convex penalties, GMCP and GSCAD, enjoy the oracle property. For the COGPER approach, the oracle estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\phi} = \boldsymbol{\gamma} e_\tau$ are given by

$$(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle}) = \arg\min_{(\boldsymbol{\beta},\boldsymbol{\phi}) \in \mathbb{R}^{2p} : \boldsymbol{\beta}_{\mathcal{A}_1^c} = \mathbf{0}, \boldsymbol{\phi}_{\mathcal{A}_2^c} = \mathbf{0}} S_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}). \tag{22}$$

where $\mathcal{A}_1$ and $\mathcal{A}_2$ are the true support set of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ respectively.

To solve the problem (20), we proceed in a similar way as in Sect. 2. That is, we focus on updating one group at a time ($\boldsymbol{\beta}_k$ or $\boldsymbol{\phi}_k$). We majorize each loss function in the right-hand side of (21) by a quadratic surrogate function. Then, for each group k ($k = 1, \ldots, K$), we obtain two upper bound approximations for updating $\boldsymbol{\beta}_k$ and $\boldsymbol{\phi}_k$, respectively, as follows

$$\begin{aligned} Q_1(\boldsymbol{\beta}_k|\tilde{\boldsymbol{\beta}}_{-k}, \tilde{\boldsymbol{\phi}}) :=& S_\tau(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\phi}}) - 2(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^\top \mathbf{U}_k^{0.5} + \\ & 2\gamma_k(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^\top(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) - (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^\top \mathbf{U}_k^\tau + \\ & 2c\gamma_k(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^\top(\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) + P_{\lambda_1}(\|\boldsymbol{\beta}_k\|_2), \end{aligned} \tag{23}$$

and

$$
\begin{aligned}
Q_2(\boldsymbol{\phi}_k|\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\phi}}_{-k}) :=& S_\tau(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\phi}}) - (\boldsymbol{\phi}_k - \tilde{\boldsymbol{\phi}}_k)^\top \mathbf{U}_k^\tau \\
&+ 2c\gamma_k(\boldsymbol{\phi}_k - \tilde{\boldsymbol{\phi}}_k)^\top(\boldsymbol{\phi}_k - \tilde{\boldsymbol{\phi}}_k) + P_{\lambda_2}(\|\boldsymbol{\phi}_k\|_2),
\end{aligned} \tag{24}
$$

where γ_k is the largest eigenvalue of the matrix $\mathbf{H}_k = \mathbf{X}_k^\top \mathbf{X}_k$, $c = 2\max(\tau, 1-\tau)$, $\mathbf{U}_k^{0.5} = -\nabla_k \Psi_{0.5}(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k})$ and $\mathbf{U}_k^\tau = -\nabla_{\boldsymbol{\beta}_k} \Psi_\tau(\tilde{\boldsymbol{\beta}}_k;\tilde{\boldsymbol{\beta}}_{-k}, \tilde{\boldsymbol{\phi}}) = -\nabla_{\boldsymbol{\phi}_k} \Psi_\tau(\tilde{\boldsymbol{\phi}}_k;\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\phi}}_{-k})$.

Proposition 5 *Let $Q_1(\boldsymbol{\beta}_k|\tilde{\boldsymbol{\beta}}_{-k}, \tilde{\boldsymbol{\phi}})$ and $Q_2(\boldsymbol{\phi}_k|\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\phi}}_{-k})$ be the surrogate loss functions given by (23) and (24). Let $P_{\lambda_1}(\|\boldsymbol{\beta}_k\|_2)$ and $P_{\lambda_2}(\|\boldsymbol{\phi}_k\|_2)$ be one of the three penalties given in (7), (8) and (9). The closed form solutions to (23) and (24) of $(\tilde{\boldsymbol{\beta}}_k^{(new)}, \tilde{\boldsymbol{\phi}}_k^{(new)})$ for COGPER-GLasso, COGPER-GMCP and COGPER-GSCAD are, respectively, given by*

$$
\tilde{\boldsymbol{\beta}}_k^{(new)} = F(\mathbf{Z}_k) \longleftarrow \frac{1}{2(1+c)\gamma_k} \frac{S(\|\mathbf{Z}_k\|_2, \lambda_1 w_k)}{\|\mathbf{Z}_k\|_2} \mathbf{Z}_k
$$

$$
\tilde{\boldsymbol{\phi}}_k^{(new)} = G(\mathbf{W}_k) \longleftarrow \frac{1}{2c\gamma_k} \frac{S(\|\mathbf{W}_k\|_2, \lambda_2 u_k)}{\|\mathbf{W}_k\|_2} \mathbf{W}_k
$$

$$
\tilde{\boldsymbol{\beta}}_k^{(new)} = F(\mathbf{Z}_k) \longleftarrow \begin{cases} \frac{1}{2(1+c)\gamma_k - 1/\theta} \frac{S(\|\mathbf{Z}_k\|_2, \lambda_1 w_k)}{\|\mathbf{Z}_k\|_2} \mathbf{Z}_k, \\ \qquad \text{if } \|\mathbf{Z}_k\|_2 \le 2(1+c)\gamma_k \theta \lambda_1 w_k \\ \frac{1}{2(1+c)\gamma_k} \mathbf{Z}_k, \\ \qquad \text{if } \|\mathbf{Z}_k\|_2 > 2(1+c)\gamma_k \theta \lambda_1 w_k \end{cases}
$$

$$
\tilde{\boldsymbol{\phi}}_k^{(new)} = G(\mathbf{W}_k) \longleftarrow \begin{cases} \frac{S(\|\mathbf{W}_k\|_2, \lambda_2 w_k)}{2c\gamma_k - 1/\theta} \frac{1}{\|\mathbf{W}_k\|_2} \mathbf{W}_k, & \text{if } \|\mathbf{W}_k\|_2 \le 2c\gamma_k \theta \lambda_2 u_k \\ \frac{1}{2c\gamma_k} \mathbf{W}_k & \text{if } \|\mathbf{W}_k\|_2 > 2c\gamma_k \theta \lambda_2 u_k, \end{cases}
$$

$$
\tilde{\boldsymbol{\beta}}_k^{(new)} = F(\mathbf{Z}_k) \longleftarrow \begin{cases} \frac{S(\|\mathbf{Z}_k\|_2, \lambda_1 w_k)}{2(1+c)\gamma_k \|\mathbf{Z}_k\|_2} \mathbf{Z}_k, \\ \qquad \text{if} \quad \|\mathbf{Z}_k\|_2 \le (1 + 2(1+c)\gamma_k)\lambda_1 w_k \\ \frac{S(\|\mathbf{Z}_k\|_2, \frac{\lambda_1 w_k \theta}{\theta - 1})}{\|\mathbf{Z}_k\|_2 (2(1+c)\gamma_k - \frac{1}{\theta - 1})} \mathbf{Z}_k, \\ \qquad \text{if} \quad (1 + 2(1+c)\gamma_k)\lambda_1 w_k < \|\mathbf{Z}_k\|_2 \le 2(1+c)\gamma_k \theta \lambda_1 w_k \\ \frac{1}{2(1+c)\gamma_k} \mathbf{Z}_k \\ \qquad \text{if} \quad \|\mathbf{Z}_k\|_2 > 2(1+c)\gamma_k \theta \lambda_1 w_k, \end{cases}
$$

$$\tilde{\boldsymbol{\phi}}_k^{(new)} = G(\mathbf{W}_k) \longleftarrow \begin{cases} \frac{1}{2c\gamma_k} \frac{S(\|\mathbf{W}_k\|_2, \lambda_2 u_k)}{\|\mathbf{W}_k\|_2} \mathbf{W}_k, \\ \qquad \text{if} \quad \|\mathbf{W}_k\|_2 \le (1+2c\gamma_k)\lambda_2 u_k \\ \frac{1}{2c\gamma_k - \frac{1}{\theta-1}} \frac{S(\|\mathbf{W}_k\|_2, \frac{\lambda_2 u_k \theta}{\theta-1})}{\|\mathbf{W}_k\|_2} \mathbf{W}_k, \\ \qquad \text{if} \quad (1+2c\gamma_k)\lambda_2 u_k < \|\mathbf{W}_k\|_2 \le 2c\gamma_k \theta \lambda_2 u_k \\ \frac{1}{2c\gamma_k} \mathbf{W}_k \qquad \text{if} \quad \|\mathbf{W}_k\|_2 > 2c\gamma_k \theta \lambda_2 u_k, \end{cases}$$

where $\mathbf{Z}_k = \mathbf{U}_k^{0.5} + \mathbf{U}_k^{\tau} + 2(1+c)\gamma_k \widetilde{\boldsymbol{\beta}}_k$ *and* $\mathbf{W}_k = \mathbf{U}_k^{\tau} + 2c\gamma_k \widetilde{\boldsymbol{\phi}}_k$.

The proof of Proposition (5) is detailed in Appendix 3.

The following algorithm summarizes the steps of the COGPER framework with GLasso, GMCP and GSCAD penalties.

Algorithm 3 The COGPER algorithm for GLasso/GMCP/GSCAD penalties

Initialize $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\phi}})$;
repeat
  for $k = 1$ to K **do**
    $\tilde{\boldsymbol{\beta}}_k^{\text{new}} \leftarrow F(\mathbf{Z}_k)$
  end for
  for $k = 1$ to K **do**
    $\tilde{\boldsymbol{\phi}}_k^{\text{new}} \leftarrow G(\mathbf{W}_k)$
  end for
until Convergence of $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\phi}})$
Return $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\phi}})$.

3.2 Coupled expectile regression with GLLA penalty

Extension of the GLLA trick to solve coupled ER for a more general form of non-convex penalties can be done in a same way as described in Sect. 2.3. Our theoretical contribution in Sect. 3.5 is focuced on a class of non-convex penalties. This class includes GMCP and GSCAD.

Using the first order Taylor expansion of the non-convex penalty functions around $\|\tilde{\boldsymbol{\beta}}_k\|_2$ and $\|\tilde{\boldsymbol{\phi}}_k\|_2$ as defined in (15) leads to the following COGPER problem with the Group LLA (GLLA) penalty

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}) = \arg\min_{(\boldsymbol{\beta},\boldsymbol{\phi}) \in \mathbb{R}^{2p}} \big(S_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}) + \sum_{k=1}^{K} w'_k \|\boldsymbol{\beta}_k\|_2 + \sum_{k=1}^{K} u'_k \|\boldsymbol{\phi}_k\|_2 \big), \qquad (25)$$

where $(w'_k, u'_k) = (w_k P'_{\lambda_1}(\|\tilde{\boldsymbol{\beta}}_k\|_2), u_k P'_{\lambda_2}(\|\tilde{\boldsymbol{\phi}}_k\|_2))$ for $k = 1, \ldots, K$. The weights w'_k and u'_k depend on the non-convex penalty function through the first derivative $P'_{\lambda}(.)$. The problem (25) can be solved using a COGPER-GLasso update similar to Algorithm 3. The details of the COGPER approaches with GLLA penalty is given in the next algorithm.

Algorithm 4 The COGPER algorithm with GLLA penalty

Initialize $i = 0$; $(\tilde{\boldsymbol{\beta}}^i, \tilde{\boldsymbol{\phi}}^i) = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\phi}})^{\text{initial}}$;
Compute the weights $(\tilde{w}_k^i, \tilde{u}_k^i) = (w_k P'_{\lambda_1}(\|\tilde{\boldsymbol{\beta}}_k^i\|_2), u_k P'_{\lambda_2}(\|\tilde{\boldsymbol{\phi}}_k^i\|_2))$ for $k = 1, \ldots, K$;
repeat
  $i \leftarrow i + 1$;
  for $k = 1$ to K **do**
    Solve the following convex optimization problem for $(\tilde{\boldsymbol{\beta}}^i, \tilde{\boldsymbol{\phi}}^i)$:

$$(\tilde{\boldsymbol{\beta}}^i, \tilde{\boldsymbol{\phi}}^i) = \underset{(\boldsymbol{\beta},\boldsymbol{\phi})\in\mathbb{R}^{2p}}{\arg\min} \left(S_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}) + \tilde{w}_k^{i-1}\|\boldsymbol{\beta}_k\|_2 + \tilde{u}_k^{i-1}\|\boldsymbol{\phi}_k\|_2 \right); \tag{26}$$

  end for
  for $k = 1$ to K **do**
    Calculate $(\tilde{w}_k^i, \tilde{u}_k^i) = (w_k P'_{\lambda_1}(\|\tilde{\boldsymbol{\beta}}_k^i\|_2), u_k P'_{\lambda_2}(\|\tilde{\boldsymbol{\phi}}_k^i\|_2))$;
  end for
until Convergence of $(\tilde{\boldsymbol{\beta}}^i, \tilde{\boldsymbol{\phi}}^i)$
Return $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\phi}})$.

To solve the problem (26), we use Algorithm 3 with GLasso (COGPER-GLasso) with $(w_k, u_k) = (\tilde{w}_k^{i-1}, \tilde{w}_k^{i-1})$ for $k = 1, \ldots, K$.

3.3 Implementation

To obtain the solution path of COGPER with the tuning parameters (λ_1, λ_2), one can choose a (relatively small) grid of values for λ_1 and then compute a grid of values of λ_2 covering the entire range, and vice versa. But, the resulting coefficients' path solution might not be smooth with several successive jumps. To remedy this problem, we follow Gu and Zou (2016) in their implementation and set a common tuning parameter in solving the problem (21) for the two penalties, as follows

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}) = \underset{\boldsymbol{\beta},\boldsymbol{\phi}}{\arg\min} \left(\nu \Psi_{0.5}(\boldsymbol{\beta}) + \Psi_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}) \right) + \sum_{k=1}^{K} w_k P_\lambda(\boldsymbol{\beta}_k) + \sum_{k=1}^{K} u_k P_\lambda(\boldsymbol{\phi}_k), \tag{27}$$

where ν is an additional weight parameter for the mean loss function, which compensates for the use of the common tuning parameter for both the mean and scale coefficients. In our implementation we set the default value of $\nu = 1$ as in the `SALES` R package (Gu and Zou 2016), but other values of ν can also be investigated.

The implementation of problem (27) has the advantage of allowing smooth path solutions for both $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$. To calculate $\lambda_{\max}$, we first obtain estimates of the intercepts $(\hat{\beta}_0, \hat{\phi}_0)$ through the null model wit all the groups' coefficients are set to be zero

$$(\hat{\beta}_0, \hat{\phi}_0) = \arg\min_{\beta_0,\phi_0} \left(S_\tau^\nu(\beta_0, \phi_0) := \nu \Psi_{0.5}(\beta_0, \mathbf{0}) + \Psi_\tau(\beta_0, \mathbf{0}, \phi_0, \mathbf{0}) \right).$$

According to KKT conditions, we have

$$\lambda_{\max} = \max \left\{ \max_{k=2,..,K} \|(\nabla_{\boldsymbol{\beta}_k} S_\tau^\nu(\hat{\beta}_0, \hat{\phi}_0))\|_2 / w_k, \ \max_{k=2,..,K} \|(\nabla_{\boldsymbol{\phi}_k} S_\tau^\nu(\hat{\beta}_0, \hat{\phi}_0))\|_2 / u_k \right\}.$$

Let $\lambda_{\min} = \eta \lambda_{\max}$, where $\eta = 0.001$ if $n \leq p$; otherwise, $\eta = 0.05$. We take $M - 2 = 98$ points uniformly in log-scale between $\lambda_{\min}$ and $\lambda_{\max}$. This sequence is denoted by $[\lambda^m]_{m=1}^M$. We use the warm-start and the strong rule tricks to speed up our code; see Sect. 2.4 and Ouhourane et al. (2021) for more details.

3.4 Theory for COGPER-GLasso

For the COGPER-GLasso approach, let $\mathcal{A}_1 \equiv \text{supp}(\boldsymbol{\beta})$ and $\mathcal{A}_2 \equiv \text{supp}(\boldsymbol{\phi})$ be the active group of $\boldsymbol{\beta}^*$ and of $\boldsymbol{\phi}^*$ respectively. Let $\mathcal{A}_0 = (\mathcal{A}_1, \mathcal{A}_2')$, where $\mathcal{A}_2' = \{k + K : \boldsymbol{\phi}_k^* \neq 0\}$. For $N \geq 1$, define $\xi_N = \{\boldsymbol{\delta} \in \mathbb{R}^{2p} : \|\boldsymbol{\delta}_{\mathcal{A}_0^c}\|_{2,1} \leq N \|\boldsymbol{\delta}_{\mathcal{A}_0}\|_{2,1}\}$, $\underline{\lambda}^{\text{GLasso}} = \lambda_1^{\text{GLasso}} \wedge \lambda_2^{\text{GLasso}} = \min(\lambda_1^{\text{GLasso}}, \lambda_2^{\text{GLasso}})$ and $\overline{\lambda}^{\text{GLasso}} = \lambda_1^{\text{GLasso}} \vee \lambda_2^{\text{GLasso}} = \max(\lambda_1^{\text{GLasso}}, \lambda_2^{\text{GLasso}})$ and $\tilde{N} = \overline{\lambda}^{\text{GLasso}} / \underline{\lambda}^{\text{GLasso}}$. For $k = 1, 2$, denote $\rho_{k,max} = \lambda_{\max}(n^{-1} \mathbf{X}_{\mathcal{A}_k}^\top \mathbf{X}_{\mathcal{A}_k})$, $\rho_{k,min} = \lambda_{\min}(n^{-1} \mathbf{X}_{\mathcal{A}_k}^\top \mathbf{X}_{\mathcal{A}_k})$, $\phi_{\min} = \rho_{k,min} \wedge \rho_{k,max}$, $\phi_{\max} = \rho_{k,min} \vee \rho_{k,max}$, and we assume $\phi_{\min} > 0$. Let $\boldsymbol{I}_2$ be a 2×2 identity matrix and $\otimes$ denotes the Kronecker product. To establish an error bound for the COGPER-GLasso estimator, the following conditions on the design matrix, $\mathbf{X}$, and the random errors, $\boldsymbol{\epsilon}$, are imposed:

- (C1') The columns of $\mathbf{X}$ are normalizable, that is, $M_0 = \max_{1 \leq j \leq p} \dfrac{\|X_j\|_2}{\sqrt{n}} \in (0, \infty)$;
- (C2') $M_1 = \|\mathbf{X}\boldsymbol{\phi}^*\|_\infty \in (0, \infty)$;
- (C3') The random errors ϵ_i are i.i.d. mean-zero sub-Gaussian random variables;
- (C4') $\overline{\kappa} = \kappa(3\tilde{N}) \in (0, \infty)$ where $\kappa = inf_{\delta \in \xi_N} \dfrac{\boldsymbol{\delta}^\top [\boldsymbol{I}_2 \otimes (n^{-1} \mathbf{X}^\top \mathbf{X})] \boldsymbol{\delta}}{n \|\boldsymbol{\delta}\|_{2,1}^2}$;
- (C5') $\overline{\varrho} = \varrho(3\tilde{N}) \in (0, \infty)$ where $\varrho = inf_{\delta \in \xi_N} \dfrac{\boldsymbol{\delta}^\top [\boldsymbol{I}_2 \otimes (n^{-1} \mathbf{X}^\top \mathbf{X})] \boldsymbol{\delta}}{n \|\boldsymbol{\delta}_{\mathcal{A}_0}\|_{2,1} \|\boldsymbol{\delta}\|_{2,\infty}}$.

As Theorem 3, both conditions $(C'4)$–$(C'5)$ are crucial assumptions to establish the estimation consistency of the COGPER-GLasso estimator.

Theorem 6 *Suppose the true parameter vectors $\boldsymbol{\beta}^*$ and $\boldsymbol{\phi}^*$ are respectively s_1-sparse and s_2-sparse and assume conditions (C1')–(C3') hold. Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\phi}}$ be optimal solutions of COGPER-GLasso. Then, with probability at least $1-\pi^*$*

$$\left\| \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\phi}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta}^* \\ \boldsymbol{\phi}^* \end{pmatrix} \right\|_{2,1} \leq \frac{(3/2)\overline{\lambda}^{\text{GLasso}}}{c_0 \overline{\kappa}}$$

if the condition (C4') holds, and

$$\left\| \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\phi}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta}^* \\ \boldsymbol{\phi}^* \end{pmatrix} \right\|_{2,\infty} \leq \frac{(3/2)\overline{\lambda}^{\text{GLasso}}}{c_0 \varrho}$$

if the condition (C5') holds, where

$$\pi^* = 2p \ \exp\left(- \frac{Cn(\lambda_1^{\text{GLasso}})^2}{4(K_1+K_2)^2 M_0^2 M_1^2 \overline{p}_m} \right) + 2p \ \exp\left(- \frac{Cn(\lambda_2^{\text{GLasso}})^2}{4K_2^2 M_0^2 M_1^2 \overline{p}_m} \right)$$

$c_0 = 2^{-1}[(1+\underline{c}) - (1+16\underline{c}^2)^{1/2}]$, $K_1 = \|\epsilon_i\|_{SG}$, $K_2 = \|S'_\tau(\epsilon_i - e_\tau)\|_{SG}$, *and* $C > 0$ *is an absolute constant.*

The proof of Theorem 6 is detailed in Appendix 4.

3.5 Theory for non-convex penalized COGPER

In this section we investigate the theoretical properties of the COGPER approach with the non-convex penalties. More precisely, in the next theorem, we show that the solution given by Algorithm 4 converges to the oracle estimator in two steps. To do this, assume the following additional assumption (A2) $\min_{k\in\mathcal{A}_1} \|\boldsymbol{\beta}_k^*\| > (a+1)\lambda_1$ and $\min_{k\in\mathcal{A}_2} \|\boldsymbol{\phi}_k^*\| > (a+1)|e_\tau|^{-1}\lambda_2$.

Theorem 7 *Suppose that $\boldsymbol{\beta}^*$ and $\boldsymbol{\phi}^*$ are respectively s_1-sparse and s_2-sparse. Take $\hat{\boldsymbol{\beta}}^{\text{GLasso}}$ and $\hat{\boldsymbol{\phi}}^{\text{GLasso}}$ as the initial values and assume conditions (C1')–(C3') hold. Take $\lambda \geq (3/2)(a_0 c_0 \overline{\kappa})^{-1}\overline{\lambda}^{\text{GLasso}}$ when (C4') holds, or take $\lambda \geq (3/2)(a_0 c_0 \overline{\varrho})^{-1}\overline{\lambda}^{\text{GLasso}}$ when (C5') holds, or take $\lambda \geq (3/2)\overline{\lambda}^{\text{GLasso}}(a_0 c_0)^{-1}\left(\kappa^{-1} \wedge \varrho^{-1}\right)$ when (C4') and (C5') hold. The COGPER-GLLA algorithm converges to the oracle estimators $(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})$ in two iterations with probability at least $1-\pi_1-\pi_2-\pi_3$, where $\pi_1 = \pi^*$ is given in Theorem 6, and*

$$\pi_2 = 2(p - s_{\mathcal{A}_1}) \exp\left(- \frac{Cn\lambda^2 a_1^2}{4M_0^2 M_1^2 (K_1 + K_2)^2 \overline{p}^2_{\mathcal{A}_1^c}} \right) + 2(p - s_{\mathcal{A}_2}) \exp\left(- \frac{Cn\lambda^2 a_1^2}{4M_0^2 M_1^2 K_2^2 \overline{p}^2_{\mathcal{A}_2^c}} \right) + \Gamma(Q_2\lambda/2; n, s_{\mathcal{A}_1}, K_1 + K_2, M_0, M_1, M_1^2\rho_{1,max}, \nu_1) + \Gamma(Q_2\lambda/2; n, s_{\mathcal{A}_2}, K_2, M_0 M_1, M_1^2\rho_{2,max}, \nu_2),$$

$$\pi_3 = \Gamma\left(c_0\phi_{\min}\frac{\overline{R}}{2\overline{p}_k}; n, s_{\mathcal{A}_1}, K_1 + K_2, M_0, M_1, M_1^2\rho_{1,max}, \nu_1\right) + \Gamma\left(c_0\phi_{\min}\frac{\overline{R}}{2\overline{p}_k}; n, s_{\mathcal{A}_2}, K_2, M_0 M_1, M_1^2\rho_{2,max}, \nu_2\right),$$

where $Q_2 = \frac{a_1}{(1+2\overline{c})M_0\phi_{\max}^{1/2}}$, $\overline{R} = (1+a)\lambda_1 \vee \lambda_2$, $\nu_1 = var(\epsilon_i + \Psi'_\tau(\epsilon_i - \mathscr{E}_\tau))$, $\nu_2 = var(\Psi'_\tau(\epsilon_i - \mathscr{E}_\tau))$, C, c_0, K_1, *and* K_2 *are given in Theorem 6, and the function* $\Gamma(.)$ *is defined in Theorem 4.*

The proof of Theorem 7 is detailed in Appendix 4.

3.6 Some solution paths of COGPER method

The motivation for the introduction of COGPER approach is illustrated in Fig. 2. This figure was provided using the same dataset that is generated under the second model (second example) of the simulation study of Sect. 2.7. Recall that under this model, G_1 was generated with effect on both the mean and scale of the response variable.

Figure 2 shows that COGPER-GLasso has a tendency to select the groups of variables that have effect on the conditional τ-mean for $\tau \in (0.5, 0.85)$. Furthermore, the heteroscedastic effect of group G_1 in the scale function is often selected as non-null effect when fitting COGPER for 0.85th conditional mean (blue-color group in the right panel), but it is not the case for $\tau = 0.5$. This shows that the COGPER not only can be used to detect the heteroscedastic group G_1, but can also estimates the amount of the heteroscedastic effect $\hat{\boldsymbol{\phi}}_1$ and separates it from the mean function effect $\hat{\boldsymbol{\beta}}_1$.

4 Numerical experiments

4.1 Simulation setting

We carried out a simulation study to illustrate the utility of the proposed approaches. We adapted the scenarios 1 and 2 of Gu and Zou (2016) to the additive model context, in which the response is modeled as a sum of functions of the covariates. That is, two scenarios were considered in the simulations. In both scenarios, we

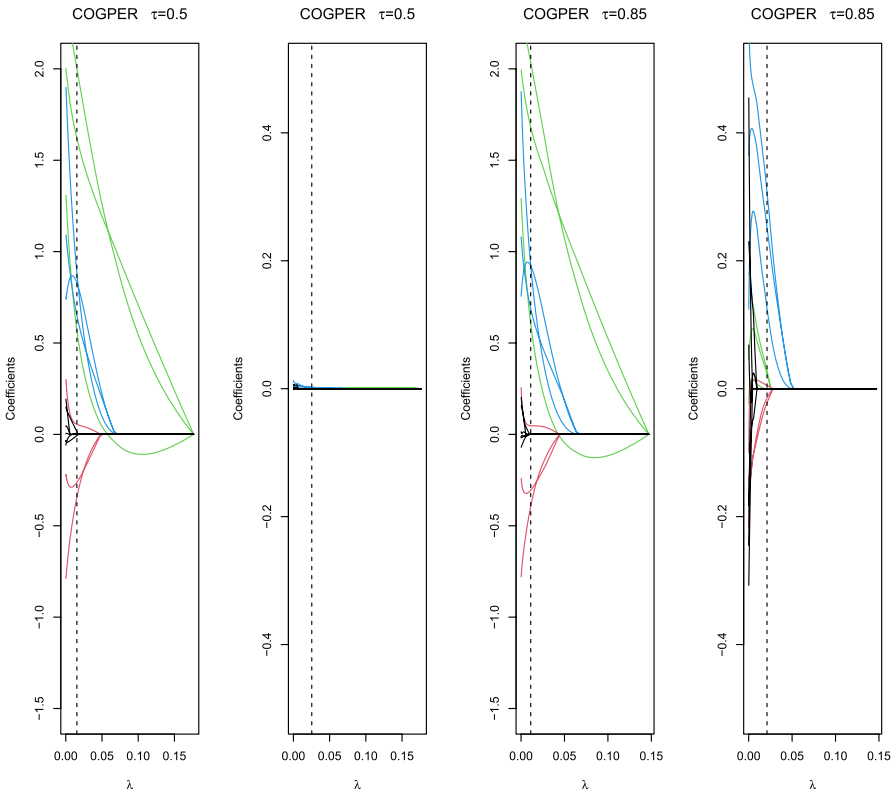

Fig. 2 From left to right, the coefficient profiles corresponding to $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}})$ obtained using COGPER-GLasso for $\tau \in (0.5, 0.85)$, respectively, are plotted as a function of the tuning parameter λ. The data are generated from the illustration example 2 of Sect. 2.7. The dashed line indicates the optimal value of λ using 5-fold CV. The group coefficients G_1, G_2 and G_3 are plotted in blue, green and red colors, respectively. The black color corresponds to the noisy groups G_4 and G_5

considered fitting an additive model of continuous factors represented by B-splines basis functions. This means that the effect of the factors is represented through non-linear functions. Our simulation results are based on a one independent dataset. This data is used to fit models and to select the tuning parameter using the 5-fold cross-validation (5-fold CV). We selected the regularization parameter by minimizing the CV error defined as

$$\frac{1}{n_{\text{validation}}} \sum_{i \in \text{validation}} \rho_\tau(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})$$

and

$$\frac{1}{n_{\text{validation}}} \sum_{i \in \text{validation}} \rho_{0.5}(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) + \rho_\tau(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} - \mathbf{x}_i^\top \hat{\boldsymbol{\phi}})$$

for GPER and COGPER, respectively.

The first scenario was considered in Wang et al. (2012) for the sparse quantile regression and in Gu and Zou (2016) for expectile regression. The predictors were generated in three steps. First, from the multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with $\Sigma = (0.5^{|i-j|})_{K\times K}$, we draw n-dimensional samples from $(Z_1, \ldots, Z_K)$, where $K = 50$ and $n = 300$. Second, for each variable $Z_k, k = 1, \ldots, K$, we derived a cubic B-spline basis (W_k^1, W_k^2, W_k^3). In the third step, we set $X_1^l = \Phi(W_1^l)$ and $X_k^l = W_k^l$ for $k = 2, 3, \ldots, K$ and $l = 1, 2, 3$, where $\Phi(.)$ is the standard normal CDF. Thus, the design matrix is $300 \times (50 * 3)$ and is defined as $\mathbf{X} = [X_k^l]_{l,k}$, $k = 1, \ldots, 50$, and $l = 1, 2, 3$. The response variable is then simulated from the following linear heteroscedastic model:

$$Y = \underbrace{Z_6}_{G_6} + \underbrace{Z_{12}}_{G_{12}} + \underbrace{Z_{15}}_{G_{15}} + \underbrace{Z_{20}}_{G_{20}} + \underbrace{\Phi(Z_1)}_{G_1^{\phi}} \epsilon,$$

where $\epsilon \sim N(0, 1)$. Our aim is to select the active variables Z_i through their representation by the cubic B-spline sets/groups.

We compared GPER and GPQR [Group Penalized Quantile Regression Ouhourane et al. (2021)] at two locations $\tau \in \{0.5, 0.85\}$ for the penalties GLasso, GMCP and GSCAD. We computed four statistics, over 100 datasets replication:

- $|\hat{\mathcal{A}}|$: the average number of nonzero group variables $\hat{\boldsymbol{\beta}}_k \neq 0$ for $k = 1, \ldots, p$.
- p_a: proportion of the event $\mathcal{A} \subset \hat{\mathcal{A}}$, where $\mathcal{A}$ is the true active set of $\boldsymbol{\beta}^*$. When $\tau = 0.5, \mathcal{A} = \{G_6, G_{12}, G_{15}, G_{20}\}$ and when $\tau = 0.85, \mathcal{A} = \{G_1, G_6, G_{12}, G_{15}, G_{20}\}$.
- p_1: proportion of the event that $\{1\} \subset \hat{\mathcal{A}}$.
- FP: False Positive, the number of groups of variables with zero coefficients incorrectly included in the estimated model.
- AE: the absolute estimate error, defined by $\sum_{j=0}^{p} |\hat{\beta}_j - \beta_j|$.

From Table 1, one can see that GPER approach with the three penalties selects the true active groups, with the p_a statistic equals to 100%. On the other hand, the Size $|\hat{\mathcal{A}}|$ and FP statistics reveal that GPER with GMCP and GSCAD has tendency to provide more accurate sparse models compared to GLasso. The statistic p_1 shows how many times the heteroscedastic group variable, represented by Z_1, is selected in each model fit. For $\tau = 0.5$, it is expected that Z_1 will be not selected since it has no effect on the center of y (p_1 is less than 22%). However, for $\tau = 0.85$, the proportion of selecting Z_1 is greater than 84%, for GPER with all penalties. The GPER approach detects the effect of heteroscedastic variable Z_1, but, it can not estimate its effect value. When comparing expectiles to quantiles, expectile regression performs better in the tails of the distribution (results for $\tau = 0.85$); however, both methods behave similarly at the center of the distribution (results for $\tau = 0.5$). For the AE statistic, there is no significant difference among all methods and penalties. In order to do that, we take $\tau = 0.85$ for easy separation of the conditional mean and scale

Table 1 Simulation results of $|\hat{\mathcal{A}}|$, p_a, p_1, FP and AE for Scenario 1 based on 100 replications

τ	Method	Penalty	$\lvert\hat{\mathcal{A}}\rvert$	p_a (%)	p_1 (%)	FP	AE
0.50	Expectile	GLasso	11.02	100	22	7.02	60.47
		GMCP	6.38	100	18	2.38	60.64
		GSCAD	4.86	100	8	0.86	60.91
		GLLA-GMCP	6.24	100	15	2.24	60.65
		GLLA-GSCAD	4.89	100	11	0.89	60.01
	Quantile	GLasso	17.98	100	52	13.98	61.12
		GMCP	3.88	96	0	0.15	61.75
		GSCAD	4.22	97	2	0.22	64.36
		GLLA-GMCP	3.96	97	0	0.09	63.92
		GLLA-GSCAD	4.25	100	1	0.25	62.11
0.85	Expectile	GLasso	15.34	100	94	11.34	59.14
		GMCP	6.26	100	86	2.26	60.17
		GSCAD	6.36	100	84	2.36	60.89
		GLLA-GMCP	6.31	100	87	2.31	61.02
		GLLA-GSCAD	5.01	100	89	1.01	60.57
	Quantile	GLasso	20.43	100	75.21	16.43	61.76
		GMCP	12.24	98	52	8.24	59.77
		GSCAD	12.62	100	76.65	8.62	62.08
		GLLA-GMCP	12.55	100	49	12.45	60.87
		GLLA-GSCAD	13.01	100	71.17	9.01	61.31

The five statistics are calculated for GPER approach with all suggested group penalties

functions. Based on 100 independent runs, the following statistics are computed to evaluate the estimation performance and the sparsity recovery of the COGPER estimators:

- $|\hat{\mathcal{A}}_1|$, $|\hat{\mathcal{A}}_2|$: the average number of nonzero group variables for $\hat{\boldsymbol{\beta}}$ and for $\hat{\boldsymbol{\phi}}$, respectively, i.e., $\hat{\mathcal{A}}_1 = \{k, \hat{\boldsymbol{\beta}}_k \neq 0\}$ and $\hat{\mathcal{A}}_2 = \{k, \hat{\boldsymbol{\phi}}_k \neq 0\}$;
- p_{a_1}, p_{a_2}: proportion of the event $\mathcal{A}_1 \subset \hat{\mathcal{A}}_1$, $\mathcal{A}_2 \subset \hat{\mathcal{A}}_2$, where $\mathcal{A}_1$ and $\mathcal{A}_2$ are the active group sets of $\boldsymbol{\beta}^*$ and $\boldsymbol{\phi}^*$, respectively. In this first scenario, we have $\mathcal{A}_1 = \{G_6, G_{12}, G_{15}, G_{20}\}$ and $\mathcal{A}_2 = \{G_1^{\boldsymbol{\phi}}\}$.

In Table 2, the statistic p_{a_1} is always equals to 100% for COGPER with all penalties; i.e., all groups of variables that have effect on the mean are selected. On the other hand, the statistic p_{a_2} shows how many times the heterogeneous group G_1, estimated as $\hat{\boldsymbol{\phi}}_1$, is selected. Thus, one can notice that p_{a_2} is always greater than 90%. This shows that the COGPER approach can be used to detect the effect of the heteroscedastic groups, and can also estimate the amount of the heteroscedastic effect and separate it from the effect on the mean function.

Table 2 Simulation results of $|\hat{\mathcal{A}}_1|, |\hat{\mathcal{A}}_2|$, p_{a_1} and p_{a_2} for Scenario 1, based on 100 replications

τ	Penalty	$\lvert\hat{\mathcal{A}}_1\rvert$	$\lvert\hat{\mathcal{A}}_2\rvert$	p_{a_1} (%)	p_{a_2} (%)
0.85	GLasso	19.22	14.66	100	98
	GMCP	4.34	3.21	100	90
	GSCAD	4.04	2.96	100	94
	GLLA-GMCP	4.39	3.13	100	92
	GLLA-GSCAD	4.41	2.89	100	95

The four statistics are calculated for the COGPER approach with all suggested group penalties

In the first scenario, the active sets of the true groups of variables do not overlap, so the GPER can detect active groups of variables in the scale. In the second scenario, we assumed that the active set of groups for the mean overlaps with the active set for the scale. More precisely, the procedure for generating the predictors and the groups in this scenario was similar to the first scenario, however we generated the response variable from the following linear heteroscedastic model

$$Y = \underbrace{Z_2}_{G_2} + \underbrace{Z_5}_{G_5} + \underbrace{Z_{10}}_{G_{10}} + \underbrace{Z_{15}}_{G_{15}} + (\underbrace{\boldsymbol{\Phi}(Z_1)}_{G_1^{\boldsymbol{\phi}}} + \underbrace{\boldsymbol{\Phi}(Z_5)}_{G_5^{\boldsymbol{\phi}}})\epsilon,$$

where $\epsilon \sim N(0,1)$. This means that the active set of groups for the mean, $\mathcal{A}_1 = \{G_2, G_5, G_{10}, G_{15}\}$, overlaps with the active set of groups for the scale, $\mathcal{A}_2 = \{G_1^{\boldsymbol{\phi}}, G_5^{\boldsymbol{\phi}}\}$. The group G_1 has effect only on the scale, but G_5 has effect on both the mean and the scale (i.e. an overlapping group effect). In this scenario we set $K = 400$ and $n = 300$. Thus, the design matrix $\mathbf{X}$ has 1200 columns and 300 rows. All results are based on 100 data replications. Table 3 shows the results of COGPER for this scenario, based on the four statistics $|\hat{\mathcal{A}}_1|, |\hat{\mathcal{A}}_2|$, p_{a_1}, and p_{a_2}, defined earlier.

From Table 3, one can derive similar conclusions for COGPER as in Table 2. The statistic p_{a_1} is always equals to 100% for COGPER with all penalties. The statistic p_{a_2} shows how many times the estimated effect of the heterogeneous groups on the scale (i.e. $G_1^{\boldsymbol{\phi}}$ and $G_5^{\boldsymbol{\phi}}$) have non-zero values. This statistic is greater than 87%

Table 3 Simulation results of $|\hat{\mathcal{A}}_1|, |\hat{\mathcal{A}}_2|$, p_{a_1} and p_{a_2} for Scenario 2, based on 100 replications

τ	Penalty	$\lvert\hat{\mathcal{A}}_1\rvert$	$\lvert\hat{\mathcal{A}}_2\rvert$	p_{a_1} (%)	p_{a_2} (%)
0.85	GLasso	8.50	8.10	100	95
	GMCP	4.44	2.21%	100	87
	GSCAD	6.13	5.32	100	93
	GLLA-GMCP	4.52	3.07%	100	94
	GLLA-GSCAD	5.96	2.99	100	95

The four statistics are calculated for the COGPER approach with all suggested group penalties

for COGPER with all penalties. This confirms good performance of the COGPER approach in disentangling the heterogeneous overlapping group.

4.2 Checking KKT condition

Since we are updating each group at a time and cycling between groups until convergence, in this section, we numerically demonstrate that the proposed algorithms satisfy the KKT conditions, which means that the algorithms converge and find the right solution.

The KKT conditions are given in Appendix 5, for each method and each penalty. Here, we design the simulation model by using a modified simulation model in Yuan and Lin (2006). We simulate first the initial matrix of predictors $X_k, (k = 1, \ldots, K)$ from multivariate normal distribution with correlation $\rho = 0.5$ among the columns in the design matrix. Then, we considered $\{X_k, X_k^2, X_k^3\}$ as a group when fitting the two models GPER and COGPER, so the final predictor matrix has the number of variables $p = 3K$. The response variable was generated as follows

$$Y = \sum_{k=1}^{K} \left(\frac{2}{3} X_k - X_k^2 + \frac{1}{3} X_k^3 \right) \beta_k + \epsilon, \ \epsilon \sim N(0, \sigma^2),$$

where σ is chosen so that the signal-to-noise ratio (SNR) is 3 (i.e. $SNR = \|\mathbf{X}\boldsymbol{\beta}\|_2 / \sqrt{n}\sigma$) and $\beta_k = (-1)^k \exp(-(2k-1)/20)$. We considered two values for $K = 1000, 3000$, and we set $n = 100$.

Table 4 shows that all group-penalized expectile methods have zero violation count for the first scenario ($K = 1000$) and has also small violation counts for the second scenario ($K = 3000$). Thus, one can argue that all the proposed approaches are accurate algorithms that pass KKT checks without sever violation.

Of note, in all the simulation scenarios, we have focused on evaluation of the proposed method on the center (i.e., $\tau = 0.5$) and high expectiles (i.e., $\tau = 0.85$) of the conditional distribution of Y. This is because the error term, in all scenarios, is assumed to follow a normal distribution, which is symmetric. This means that the theoretical expectile is the same for the lower and upper locations that are symmetric to $\tau = 0.5$ (e.g., $\tau = 0.15$ and $\tau = 0.85$). We have conducted similar analysis for lower expectiles ($\tau = 0.15$). As expected, the results (not reported here) of both GPER and COGPER were similar to the results presented in Tables 1, 2, 3, and 4.

4.3 Comparison of running times

We compared the computational time of GPER and COGPER approaches versus the group penalized quantile regression (GPQR) method using the same example employed in the checking KKT conditions section, with $n = 100$ and $K = 1000$ ($p = 3000$). Table 5 shows the results of the three methods over a single data generation to estimate the optimal model, over a grid of 100 values of λ, based on the 5-fold cross-validation procedure. GPER has a faster running time than GPQR in the tails of the outcome distribution, and shows a similar running time to group

Table 4 Reported numbers are the average number of groups among K groups of variables that violated the KKT conditions check using GPER-GLasso, GPER-GMCP, GPER-GSCAD, COGPER-GLasso, COGPER-GMCP and COGPER-GSCAD

τ	*GLasso*	*GMCP*	*GSCAD*	*GLLA−GMCP*	*GLLA−GSCAD*
GPER					
$n = 100,\ K = 1000$					
0.50	0.01	0.01	0.01	0.00	0.01
0.85	0.00	0.00	0.01	0.00	0.00
$n = 100,\ K = 3000$					
0.50	0.03	0.05	0.04	0.04	0.03
0.85	0.09	0.07	0.05	0.10	0.10
COGPER					
$n = 100,\ K = 1000$					
0.50	0.78	0.63	0.64	0.71	0.72
0.85	2.29	2.12	2.15	2.98	2.05
$n = 100,\ K = 3000$					
0.50	1.12	1.10	1.10	1.12	1.12
0.85	3.02	2.94	3.25	3.27	2.90

Results are averaged over the λ sequence of 100 values and averaged over 50 independent runs

Table 5 Comparison of the running time of GPER, GPQR and COGPER

τ	*GLasso*	*GMCP*	*GSCAD*	*GLLA−GMCP*	*GLLA−GSCAD*
GPER					
$n = 100,\ K = 1000$					
0.50	4.91	16.37	10.22	6.02	6.18
0.85	4.87	12.42	8.07	4.97	4.85
GPQR					
$n = 100,\ K = 1000$					
0.50	5.89	11.42	11.48	19.75	14.19
0.85	6.93	14.00	14.31	20.25	16.38
COGPER					
$n = 100,\ K = 1000$					
0.50	78.81	129.0	86.61	73.61	75.46
0.85	104.5	119.5	92.71	71.03	72.13

Results are based on a single data generation, with $n = 100$ and $K = 1000$ ($p = 3 \times 1000$), to estimate the optimal model over a grid of 100 values of λ, using the 5-fold cross-validation procedure

penalized quantile in the center of the distribution. However, the running time of COGPER is significantly greater compared to GPER because COGPER estimates a vector $(\boldsymbol{\beta}, \boldsymbol{\phi})$ that is twice the size of the estimate vector of GPER.

5 Real data

5.1 The birth weight data

This dataset was collected by the Medical Center in Springfield, Massachusetts. It was used as an illustration example for demonstrating various aspects of regression modeling (Hosmer Jr et al. 2013; Venables and Ripley 2013). It was also used to illustrate both the group penalized least squares and quantile regression models (Yuan and Lin 2006; Hashem et al. 2016). The dataset records the birth weights of 189 babies in kilograms and eight predictors concerning their mothers. Among the eight predictors, two are continuous (mother's age in years and mother's weight in pounds at the last menstrual period), and six are categorical: mother's race with three levels (white, black or other), smoking status during pregnancy (yes = 1 or no = 0), number of previous premature labours with three levels (0, 1 or 2 or more), history of hypertension (yes = 1 or no = 0), presence of uterine irritability (yes = 1 or no = 0), number of physician visits during the first trimester with four levels (0, 1, 2 or 3 or more).

A preliminary analysis conducted in Venables and Ripley (2013) suggests that non-linear effects of both mother's age and weight may exist. Thus, in our analysis the two continuous variables were represented through two third-order polynomials, i.e. the two continuous variables were considered as groups of three predictors. The categorical variables were considered as groups using dummy variables. So, each categorical predictor of l levels is represented by $l-1$ dummy variables. In summary, we have a total $p = 16$ predictors (i.e. 6 continuous and 10 dummy variables) that are grouped in $K = 8$ groups. A preliminary analysis of this data can be found in the `grpreg` R package.

The goal of this study is to identify the risk factors associated with the baby birth weight response variable. In particular, we aim to explore the effect of the mother smoking during pregnancy, which is known to have a heterogeneous effect on the baby birth weight (Tang et al. 2021). An ANOVA analysis can be investigated to evaluate differences in the birth weight of the babies between the two groups: smoking versus non-smoking mothers. An F-value of 7.038 leads to a p-value of 0.00867; we can conclude that the mother's smoking status is significantly associated with the baby birth weight. However, ANOVA is a mean-based test. Thus, we adjusted both GPER and COGPER for three values of $\tau = 0.15, 0.50, 0.85$, with aim to capture the heterogeneous effect of the mother smoking in different expectiles/locations of the conditional distribution of the response variable.

We conducted two analyses for this data. Firstly, we fitted the GPER and COGPER with the group Lasso penalty for all 189 babies with 5-fold CV to obtain the optimal models for the three locations ($\tau = 0.15, 0.50, 0.85$). Figure 3 shows the coefficient path solutions of GPER and COGPER for this analysis. The solution of the optimal models is indicated by vertical lines, which indicate the optimal values of λ for each model fit. From this figure, one can notice that both GPER and COGPER tend to select three important groups of variables: mother's race (green),

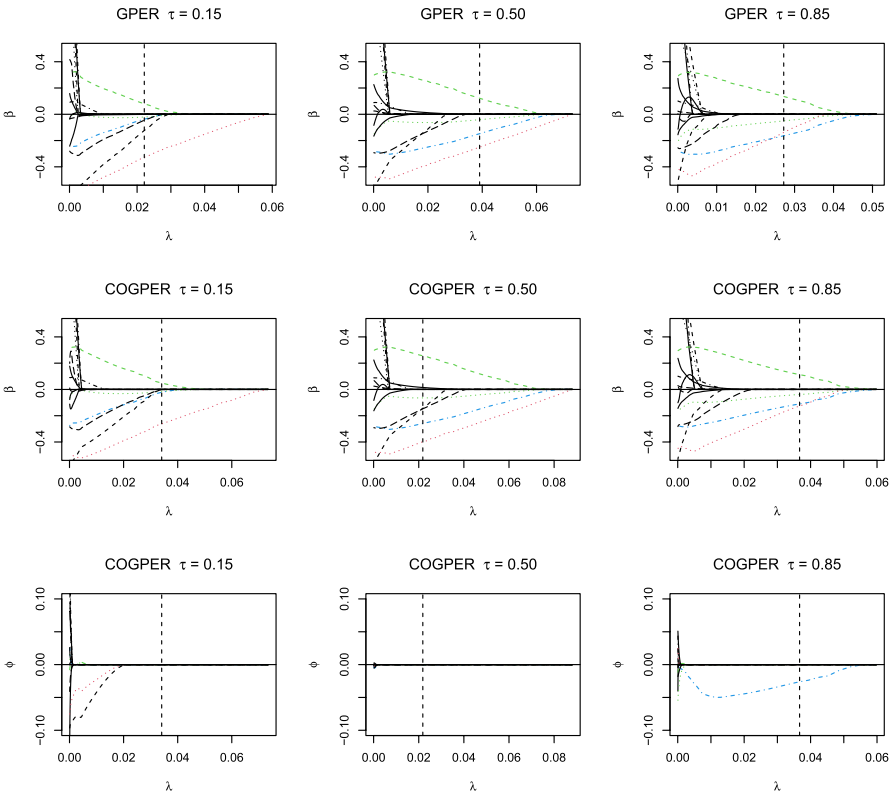

Fig. 3 At the left and from top to bottom, the coefficient paths of GPER with $\tau = 0.15$, 0.50 and 0.85 respectively, are shown as a function of the tuning parameter. At the middle and right columns, the coefficients paths $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ of COGPER respectively with $\tau \in \{0.15, 0.55, 0.85\}$

smoking status (blue) and uterine irritability (red) for all $\tau = 0.15, 0.50, 0.85$. The coefficient values corresponding to smoking-status effects are estimates of the total effect on both mean and scale functions when using GPER. This cannot tell us if this predictor has an overlapping effect (i.e. if this predictor is also relevant or not to the scale function). Interestingly, Fig. 3 (bottom right panel) shows that COGPER selects the scale coefficient, $\hat{\phi}$, corresponding to smoking-status as a non zero effect, for $\tau = 0.85$ (blue path solution). This indicates that smoking-status might be a heterogeneous overlapping predictor. COGPER provides also estimates of the scale effect and thus it distinguishes it from the mean function effect.

In the second analysis, we randomly divided the data into a training sample of two-thirds observations and the remainder making up a test data. For the three values of τ, both GPER and COGPER were fitted to the training data to obtain the parameter estimates of the optimal models, where the optimal λ values were selected by 5-fold CV. The performance of the methods in this analysis is based on the following statistics, which are calculated on the test data:

- the estimates of the effects of the three variables that have been selected as relevant in the first analysis: Smoking status during pregnancy (blue), mother's race (green) and presence of uterine irritability (red). These estimates are calculated based on the optimal model of the training data analysis.
- the model-size statistic (*MS*), which is defined as the number of selected groups for GPER. For COGPER, two MS statistics are needed: MS_1 to count the relevant groups for the mean function (i.e. $\hat{\boldsymbol{\beta}}_k \neq 0$), and MS_2 for counting the relevant groups for the scale function (i.e. $\hat{\boldsymbol{\phi}}_k \neq 0$). The MS statistic estimation is also based on the results of the optimal model of the training data analysis.
- the expectile-based prediction error (EPE), which is calculated on the test data, and is defined as

$$EPE_{\text{gper}} = \frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} \rho_\tau(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})$$

for GPER, and

$$EPE_{\text{cogper}} = \frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} (y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}})^2 + \rho_\tau(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} - \mathbf{x}_i^\top \hat{\boldsymbol{\phi}})$$

for COGPER.

The whole procedure was repeated 100 times, and we reported the empirical distribution (boxplots) of the aforementioned statistics in Figs. 4 and 5.

Figure 4 highlights the results of the second analysis of GPER and COGPER for the estimates of the effects of the three variables that have been selected as relevant in the first analysis. That is, based on 100 replications, this figure reports the empirical distribution of the point estimates $\hat{\boldsymbol{\beta}}_k$'s of the three important variables for the optimal model, which is obtained using 5-fold CV in the analysis of the training datasets. The empirical distribution of the estimates of the predictors' effects through the 100 replications demonstrates the consistency of both approaches to select the three variables as relevant predictors, in particular for $\tau \in \{0.5, 0.85\}$. Interestingly, Fig. 4 (bottom middle panel) shows also that the distribution of the estimates of the smoking-status effect on the variance is non-null when fitting COGPER for the location $\tau = 0.85$.

Figure 5 shows the results of the second analysis of GPER and COGPER for the MS and EPE statistics. The top, middle, and right panels in the left of Fig. 5, which report the empirical distribution of the MS statistic for both GPER (MS) and COGPER (MS_1 and MS_2), confirms also the results of Fig. 4. In fact, for both methods the average, over 100 replication, of the MS statistic equals to 3 (i.e. the average over 100 run, of the number of active groups at each run, for which $\hat{\boldsymbol{\beta}}_k \neq 0$, is approximately equals to 3). This corresponds to the three variables that have been declared as relevant in the first analysis. The MS_2 distribution of COGPER confirms also the presence of an overlapping group, which corresponds to the heterogeneous effect of the mother smoking-status predictor. Finally, in terms of prediction, the EPE statistic results of Fig. 5 (right two panels) show that the better fit for both models seems to be at location $\tau = 0.85$. This might again emphasize the usefulness of

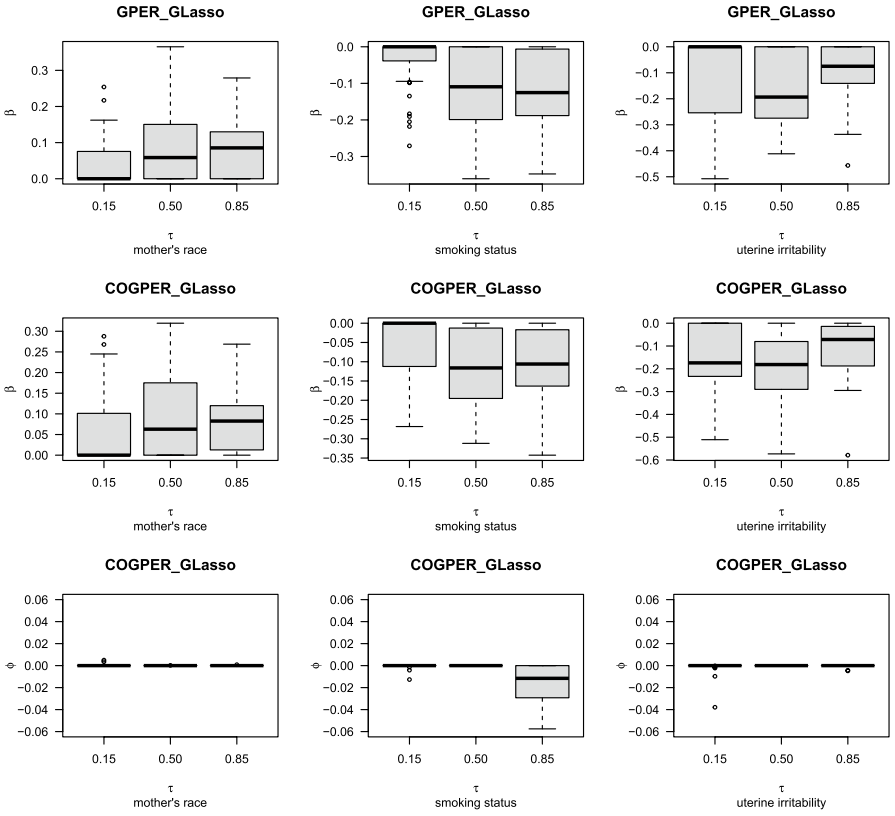

Fig. 4 From left to right column, the box-plot of the coefficient values for mother's race, smoking status and uterine irritability respectively. The GPER and COGPER are fitted with $\tau \in \{0.15, 0.50, 0.85\}$

both methods for allowing flexible exploration of the response–predictors relationship. All these results are also in agreements with several studies that have revealed a strong and heterogeneous relationship between mother smoking-status and baby birth weight (Spady et al. 1986; Wilcox 1993; Chiolero et al. 2005).

5.2 Gene-based analysis of DNA methylation data near *BLK* gene

This section considers illustration of the proposed approach via a DNA methylation data analysis. DNA methylation is an important epigenetic modification that can modulates gene expression (either activate or repress gene expression) (Yousefi et al. 2022). The methylation level at a genomic position is measured as a proportion between 0 and 1, and it refers to the extent to which this specific position is methylated. Methylation occurs, in general, at CpG sites, defined as specific genomic locations where a cytosine nucleotide is followed immediately by a guanine nucleotide in the DNA sequence. DNA methylation is known to exhibit differentiation across cell types (McGregor et al. 2016). Thus, in this section we analyse DNA methylation

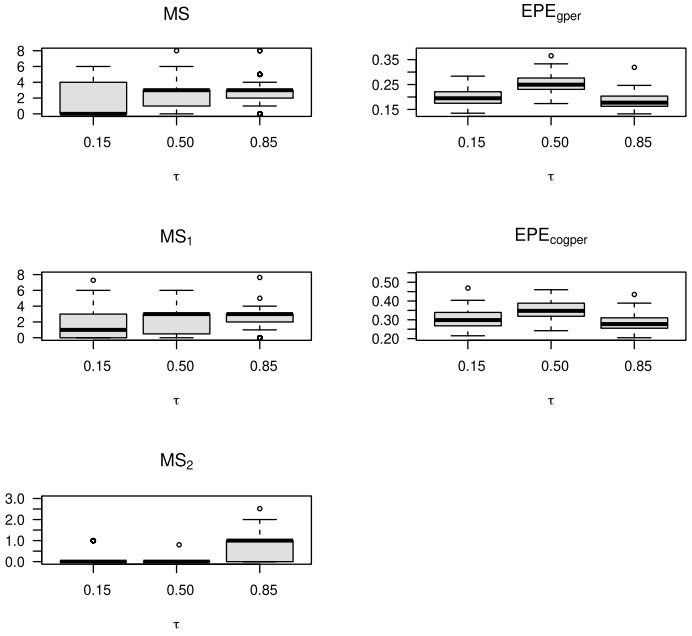

Fig. 5 Comparison of the number of selected groups (model size) and the expectile-based prediction error (EPE), based on 100 replications, for the birth weight dataset. The GPER and COGPER with GLasso penalty are fitted for three locations $\tau \in \{0.15, 0.50, 0.85\}$

data to validate the performance of our method on detecting genomic regions (groups of predictors) that are differentially methylated between three cell types.

The data considered in this analysis consists of methylation levels of 5,986 CpG sites (i.e. predictors) within a genomic region with around 2 Millions base (Mb) pairs of Chromosome 8 (2Mb region start-end positions: 10321522–12391296), measured on 40 samples using bisulfite sequencing (Lakhal-Chaieb et al. 2017). Each sample corresponds to one of three cell types: B cells (8 samples), T cells (19 samples), or Monocytes (13 samples). The 40 samples are obtained from whole blood collected on a cohort of healthy individuals from Sweden. The methylation levels vary between B-cell types and T-/Monocyte-cell types around the studied genomic region. In fact, B-cells are known to be hypomethylated near the *BLK* gene, compared to the other two cell types (Hertz et al. 1999). Thus, the cell type is considered as the response variable, where $y = 1$ corresponds to B-cell samples and $y = 0$ corresponds to T- and Monocyte-cell type samples.

We proceeded as follows in order to form groups of predictors (CpGs sites): (1) we extracted all genes belongings to the 2Mb region and their start-end genomic positions using `biomart` R package. We obtained $K = 36$ genes fall within this region in total. (2) We used prior information about the genomic position of each CpG site and assigned each CpG to a corresponding gene/group based on its base pair coordinate. Specifically, we considered that a CpG belongs to a gene/group if its genomic position is between the start and end positions of that gene. In total, 4,427 of all the 5,986 CpG sites spread over the $K = 36$ genes. The size of the studied

groups ranges between 1 and 756, with 398 CpG sites falling between the start-end coordinates of the BLK gene.

This analysis aims to validate the performance of our methods on detecting the group of CpG sites belonging to the *BLK* gene as a Differentially Methylated Region (DMR) for the 0–1 response variable, and to test the power of GPER in classification. The classification function is $\mathbb{1}(\text{fitted value} > 0.5)$, where $\mathbb{1}(A)$ is the indicator function which equals 1 if A is true and 0 if A is false.

Notice that the analysis results of this dataset using GPER with the non-convex penalties, GMCP and GSCAD, were very similar. Thus, we reported only the results of GPER with GMCP penalty.

In Fig. 6, the x-axis and the y-axis correspond respectively to the genomic position of the CpGs and the coefficient values of the optimal solutions chosen by 5-fold CV. We can observe that the region around 11.3 Mb with size 150kb is significantly detected/selected by the group expectile methods with $\tau = 0.85$ but not with GPER with $\tau \in (0.15, 0.5)$. This region is known as a DMR between DNA methylation profiles of B-cells and T/Mono cells (Turgeon et al. 2016). This observation is consistent with our analysis of this data using group quantile regression (Ouhourane et al. 2021) and a study conducted by Lakhal-Chaieb et al. (2017).

A second analysis of this DNA methylation data aims to show the advantages of the proposed group penalized expectile regression approaches for classification. This analysis emphasises GPER and COGPER utility when predicting the sample cell type (i.e. observation's class) from a tail location of the distribution (i.e. $\tau = 0.85$), in comparison with group penalized least-squares regression

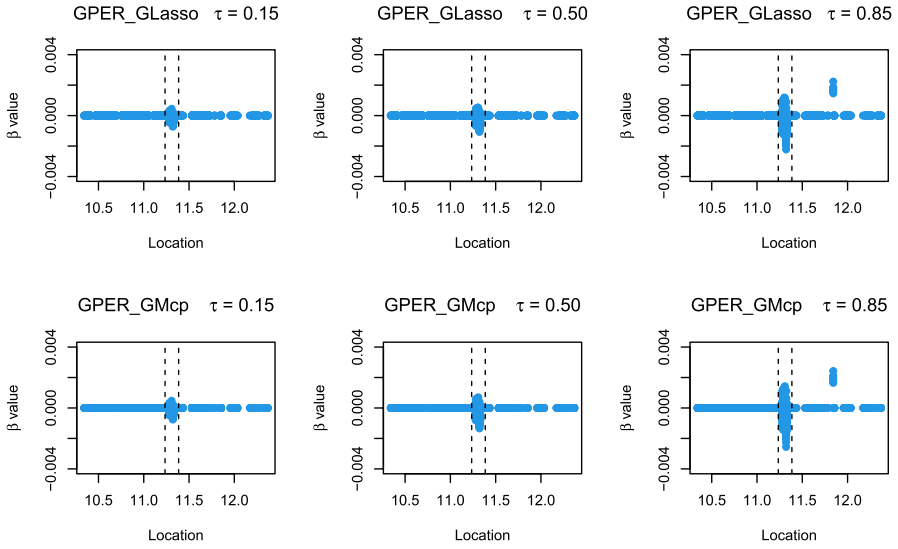

Fig. 6 At the top and from left to right, the optimal values (5-fold CV) for the regression coefficients of GLasso with $\tau = 0.15, 0.50, 0.85$ respectively, are shown as a function of a real genomic position. At the bottom, the coefficients' values of the GMcp with the same values of τ are displayed

($\tau = 0.5$) and two well-known group supervised classification methods: Group Support Vector Machine (GSVM) and group logistic regression (GLogit). The latter are both implemented in the `gglasso` R package (Yang and Zou 2015). We randomly divided the data into a training sample of 30 observations with the remainder making up a test data. The model is fitted to the training data and the misclassification error rate (MER) is calculated on the test data. The MER is defined as the ratio of the number of misclassified observations to the total number of observations in the test data. The tuning parameters are selected by 5-fold CV on the training data, and we adjust our models at two different locations $\tau \in \{0.5, 0.85\}$ in this analysis. The whole procedure is executed 100 times. The results of this analysis are presented in Fig. 7.

In Fig. 7, the DMR region is 100% selected by the GPER approach with both GLasso and GMCP penalties, for the location $\tau = 0.85$. However, it is selected with a rate less than 80% with the group least squares (GPER with $\tau = 0.5$), GSVM and GLogit methods. In terms of the MSE performance (last panel of Fig. 7), the fit of GPER with GLasso at the tail of the response variable ($\tau = 0.85$) gives the best the best classification error. This, again, confirms the utility of GPER for the classification regression framework. The results of GPER with $\tau = 0.15$ (not reported here) were similar to those of GPER with $\tau = 0.5$.

Of note, the analysis of this DNA methylation data using COGPER does not reveal any overlapping predictor (i.e. $\hat{\boldsymbol{\phi}}_k = 0$, for all k). The conclusions of COGPER in terms of selecting the *BLK* gene as a DMR were relatively similar to those GPER; but COGPER seemed to be less consistent compared to GPER in this DNA methylation analysis (results not reported here). We decided to not report COGPER to provide a clear summary analysis of this data and to avoid results redundancy.

6 Discussion

In this paper, we have proposed the group penalized expectile regression approaches (GPER and COGPER) for selection of grouped variables. Both approaches, (CO) GPER, handle most known group penalties in the literature, namely, group Lasso, group MCP, group SCAD, and group LLA penalties. (CO)GPER are implemented in computationally-efficient groupwise-majorizatoin-descent algorithms. We have showed theoretically that, under some regularity conditions, our proposed methods enjoy the consistency property for the group Lasso penalty, and we have proved the convergence of (CO)-GPER-GLLA algorithms to the oracle estimator in two steps for the non-convex penalties. The results from our simulation studies have shown that the proposed methods provide appropriate sparse group-variable selection and accurate estimation.

GPER and COGPER approaches (and their corresponding penalties) have their specific characteristics. When analyzing real datasets, we recommend that users first apply GPER with GLasso at different locations to detect heteroscedastic predictors, which affect the variance of the outcome and might potentially affect also its mean. To further refine the analysis, COGPER with non-convex penalties can be used to identify predictors that impact only the variance, if any exist. Additionally, if the

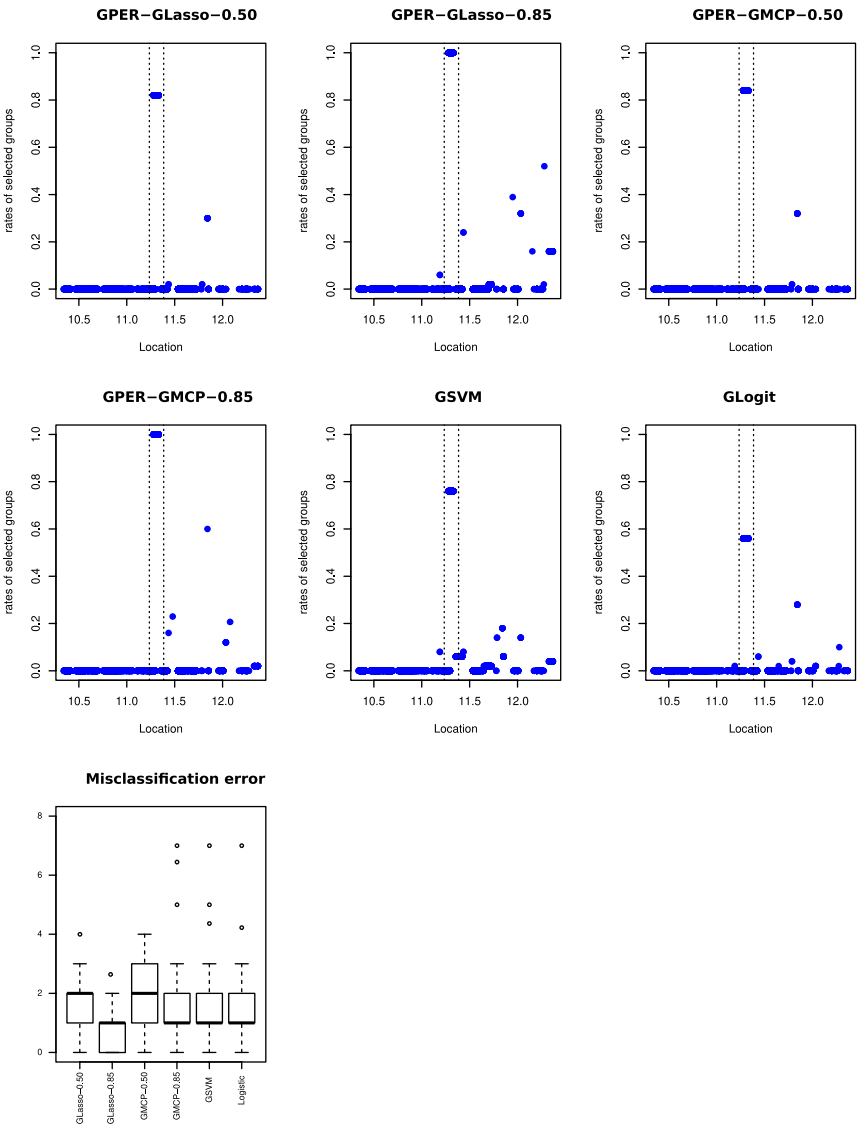

Fig. 7 Comparison of the proportion of selected genes for the DNA methylation data. At the top from left to right, the proportion of GPER-GLasso for $\tau \in \{0.50, 0.85\}$ and GPER-GMCP with $\tau = 0.50$ are shown as a function of the genomic position. The middle from left to right shows the proportion of GPER-GMCP with $\tau = 0.50$, GSVM and GLogit. The bottom row shows the misclassification error for all these methods cited above. The x and y axes correspond respectively to the genomic position, t_j, of the j-th CpG site and the proportion of non-zero $(\hat{\beta}_j)_{1 \leq j \leq 4427}$

primary aim of the analysis is prediction accuracy, then one can use (CO)GPQR with the Glasso penalty. If selection consistency and sparsity are the main goals, then GPQR with the GSCAD or GMCP penalties would be preferred.

It is well known that the asymmetric squared loss function in (CO)GPER might be sensitive to outliers either in the response and/or in the covariates. Recently, Zhao et al. (2022) have developed a robust expectile regression approach for ultrahigh dimensional heavy-tailed heterogeneous data. The proposed loss function in Zhao et al. (2022) differs from the asymmetric squared loss function of (CO)GPER only in the tail, where it is peace-wise linear in the extremes to down-weight the outliers. Zhao et al. (2022) have also provided attractive theoretical results for their proposed estimator. The extension of our framework to robust expectile regression in the presence of group structure among the covariates could be an interesting avenue to explore.

Asymmetric regression models have been increasingly investigated in the last decade. The extremile regression (Daouia et al. 2019, 2021) is a new attractive asymmetric least squares analog of quantile and expectile regression, which is found to be a useful descriptor of the tail of the response distribution, especially for long-tailed distributions. Although the extremile loss function is defined through a power transformation of the cumulative distribution function of the response variable, the extremile regression estimator has a closed form and can be calculated using an iterative reweighted least squares algorithm. This makes it computationally very attractive. Investigating penalized extremile regression in high-dimensional settings might be an interesting avenue of research in penalized asymmetric regression.

Appendix 1: Proof of Proposition 1

Proof Notice first that the expectile loss function $\Psi_\tau(.)$ has a Lipschitz continuous derivative $\Psi'_\tau(.)$. That is, one can verify that

$$|\Psi'_\tau(u) - \Psi'_\tau(v)| \leq c|u - v| \quad \forall u, v \in \mathbb{R}, \tag{28}$$

where $c = 2\max(\tau, 1-\tau)$.

For $\boldsymbol{\beta}_k$ and $\tilde{\boldsymbol{\beta}}_k$, let $\mathbf{V}_k = \boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k$ and define $h(t) = \Psi_\tau(\tilde{\boldsymbol{\beta}}_k + t\mathbf{V}_k, \tilde{\boldsymbol{\beta}}_{-k})$.

Then, we have $h(0) = \Psi_\tau(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k})$ and $h(1) = \Psi_\tau(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k})$.

By the mean value theorem, there exits $a \in (0, 1)$ such that

$$h(1) = h(0) + h'(a) = h(0) + h'(0) + (h'(a) - h'(0)). \tag{29}$$

Noticed that

$$h'(t) = n^{-1} \sum_{i=1}^{n} \mathbf{x}_{i,k}^\top \mathbf{V}_k \Psi'_\tau(y_i - \mathbf{x}_{i,-k}^\top \tilde{\boldsymbol{\beta}}_{-k} - \mathbf{x}_{i,k}^\top (\tilde{\boldsymbol{\beta}}_k + t\mathbf{V}_k)),$$

which leads to

$$h'(0) = (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^\top \nabla_k \Psi_\tau(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}),$$

and

$$
\begin{aligned}
|\, h'(a) - h'(0)| &= |n^{-1} \sum_{i=1}^{n} \mathbf{x}_{i,k}^{\top} \mathbf{V}_k [\Psi'_{\tau}(y_i - \mathbf{x}_{i,-k}\tilde{\boldsymbol{\beta}}_{-k} - \mathbf{x}_{i,k}(\tilde{\boldsymbol{\beta}}_k + a\mathbf{V}_k)) \\
&\quad - \Psi'_{\tau}(y_i - \mathbf{x}_{i,-k}\tilde{\boldsymbol{\beta}}_{-k} - \mathbf{x}_{i,k}\tilde{\boldsymbol{\beta}}_k)]| \\
&\leq n^{-1} \sum_{i=1}^{n} |\mathbf{x}_{i,k}^{\top} \mathbf{V}_k| |\Psi'_{\tau}(y_i - \mathbf{x}_{i,-k}\tilde{\boldsymbol{\beta}}_{-k} - \mathbf{x}_{i,k}(\tilde{\boldsymbol{\beta}}_k + a\mathbf{V}_k)) \\
&\quad - \Psi'_{\tau}(y_i - \mathbf{x}_{i,-k}\tilde{\boldsymbol{\beta}}_{-k} - \mathbf{x}_{i,k}\tilde{\boldsymbol{\beta}}_k)| \\
&\overset{(a)}{\leq} n^{-1} \sum_{i=1}^{n} |\mathbf{x}_{i,k}^{\top} \mathbf{V}_k| c |a\mathbf{x}_{i,k}^{\top} \mathbf{V}_k| \\
&\leq cn^{-1} \sum_{i=1}^{n} |\mathbf{x}_{i,k}^{\top} \mathbf{V}_k|^2 \\
&\leq cn^{-1} \mathbf{V}_k^{\top} \mathbf{x}_k^{\top} \mathbf{x}_k \mathbf{V}_k.
\end{aligned}
$$

Inequality (a) is due to the Eq. (28). Plugging the last inequality into (29), we have

$$
\begin{aligned}
\Psi_{\tau}(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) \leq & \Psi_{\tau}(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^{\top} \nabla_k \Psi_{\tau}(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) \\
& + cn^{-1} (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^{\top} \mathbf{x}_k^{\top} \mathbf{x}_k (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k).
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
R_{\tau}(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) = & \Psi_{\tau}(\boldsymbol{\beta}_k, \tilde{\boldsymbol{\beta}}_{-k}) + P_{\lambda}(\|\boldsymbol{\beta}_k\|_2) \\
\leq & \Psi_{\tau}(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^{\top} \nabla_k \Psi_{\tau}(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) \\
& + cn^{-1} (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^{\top} x_k^{\top} \mathbf{x}_k (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) + P_{\lambda}(\|\boldsymbol{\beta}_k\|_2) \\
\leq & \Psi_{\tau}(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^{\top} \nabla_k \Psi_{\tau}(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) \\
& + \frac{\gamma_k}{2} (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k)^{\top} (\boldsymbol{\beta}_k - \tilde{\boldsymbol{\beta}}_k) + P_{\lambda}(\|\boldsymbol{\beta}_k\|_2),
\end{aligned}
$$

where γ_k is the largest eigenvalue of the matrix $2\max(1-\tau, \tau)\dfrac{\mathbf{x}_k^{\top}\mathbf{x}_k}{n}$.

This ends the proof of Proposition 1. □

Appendix 2: Proof of Proposition 2

For GSCAD penalty:

The KKT conditions of the objective function in Eq. (11) of the main manuscript can be written as

$$
-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \frac{\partial P_{\lambda}(\|\boldsymbol{\beta}_k\|_2)}{\partial \boldsymbol{\beta}_k} = 0,
$$

where $\mathbf{Z}_k = -\nabla_k \Psi_{\tau}(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + \gamma_k \tilde{\boldsymbol{\beta}}_k$.

- If $\|\boldsymbol{\beta}_k\|_2 \leq \lambda$, then $-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \lambda w_k \mathbf{u} = 0$, where $\mathbf{u}$ is the sub-gradient and $\|\mathbf{u}\|_2 \leq 1$.
  - If $\boldsymbol{\beta}_k = 0$, then we have

    $$-\mathbf{Z}_k + \lambda w_k \mathbf{u} = 0,$$

    which implies

    $$\|\mathbf{Z}_k\|_2 \leq \lambda w_k.$$

  - If $\boldsymbol{\beta}_k \neq 0$, then we have

    $$-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \lambda w_k \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} = 0.$$

    Applying the l_2 norm to the least equality, we have

    $$\|\mathbf{Z}_k\|_2 = \gamma_k \|\boldsymbol{\beta}_k\|_2 + \lambda w_k,$$

    which implies

    $$\|\mathbf{Z}_k\|_2 \leq \lambda(w_k + \gamma_k).$$

    Moreover, we have

    $$-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \lambda w_k \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = 0 \text{ (since } \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2}).$$

    Then, we obtain

    $$\boldsymbol{\beta}_k = \frac{1}{\gamma_k} \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} (\|\mathbf{Z}_k\|_2 - \lambda w_k).$$

- If $\lambda \leq \|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda$, then

  $$-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \frac{\theta \lambda w_k}{\theta - 1} \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} - \frac{w_k}{\theta - 1} \boldsymbol{\beta}_k = 0.$$

  It follows that

  $$\mathbf{Z}_k = [\gamma_k + \frac{w_k}{\theta - 1} (\frac{\theta \lambda}{\|\boldsymbol{\beta}_k\|_2} - 1)] \boldsymbol{\beta}_k,$$

  which implies that

  $$\|\mathbf{Z}_k\|_2 = (\gamma_k - \frac{w_k}{\theta - 1}) \|\boldsymbol{\beta}_k\|_2 + \frac{w_k \lambda \theta}{\theta - 1} \tag{30}$$

  and

$$\frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2}.$$

Thus, combining the condition $\lambda \leq \|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda$ and Eq. (30), we get

$$\lambda(\gamma_k + w_k) \leq \|\mathbf{Z}_k\|_2 \leq \gamma_k \theta \lambda$$

and

$$\boldsymbol{\beta}_k = \frac{1}{\gamma_k - \frac{w_k}{\theta - 1}} \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} (\|\mathbf{Z}_k\|_2 - \lambda w_k \frac{\theta}{\theta - 1}).$$

- If $\|\boldsymbol{\beta}_k\|_2 \geq \theta\lambda$, then we have

$$-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k = 0,$$

which implies

$$\|\mathbf{Z}_k\|_2 \geq \gamma_k \theta \lambda$$

and

$$\boldsymbol{\beta}_k = \frac{1}{\gamma_k} \mathbf{Z}_k.$$

This ends the proof of Proposition 2 for GSCAD penalty. □

For GMCP penalty

Again, the KKT conditions of the objective function in Eq. (11) of the main manuscript can be written as

$$-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \frac{\partial P_\lambda(\|\boldsymbol{\beta}_k\|_2)}{\partial \boldsymbol{\beta}_k} = 0,$$

where $\mathbf{Z}_k = -\nabla_k \boldsymbol{\Psi}_\tau(\tilde{\boldsymbol{\beta}}_k, \tilde{\boldsymbol{\beta}}_{-k}) + \gamma_k \tilde{\boldsymbol{\beta}}_k$.

- If $\|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda$, then we have

$$-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \lambda \mathbf{u} - \frac{w_k}{\theta} \boldsymbol{\beta}_k = 0,$$

where $\mathbf{u}$ is the sub-gradient and $\|\mathbf{u}\|_2 \leq 1$.

- If $\boldsymbol{\beta}_k = 0$, then we obtain

$$-\mathbf{Z}_k + \lambda w_k \mathbf{u} = 0.$$

Thus, we have

$$\|\mathbf{Z}_k\|_2 \leq \lambda w_k.$$

- If $\boldsymbol{\beta}_k \neq 0$, then we get

$$-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k + \lambda w_k \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} - \frac{w_k}{\theta}\boldsymbol{\beta}_k = 0.$$

Applying the l_2 norm to the last equality, we obtain

$$\|\mathbf{Z}_k\|_2 = (\gamma_k - \frac{w_k}{\theta})\|\boldsymbol{\beta}_k\|_2 + \lambda w_k.$$

Combining the condition $\|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda$ and the last two equations, we have

$$\|\mathbf{Z}_k\|_2 \leq \gamma_k \theta \lambda$$

and

$$\boldsymbol{\beta}_k = \frac{1}{\gamma_k - \frac{w_k}{\theta}} \frac{\mathbf{Z}^{(k)}}{\|\mathbf{Z}_k\|_2}(\|\mathbf{Z}_k\|_2 - \lambda w_k).$$

- If $\|\boldsymbol{\beta}_k\|_2 \geq \theta\lambda$, then we have $-\mathbf{Z}_k + \gamma_k \boldsymbol{\beta}_k = 0$. This implies that

$$\|\mathbf{Z}_k\|_2 \geq \gamma_k \theta \lambda \quad \text{and} \quad \boldsymbol{\beta}_k = \frac{1}{\gamma_k}\mathbf{Z}_k.$$

This ends the proof of Proposition 2 for GMCP penalty. □

Appendix 3: Proof of Proposition 5

The KKT conditions of the objective functions in (23) and (24) of the main manuscript can be written as:

$$-\mathbf{Z}_k + 2(1+c)\gamma_k \boldsymbol{\beta}_k + \frac{\partial P_{\lambda_1}(\|\boldsymbol{\beta}_k\|_2)}{\partial \boldsymbol{\beta}_k} = 0$$

and

$$-\mathbf{W}_k + 2c\gamma_k \boldsymbol{\phi}_k + \frac{\partial P_{\lambda_2}(\|\boldsymbol{\phi}_k\|_2)}{\partial \boldsymbol{\phi}_k} = 0,$$

where $Z_k = \mathbf{U}_k^{0.5} + \mathbf{U}_k^{\tau} + 2(1+c)\gamma_k \widetilde{\boldsymbol{\beta}}_k$ and $\mathbf{W}_k = \mathbf{U}_k^{\tau} + 2c\gamma_k \widetilde{\boldsymbol{\phi}}_k$.

For GLasso penalty

We have

$$-\mathbf{Z}_k + 2(1+c)\gamma_k \boldsymbol{\beta}_k + \lambda_1 \omega_k \mathbf{u} = 0$$

and

$$-\mathbf{W}_k + 2c\gamma_k \boldsymbol{\phi}_k + \lambda_2 u_k \mathbf{v} = 0,$$

where $\mathbf{u}$ and $\mathbf{v}$ are the sub-gradient of $P_{\lambda_1}(.)$ at $\boldsymbol{\beta}_k$ and $P_{\lambda_2}(.)$ at $\boldsymbol{\phi}_k$ repectively. So, we have $\|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1$.

- If $\boldsymbol{\beta}_k = 0$ and $\boldsymbol{\phi}_k = 0$, then we get

$$-\mathbf{Z}_k + \lambda_1 \omega_k \mathbf{u} = 0, \text{ and } -\mathbf{W}_k + \lambda_2 u_k \mathbf{v} = 0,$$

which implies that

$$\|\mathbf{Z}_k\|_2 \leq \lambda_1 \omega_k, \text{ and } \|\mathbf{W}_k\|_2 \leq \lambda_2 u_k.$$

- If $\boldsymbol{\beta}_k \neq 0$ and $\boldsymbol{\phi}_k \neq 0$, then

$$-\mathbf{Z}_k + 2(1+c)\gamma_k \boldsymbol{\beta}_k + \lambda_1 \omega_k \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} = 0$$

and

$$-\mathbf{W}_k + 2c\gamma_k \boldsymbol{\phi}_k + \lambda_2 u_k \frac{\boldsymbol{\phi}_k}{\|\boldsymbol{\phi}_k\|_2} = 0.$$

Moreover, we have

$$-\mathbf{Z}_k + 2(1+c)\gamma_k \boldsymbol{\beta}_k + \lambda_1 w_k \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = 0 \left(\text{since } \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2}\right)$$

and

$$-\mathbf{W}_k + 2c\gamma_k \boldsymbol{\phi}_k + \lambda_2 u_k \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_2} = 0 \left(\text{since } \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_2} = \frac{\boldsymbol{\phi}_k}{\|\boldsymbol{\phi}_k\|_2}\right),$$

which implies

$$\boldsymbol{\beta}_k = \frac{1}{2(1+c)\gamma_k} \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} (\|\mathbf{Z}_k\|_2 - \lambda_1 w_k)$$

and

$$\boldsymbol{\phi}_k = \frac{1}{2c\gamma_k} \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_2} (\|\mathbf{W}_k\|_2 - \lambda_2 u_k).$$

For GSCAD penalty

- If $\|\boldsymbol{\beta}_k\|_2 \leq \lambda_1$ and $\|\boldsymbol{\phi}_k\|_2 \leq \lambda_2$, then

$$-\mathbf{Z}_k + 2(1+c)\gamma_k \boldsymbol{\beta}_k + w_k \lambda_1 \mathbf{u} = 0$$

and

$$-\mathbf{W}_k + 2c\gamma_k \boldsymbol{\phi}_k + u_k \lambda_2 \mathbf{v} = 0,$$

where $\mathbf{u}$ and $\mathbf{v}$ are the sub-gradient of $P_{\lambda_1}(.)$ at $\boldsymbol{\beta}_k$ and $P_{\lambda_2}(.)$ at $\boldsymbol{\phi}_k$ repectively. So, we have $\|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1$.

- If $\boldsymbol{\beta}_k = 0$ and $\boldsymbol{\phi}_k = 0$, then we obtain

$$-\mathbf{Z}_k + \lambda_1 \omega_k \mathbf{u} = 0 \quad \text{and} \quad -\mathbf{W}_k + \lambda_2 u_k \mathbf{v} = 0,$$

which implies that

$$\|\mathbf{Z}_k\|_2 \leq \lambda_1 \omega_k \quad \text{and} \quad \|\mathbf{W}_k\|_2 \leq \lambda_2 u_k.$$

- If $\boldsymbol{\beta}_k \neq 0$ and $\boldsymbol{\phi}_k \neq 0$, then we obtain

$$-\mathbf{Z}_k + 2(1+c)\gamma_k \boldsymbol{\beta}_k + \lambda_1 \omega_k \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} = 0$$

and

$$-\mathbf{W}_k + 2c\gamma_k \boldsymbol{\phi}_k + \lambda_2 u_k \frac{\boldsymbol{\phi}_k}{\|\boldsymbol{\phi}_k\|_2} = 0.$$

Applying the l_2 norm to the last two equations, we get

$$\|\mathbf{Z}_k\|_2 = 2(1+c)\gamma_k \boldsymbol{\beta}_k + \lambda_1 \omega_k$$

and

$$\|\mathbf{W}_k\|_2 = 2c\gamma_k \boldsymbol{\phi}_k + \lambda_2 u_k,$$

which implies

$$\|\mathbf{Z}_k\|_2 \leq (1 + 2(1+c)\gamma_k)\lambda_1 \omega_k$$

and

$$\|\mathbf{W}_k\|_2 \leq (1 + 2c\gamma_k)\lambda_2 u_k.$$

Moreover, we have

$$-\mathbf{Z}_k + 2(1+c)\gamma_k \boldsymbol{\beta}_k + \lambda_1 w_k \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = 0 \left(\text{since } \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} \right)$$

and

$$-\mathbf{W}_k + 2c\gamma_k \boldsymbol{\phi}_k + \lambda_2 u_k \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_2} = 0 \left(\text{since } \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_2} = \frac{\boldsymbol{\phi}_k}{\|\boldsymbol{\phi}_k\|_2} \right),$$

which implies

$$\boldsymbol{\beta}_k = \frac{1}{2(1+c)\gamma_k} \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} (\|\mathbf{Z}_k\|_2 - \lambda_1 w_k),$$
$$\boldsymbol{\phi}_k = \frac{1}{2c\gamma_k} \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_2} (\|\mathbf{W}_k\|_2 - \lambda_2 u_k).$$

- If $\lambda_1 \leq \|\boldsymbol{\beta}_k\|_2 \leq \theta \lambda_1$ and $\lambda_2 \leq \|\boldsymbol{\phi}_k\|_2 \leq \theta \lambda_2$, then we get

$$-\mathbf{Z}_k + 2(1+c)\gamma_k\boldsymbol{\beta}_k + \frac{\theta\lambda_1 w_k}{\theta-1}\mathbf{u} - \frac{w_k}{\theta-1}\boldsymbol{\beta}_k = 0$$

and

$$-\mathbf{W}_k + 2c\gamma_k\boldsymbol{\phi}_k + \frac{\theta\lambda_2 u_k}{\theta-1}\mathbf{v} - \frac{u_k}{\theta-1}\boldsymbol{\phi}_k = 0,$$

where $\mathbf{u}$ and $\mathbf{v}$ are the sub-gradient of $P_{\lambda_1}(.)$ at $\boldsymbol{\beta}_k$ and $P_{\lambda_2}(.)$ at $\boldsymbol{\phi}_k$ repectively. So, we have $\|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1$.

- If $\boldsymbol{\beta}_k = 0$ and $\boldsymbol{\phi}_k = 0$, then we have

  $$-\mathbf{Z}_k + \frac{\theta\lambda_1 w_k}{\theta-1}\mathbf{u} = 0 \quad \text{and} \quad -\mathbf{W}_k + \frac{\theta\lambda_2 u_k}{\theta-1}\mathbf{v} = 0,$$

  which implies that

  $$\|\mathbf{Z}_k\|_2 \leq \frac{\theta\lambda_1 w_k}{\theta-1}\mathbf{u} \quad \text{and} \quad \|\mathbf{W}_k\|_2 \leq \frac{\theta\lambda_2 u_k}{\theta-1}\mathbf{v} = 0.$$

- If $\boldsymbol{\beta}_k \neq 0$ and $\boldsymbol{\phi}_k \neq 0$, then we obtain

  $$-\mathbf{Z}_k + 2(1+c)\gamma_k\boldsymbol{\beta}_k + \frac{\theta\lambda_1 w_k}{\theta-1}\frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} - \frac{w_k}{\theta-1}\boldsymbol{\beta}_k = 0$$

  and

  $$-\mathbf{W}_k + 2c\gamma_k\boldsymbol{\phi}_k + \frac{\theta\lambda_2 u_k}{\theta-1}\frac{\boldsymbol{\phi}_k}{\|\boldsymbol{\phi}_k\|_2} - \frac{u_k}{\theta-1}\boldsymbol{\phi}_k = 0,$$

  which implies that

$$\|\mathbf{Z}_k\|_2 = (2(1+c)\gamma_k - \frac{w_k}{\theta-1})\|\boldsymbol{\beta}_k\|_2 + \frac{w_k\lambda_1\theta}{\theta-1} \quad \left(\text{since} \quad \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2}\right) \tag{31}$$

  and

$$\|\mathbf{W}_k\|_2 = (2c\gamma_k - \frac{u_k}{\theta-1})\|\boldsymbol{\phi}_k\|_2 + \frac{u_k\lambda_2\theta}{\theta-1} \quad \left(\text{since} \quad \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_2} = \frac{\boldsymbol{\phi}_k}{\|\boldsymbol{\phi}_k\|_2}\right). \tag{32}$$

  For $\boldsymbol{\beta}_k$, combining the condition $\lambda_1 \leq \|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda_1$ and Eq. (31), we obtain

  $\lambda_1 w_k(\gamma_k 2(1+c)+1) \leq \|\mathbf{Z}_k\|_2 \leq 2(1+c)\gamma_k\theta\lambda w_k$.

If $\|\mathbf{Z}_k\|_2 \geq \dfrac{w_k\lambda_1\theta}{\theta-1}$, then we have

$$\boldsymbol{\beta}_k = \frac{1}{2(1+c)\gamma_k - \frac{w_k}{\theta-1}}\frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2}(\|\mathbf{Z}_k\|_2 - \lambda_1 w_k\frac{\theta}{\theta-1}).$$

For $\boldsymbol{\phi}_k$, combining the condition $\lambda_2 \leq \|\boldsymbol{\phi}_k\|_2 \leq \theta\lambda_2$ and Eq. (32), we obtain

$$\lambda_2 u_k(2c\gamma_k + 1) \leq \|\mathbf{W}_k\|_2 \leq 2c\gamma_k\theta\lambda_2 u_k.$$

If $\|\mathbf{W}_k\|_2 \geq \dfrac{u_k\lambda_2\theta}{\theta - 1}$, then we have

$$\boldsymbol{\phi}_k = \frac{1}{2c\gamma_k - \frac{u_k}{\theta-1}} \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_2}(\|\mathbf{W}_k\|_2 - \lambda_2 u_k \frac{\theta}{\theta - 1}).$$

- If $\|\boldsymbol{\beta}_k\|_2 \geq \theta\lambda_1$ and $\|\boldsymbol{\phi}_k\|_2 \geq \theta\lambda_2$, then we have

$$-\mathbf{Z}_k + 2(1 + c)\gamma_k\boldsymbol{\beta}_k = 0 \quad \text{and} \quad \mathbf{W}_k + 2c\gamma_k\boldsymbol{\beta}_k = 0.$$

This implies that

$$\|\mathbf{Z}_k\|_2 \geq 2(1 + c)\gamma_k\theta\lambda_1 w_k \quad \text{and} \quad \boldsymbol{\beta}_k = \frac{1}{2(1 + c)\gamma_k}\mathbf{Z}_k$$

and

$$\|\mathbf{W}_k\|_2 \geq 2c\gamma_k\theta\lambda_2 u_k \quad \text{and} \quad \boldsymbol{\phi}_k = \frac{1}{2c\gamma_k}\mathbf{W}_k.$$

For GMCP penalty

- If $\|\boldsymbol{\beta}_k\|_2 \leq \theta\lambda_1$ and $\|\boldsymbol{\phi}_k\|_2 \leq \theta\lambda_2$, then we get

$$-\mathbf{Z}_k + 2(1 + c)\gamma_k\boldsymbol{\beta}_k + \lambda_1 w_k\mathbf{u} - \frac{w_k}{\theta}\boldsymbol{\beta}_k = 0$$

and

$$-\mathbf{W}_k + 2c\gamma_k\boldsymbol{\phi}_k + \lambda_2 u_k\mathbf{v} - \frac{u_k}{\theta}\boldsymbol{\phi}_k = 0,$$

where $\mathbf{u}$ and $\mathbf{v}$ are the sub-gradient of $P_{\lambda_1}(.)$ at $\boldsymbol{\beta}_k$ and $P_{\lambda_2}(.)$ at $\boldsymbol{\phi}_k$ repectively. So, we have $\|\mathbf{u}\|_2 \leq 1, \|\mathbf{v}\|_2 \leq 1$.

- If $\boldsymbol{\beta}_k = 0$ and $\boldsymbol{\phi}_k = 0$, then we have

$$-\mathbf{Z}_k + \lambda_1 w_k\mathbf{u} = 0 \quad \text{and} \quad -\mathbf{W}_k + \lambda_2 u_k\mathbf{v} = 0,$$

which implies that

$$\|\mathbf{Z}_k\|_2 \leq \lambda_1 w_k\mathbf{u} \quad \text{and} \quad \|\mathbf{W}_k\|_2 \leq \lambda_2 u_k\mathbf{v} = 0.$$

- If $\boldsymbol{\beta}_k \neq 0$ and $\boldsymbol{\phi}_k \neq 0$, then

$$-\mathbf{Z}_k + 2(1 + c)\gamma_k\boldsymbol{\beta}_k + \lambda_1 w_k\frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} - \frac{w_k}{\theta}\boldsymbol{\beta}_k = 0$$

and

$$-\mathbf{W}_k + 2c\gamma_k\boldsymbol{\phi}_k + \lambda_2 u_k \frac{\boldsymbol{\phi}_k}{\|\boldsymbol{\phi}_k\|_2} - \frac{u_k}{\theta}\boldsymbol{\phi}_k = 0,$$

which implies that

$$\|\mathbf{Z}_k\|_2 = (2(1+c)\gamma_k - \frac{w_k}{\theta})\|\boldsymbol{\beta}_k\|_2 + w_k\lambda_1 \quad \left(\text{since} \quad \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2} = \frac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2}\right) \tag{33}$$

and

$$\|\mathbf{W}_k\|_2 = (2c\gamma_k - \frac{u_k}{\theta})\|\boldsymbol{\phi}_k\|_2 + u_k\lambda_2 \quad \left(\text{since} \quad \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_2} = \frac{\boldsymbol{\phi}_k}{\|\boldsymbol{\phi}_k\|_2}\right). \tag{34}$$

For $\boldsymbol{\beta}_k$, combining the condition $\|\boldsymbol{\beta}_k\|_2 \le \theta\lambda_1$ and Eq. (33), we obtain

$$\|\mathbf{Z}_k\|_2 \le 2(1+c)\gamma_k\theta\lambda w_k.$$

If $\|\mathbf{Z}_k\|_2 \ge w_k\lambda_1$, then we have

$$\boldsymbol{\beta}_k = \frac{1}{2(1+c)\gamma_k - \frac{w_k}{\theta}} \frac{\mathbf{Z}_k}{\|\mathbf{Z}_k\|_2}(\|\mathbf{Z}_k\|_2 - \lambda_1 w_k).$$

For $\boldsymbol{\phi}_k$, combining the condition $\|\boldsymbol{\phi}_k\|_2 \le \theta\lambda_2$ and Eq. (34), we obtain

$$\|\mathbf{W}_k\|_2 \le 2c\gamma_k\theta\lambda_2 u_k.$$

If $\|\mathbf{W}_k\|_2 \ge u_k\lambda_2$, then we have

$$\boldsymbol{\phi}_k = \frac{1}{2c\gamma_k - \frac{u_k}{\theta}} \frac{\mathbf{W}_k}{\|\mathbf{W}_k\|_2}(\|\mathbf{W}_k\|_2 - \lambda_2 u_k).$$

- If $\|\boldsymbol{\beta}_k\|_2 \ge \theta\lambda_1$ and $\|\boldsymbol{\phi}_k\|_2 \ge \theta\lambda_2$, then we have

  $$-\mathbf{Z}_k + 2(1+c)\gamma_k\boldsymbol{\beta}_k = 0 \quad \text{and} \quad -\mathbf{W}_k + 2c\gamma_k\boldsymbol{\beta}_k = 0,$$

  which implies

  $$\|\mathbf{Z}_k\|_2 \ge 2(1+c)\gamma_k\theta\lambda_1 w_k \quad \text{and} \quad \boldsymbol{\beta}_k = \frac{1}{2(1+c)\gamma_k}\mathbf{Z}_k$$

  and

  $$\|\mathbf{W}_k\|_2 \ge 2c\gamma_k\theta\lambda_2 u_k \quad \text{and} \quad \boldsymbol{\phi}_k = \frac{1}{2c\gamma_k}\mathbf{W}_k.$$

Appendix 4: Proof of Theorems 3, 4, 6, and 7

Let us state two Lemmas; the first lemma is on the properties of the expectile loss function $\Psi_\tau(.)$ and coupled loss function $S_\tau(.)$. The second lemma deals with sub-Gaussian random variables.

Lemma 1 (1) *For any* $\boldsymbol{\beta}, \boldsymbol{\delta} \in \mathbb{R}^p$, $2\underline{c}\|\mathbf{X}\boldsymbol{\delta}\|_2^2/n \leq \langle \nabla\Psi_\tau(\boldsymbol{\beta}+\boldsymbol{\delta}) - \nabla\Psi_\tau(\boldsymbol{\beta}), \boldsymbol{\delta}\rangle$. (2) *Let* $\boldsymbol{\epsilon} = (\epsilon_i, 1 \leq i \leq n)^\top$ *and* $\boldsymbol{\eta} = (\eta_i, 1 \leq i \leq n)^\top$, *where* $\eta_i = S'_\tau(\epsilon_i - e_\tau)$.

For $\boldsymbol{\theta}, \boldsymbol{\delta} \in \mathbb{R}^{2p}$, we have

$$n^{-1}c_0\|(\boldsymbol{I}_2 \otimes \mathbf{X})\boldsymbol{\delta}\|_2^2/n \leq \langle \nabla S_\tau(\boldsymbol{\theta}+\boldsymbol{\delta}) - \nabla S_\tau(\boldsymbol{\theta}), \boldsymbol{\delta}\rangle,$$

where $\boldsymbol{I}_2$ is a 2×2 is the identity matrix, and $c_0 = 2^{-1}[(1+\underline{c}) - (1+16\underline{c}^2)^{1/2}] > 0$.

Proof of Lemma 1 The parts (1) and (2) of Lemma 1 follow from the part (1) of Lemma 4 and the part (1) of Lemma 6 in Gu and Zou (2016), respectively.

Lemma 2 *Suppose that* $Z_1, \ldots, Z_n \in \mathbb{R}$ *are i.i.d sub-Gaussian random variables. Let* $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$, $K = \|\mathbf{Z}\|_{SG}$, $\mathbf{Z}^+ = \max(\mathbf{Z}, \mathbf{0})$ *and* $\mathbf{Z}^- = \min(-\mathbf{Z}, \mathbf{0})$.

(1) *If* $\mathbb{E}(\mathbf{Z}) = 0$, *then there exists an absolute constant* C *such that for any* $\boldsymbol{a} = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$ *and any* $t \geq 0$, *we have*

$$P(|\boldsymbol{a}^\top\mathbf{Z}| \geq t) \leq 2\exp\Big(-\frac{Ct^2}{K^2\|\boldsymbol{a}\|_2^2}\Big).$$

(2) *For any* $a_1, a_2 \in \mathbb{R}$, *the random variable* $a_1\mathbf{Z}^+ + a_2\mathbf{Z}^-$ *is sub-Gaussian*

(3) *Let* $\boldsymbol{A}$ *be a fixed* $m \times n$ *matrix. If* $\mathbb{E}(\mathbf{Z}) = 0$ *and* $var(\mathbf{Z}) = 1$, *then there exists an absolute constant* $C > 0$ *such that for any* $t \geq 0$

$$P\big(|\|\boldsymbol{A}\mathbf{Z}\|_2 - \|\boldsymbol{A}\|_F| \geq t\big) \leq 2exp\bigg(-\frac{Ct^2}{K^2\|\boldsymbol{A}\|_2^2}\bigg),$$

where $\|\boldsymbol{A}\|_F$ *and* $\|\boldsymbol{A}\|_2$ *represent the Frobenius and* l_2 *norms of matrix* $\boldsymbol{A}$, *respectively*.

Proof of Lemma 2 The part (1) follows from Proposition 5.10 of Vershynin (2010), and the parts (2) and (3) follow from the parts (4) and (2) of Lemma 3 of Gu and Zou (2016).

Proof of Theorem 3 Let $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ and $z_\infty^* = \|\nabla\Psi_\tau(\boldsymbol{\beta}^*)\|_{2,\infty}$, then $\hat{\boldsymbol{\beta}}$ satisfies the KKT conditions

$$\nabla_k\Psi_\tau(\boldsymbol{\beta}_k, \boldsymbol{\beta}_{-k}) + \mathbf{u}_k = 0, \quad \text{for } k = 1, \ldots, K,$$

where

$$\mathbf{u}_k = \begin{cases} \lambda^{\text{GLasso}} \dfrac{\boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_k\|_2} & \text{for } \boldsymbol{\beta}_k \neq 0, \\ \mathbf{u}_k, \quad \|\mathbf{u}_k\|_2 \in [-\lambda^{\text{GLasso}}, \lambda^{\text{GLasso}}], & \text{for } \boldsymbol{\beta}_k = 0. \end{cases}$$

It follows that

$$\langle \hat{\boldsymbol{\beta}}_k, \mathbf{u}_k \rangle = \lambda^{\text{GLasso}} \|\hat{\boldsymbol{\beta}}_k\|_2, \quad \forall k = 1, \ldots, K. \tag{35}$$

Lemma 1 and Holder's inequality lead to

$$\begin{aligned} 0 \leq 2\underline{c}\|X\hat{\boldsymbol{\delta}}\|_2^2/n &\leq \langle \nabla\Psi_\tau(\hat{\boldsymbol{\beta}}) - \nabla\Psi_\tau(\boldsymbol{\beta}^*), \hat{\boldsymbol{\delta}} \rangle \\ &= \sum_{k=1}^{K} \langle \nabla\Psi_\tau(\hat{\boldsymbol{\beta}}_k) - \nabla\Psi_\tau(\boldsymbol{\beta}_k^*), \hat{\boldsymbol{\delta}}_k \rangle \\ &= \sum_{k\in\mathcal{A}} \langle \nabla\Psi_\tau(\hat{\boldsymbol{\beta}}_k) - \nabla\Psi_\tau(\boldsymbol{\beta}_k^*), \hat{\boldsymbol{\delta}}_k \rangle + \sum_{k\in\mathcal{A}^c} \langle \nabla\Psi_\tau(\hat{\boldsymbol{\beta}}_k) - \nabla\Psi_\tau(\boldsymbol{\beta}_k^*), \hat{\boldsymbol{\delta}}_k \rangle \\ &\overset{(a)}{=} \sum_{k\in\mathcal{A}} \langle -\mathbf{u}_k - \nabla\Psi_\tau(\boldsymbol{\beta}_k^*), \hat{\boldsymbol{\delta}}_k \rangle + \sum_{k\in\mathcal{A}^c} \langle -\mathbf{u}_k, \hat{\boldsymbol{\beta}}_k \rangle + \sum_{k\in\mathcal{A}^c} \langle -\nabla\Psi_\tau(\boldsymbol{\beta}_k^*), \hat{\boldsymbol{\beta}}_k \rangle \\ &\leq \sum_{k\in\mathcal{A}} \| -\mathbf{u}_k - \nabla\Psi_\tau(\boldsymbol{\beta}_k^*)\|_2 \|\hat{\boldsymbol{\delta}}_k\|_2 - \sum_{k\in\mathcal{A}^c} \lambda^{\text{GLasso}} \|\hat{\boldsymbol{\beta}}_k\|_2 \\ &\quad + \sum_{k\in\mathcal{A}^c} \| -\nabla\Psi_\tau(\boldsymbol{\beta}_k^*)\|_2 \|\hat{\boldsymbol{\beta}}_k\|_2 \\ &\leq \sum_{k\in\mathcal{A}} \| -\mathbf{u}_k\|_2 \|\hat{\boldsymbol{\delta}}_k\|_2 + \sum_{k\in\mathcal{A}} \| -\nabla\Psi_\tau(\boldsymbol{\beta}_k^*)\|_2 \|\hat{\boldsymbol{\delta}}_k\|_2 - \lambda^{\text{GLasso}} \sum_{k\in\mathcal{A}^c} \|\hat{\boldsymbol{\beta}}_k\|_2 \\ &\quad + \sum_{k\in\mathcal{A}^c} \| -\nabla\Psi_\tau(\boldsymbol{\beta}_k^*)\|_2 \|\hat{\boldsymbol{\beta}}_k\|_2. \end{aligned}$$

Equality (a) is due to Eq. (35) and $\hat{\boldsymbol{\delta}}_k = \hat{\boldsymbol{\beta}}_k \quad \text{for } k \in \mathcal{A}^c$.

From the last inequality, we get

$$0 \leq 2\underline{c}\|X\hat{\boldsymbol{\delta}}\|_2^2/n \leq (z_\infty^* + \lambda^{\text{GLasso}})\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_{2,1} + (z_\infty^* - \lambda^{\text{GLasso}})\|\hat{\boldsymbol{\delta}}_{\mathcal{A}^c}\|_{2,1}. \tag{36}$$

Under the event $\xi_N = \{z_\infty^* \leq \lambda^{\text{GLasso}}/2\}$, we have

$$\|\hat{\boldsymbol{\delta}}_{\mathcal{A}^c}\|_{2,1} \leq \frac{z_\infty + \lambda^{\text{GLasso}}}{-z_\infty + \lambda^{\text{GLasso}}} \|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_{2,1} \leq 3\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_{2,1},$$

which implies that $\hat{\boldsymbol{\delta}} \in \mathcal{C}_3$ satisfies the condition (C3).

It follows that

$$2\underline{c}\kappa\|\hat{\boldsymbol{\delta}}\|_{2,1}^2 \leq 2\underline{c}\|X\hat{\boldsymbol{\delta}}\|_2^2/n \leq \frac{3}{2}\lambda^{\text{GLasso}}\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_{2,1} \leq \frac{3}{2}\lambda^{\text{GLasso}}\|\hat{\boldsymbol{\delta}}\|_{2,1}.$$

Thus, one has

$$\|\hat{\boldsymbol{\delta}}\|_{2,1} \leq 3\lambda^{\text{GLasso}}(4\kappa\underline{c})^{-1}.$$

Similarly, by condition (C4) and Eq. (36), we deduce

$$2n\underline{c}\varrho\|\hat{\boldsymbol{\delta}}\|_{2,1}\|\hat{\boldsymbol{\delta}}\|_{2,\infty} \leq 2\underline{c}\|\mathbf{X}\hat{\boldsymbol{\delta}}\|_2^2/n \leq \frac{3}{2}\lambda^{\text{GLasso}}\|\hat{\boldsymbol{\delta}}_{\mathcal{A}}\|_{2,1} \leq \frac{3}{2}\lambda^{\text{GLasso}}\|\hat{\boldsymbol{\delta}}\|_{2,1};$$

then

$$\|\hat{\boldsymbol{\delta}}\|_{2,\infty} \leq 3\lambda^{\text{GLasso}}(4\underline{c}\varrho)^{-1}.$$

Thus, we have

$$\begin{aligned} &P\Big(\big[\|\hat{\boldsymbol{\delta}}\|_{2,1} \leq 3\lambda^{\text{GLasso}}(4\kappa\underline{c})^{-1}\big] \cap \big[\|\hat{\boldsymbol{\delta}}\|_{2,\infty} \leq 3\lambda^{\text{GLasso}}(4\underline{c}\varrho)^{-1}\big]\Big) \\ &\geq P(z_{\infty}^* \leq \lambda^{\text{GLasso}}/2) \\ &\geq 1 - P\Big(\|\nabla\Psi_{\tau}(\boldsymbol{\beta}^*)\|_{2,\infty} \geq \lambda^{\text{GLasso}}/2\Big). \end{aligned} \tag{37}$$

Developing the last term of (37), we have

$$\begin{aligned} &P\Big(\|\nabla\Psi_{\tau}(\boldsymbol{\beta}^*)\|_{2,\infty} \geq \lambda^{\text{GLasso}}/2\Big) \\ &\leq \sum_{k=1}^{K} P\Big(\|\nabla\Psi_{\tau}(\boldsymbol{\beta}_k^*)\|_2 \geq \lambda^{\text{GLasso}}/2\Big) \\ &\leq \sum_{k=1}^{K} P\Big(\|\frac{\mathbf{x}_k^{\top}}{n}\mathbf{Z}\|_2 \geq \lambda^{\text{GLasso}}/2\Big) \\ &\leq \sum_{k=1}^{K} P\Big(\|\frac{\mathbf{x}_k^{\top}}{\sqrt{n}}\mathbf{Z}\|_{\infty} \geq \frac{\sqrt{n}\lambda^{\text{GLasso}}}{2\sqrt{p_k}}\Big) \\ &\leq \sum_{k=1}^{K} p_k \max_{1\leq j\leq p_k} P\Big(|\frac{\mathbf{x}_k^{\top}}{\sqrt{n}}\mathbf{Z}| \geq \frac{\sqrt{n}\lambda^{\text{GLasso}}}{2\sqrt{p_k}}\Big). \end{aligned}$$

Note that $Z_i = 2\tau\epsilon_i^+ - 2(1-\tau)\epsilon_i^-$, where $\epsilon^+ = \max(\epsilon, 0)$, $\epsilon^- = \max(-\epsilon, 0)$. It follows by part (2) of Lemma 2 and $\mathscr{E}^{\alpha}(\epsilon_i) = 0$ that Z_i are i.i.d sub-Gaussian random variables. Now by part (1) of Lemma 2 we have

$$\begin{aligned} P\Big(\|\nabla\Psi_{\tau}(\boldsymbol{\beta}_k)\|_{2,\infty} \geq \lambda^{\text{GLasso}}/2\Big) &\leq \sum_{k=1}^{K} 2p_k \ \exp\big(-\frac{Cn(\lambda^{\text{GLasso}})^2}{4K_0^2 M_0 p_k}\big) \\ &\leq 2p \ \exp\big(-\frac{Cn(\lambda^{\text{GLasso}})^2}{4K_0^2 M_0^2 \overline{p}_m}\big). \end{aligned} \tag{38}$$

From (37) and (38) we deduce

$$P\Big((\|\hat{\boldsymbol{\delta}}\|_{2,1} \le 3\lambda^{\mathrm{GLasso}}(4\kappa\underline{c})^{-1}) \cap (\|\hat{\boldsymbol{\delta}}\|_{2,\infty} \le 3\lambda^{\mathrm{GLasso}}(4\underline{c}\varrho)^{-1})\Big)$$
$$\ge 1 - 2p\ \exp\left(-\frac{Cn(\lambda^{\mathrm{GLasso}})^2}{4K_0^2 M_0^0 \overline{p}_m}\right).$$

This ends the proof of Theorem 3. □

Proof of Theorem 4 Let $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}^{\mathrm{GLasso}}$, under the condition $\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_{2,\infty} \le a_0\lambda$ and by assumptions (A1) and $a_0 = 1 \wedge a_2$, we have

For $k \in \mathcal{A}^c$

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_k^{(0)}\|_2 &\le \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_{2,\infty}\ \textit{for}\ k \in \mathcal{A}^c\ (\boldsymbol{\beta}_k^* = 0)\\ &\le a_0\lambda\\ &\le a_2\lambda.\end{aligned} \tag{39}$$

For $k \in \mathcal{A}$

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}_k^{(0)}\|_2 &\ge \min_{k\in\mathcal{A}}\|\boldsymbol{\beta}_k^*\|_2 - \|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}^*\|_{2,\infty}\\ &> (1+a)\lambda - a_0\lambda\\ &> a\lambda.\end{aligned}$$

By the last inequality and under property (P5), we have $P_\lambda(\|\hat{\boldsymbol{\beta}}_k^{(0)}\|_2) = 0$ for all $k \in \mathcal{A}$. Then, $\hat{\boldsymbol{\beta}}^{(1)}$ is solution to the following problem

$$\hat{\boldsymbol{\beta}}^{(1)} = \arg\min_{\boldsymbol{\beta}}\Big(\Psi_\tau(\boldsymbol{\beta}) + \sum_{k\in\mathcal{A}^c} P'_\lambda(\|\boldsymbol{\beta}_k^{(0)}\|_2)\|\boldsymbol{\beta}_k\|_2\Big). \tag{40}$$

By properties $(P3)$ and $(P4)$ and inequation (39), the inequality $P'_\lambda(\|\boldsymbol{\beta}_k^{(0)}\|_2) \ge a_1\lambda$ holds for $k \in \mathcal{A}^c$. Under the event $\{\|\nabla_k\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_2 < a_1\lambda, \forall k \in \mathcal{A}^c\}$, we demonstrate that $\hat{\boldsymbol{\beta}}^{oracle}$ is the unique global solution to (40). Indeed, from the convexity of Ψ_τ we obtain

$$\begin{aligned}\Psi_\tau(\boldsymbol{\beta}) &\ge \Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle}) + \sum_{k=1}^{K}\langle\nabla_k\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle}), \boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k^{oracle}\rangle;\\ &\overset{(a)}{=} \Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle}) + \sum_{k\in\mathcal{A}^c}\langle\nabla_k\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle}), \boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k^{oracle}\rangle.\end{aligned} \tag{41}$$

Equality (a) is due to $\nabla_k\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle}) = 0$ for all $k \in \mathcal{A}$ (KKT conditions of problem (10)). Using the inequality (41) leads to the following inequality

$$\begin{aligned}\left(\Psi_\tau(\boldsymbol{\beta})+\sum_{k\in\mathcal{A}^c}P'_\lambda(\|\boldsymbol{\beta}_k^{(0)}\|_2)\|\boldsymbol{\beta}_k\|_2\right)-\left(\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle})+\sum_{k\in\mathcal{A}^c}\langle\nabla_k\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle}),\hat{\boldsymbol{\beta}}_k^{oracle}\rangle\right)\\ \overset{(a)}{\geq}\sum_{k\in\mathcal{A}^c}\left(P'_\lambda(\|\boldsymbol{\beta}_k^{(0)}\|_2-\|\nabla_k\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_2\right)\|\boldsymbol{\beta}_k\|_2\\ \overset{(b)}{\geq}\sum_{k\in\mathcal{A}^c}\left(a_1\lambda-\|\nabla_k\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_2\right)\|\boldsymbol{\beta}_k\|_2\\ \geq 0.\end{aligned}$$

Inequalities (a) and (b) are due to the fact that $\langle\nabla_k\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle}),\hat{\boldsymbol{\beta}}_k\rangle\geq-\|\nabla_k\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_2\|\boldsymbol{\beta}_k\|_2$ and the condition (P4). Combining the last inequality with the uniqueness of the solution of problem (10), we conclude that $\hat{\boldsymbol{\beta}}^{oracle}$ is the unique solution to (40). Hence $\hat{\boldsymbol{\beta}}^{(1)}=\hat{\boldsymbol{\beta}}^{oracle}$. We start the second iteration of GLLA algorithm with the initial value $\hat{\boldsymbol{\beta}}^{oracle}$ solution of the problem (40) at the first iteration.

Let $\hat{\boldsymbol{\beta}}$ be the solution to the convex optimization problem in the second iteration of the GLLA algorithm. Under the event $\{\min_{k\in\mathcal{A}}\|\hat{\boldsymbol{\beta}}_k^{oracle}\|_2>a\lambda\}$, we have $P'_\lambda(\|\hat{\boldsymbol{\beta}}_k^{oracle}\|_2)=0,\ \forall k\in\mathcal{A}$ (condition (P5)). So, we obtain

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}}\left(\Psi_\tau(\boldsymbol{\beta})+\sum_{k\in\mathcal{A}^c}P'_\lambda(\hat{\boldsymbol{\beta}}_k^{oracle})\|\boldsymbol{\beta}_k\|_2\right).\tag{42}$$

Using $\hat{\boldsymbol{\beta}}_k^{oracle}=0,\ \forall k\in\mathcal{A}^c$ and condition $(P4)$, we have $P'_\lambda(\|\hat{\boldsymbol{\beta}}_k^{oracle}\|_2)=P'_\lambda(0)\geq a_1\lambda$. Hence, the problem (42) is very similar to (40). We deduce that $\hat{\boldsymbol{\beta}}^{oracle}$ is the unique solution to (42) under the event $\{\|\nabla_{\mathcal{A}}\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{2,\infty}<a_1\lambda\}$. Then, under the assumption of Theorem 4, the probability that Algorithm 2 initialized by $\hat{\boldsymbol{\beta}}^{\text{GLasso}}$ given by Theorem 3 converges to $\hat{\boldsymbol{\beta}}^{oracle}$ after two iterations is at least $1-p_1-p_2-p_3$, where

$$\begin{aligned}p_1&=P(\|\hat{\boldsymbol{\beta}}^{\text{GLasso}}-\boldsymbol{\beta}^*\|_{2,\infty}>a_0\lambda),\\ p_2&=P(\{\|\nabla_{\mathcal{A}^c}\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{2,\infty}\geq a_1\lambda\}),\\ p_3&=P(\{\min_{k\in\mathcal{A}}\|\hat{\boldsymbol{\beta}}_k^{oracle}\|_2\leq a\lambda\}).\end{aligned}$$

Let $\hat{\boldsymbol{\delta}}=\hat{\boldsymbol{\beta}}^{\text{GLasso}}-\boldsymbol{\beta}^*$. By the assumption $\lambda\geq3\lambda^{\text{GLasso}}a_0^{-1}\left((4\underline{c}\varrho)^{-1}\wedge(4\kappa\underline{c})^{-1}\right)$ and Theorem 3, we immediately get

$$\begin{aligned}p_1&\leq P(\|\hat{\boldsymbol{\delta}}\|_{2,\infty}>3\lambda^{\text{GLasso}}\left((4\underline{c}\varrho)^{-1}\wedge(4\kappa\underline{c})^{-1}\right))\\ &\leq P\left(\|\hat{\boldsymbol{\delta}}\|_{2,1}>3\lambda^{\text{GLasso}}(4\kappa\underline{c})^{-1}\right)\vee P\left(\|\hat{\boldsymbol{\delta}}\|_{2,\infty}>3\lambda^{\text{GLasso}}(4c_0\varrho)^{-1}\right)\\ &\leq p^*.\end{aligned}$$

To establishes the bound for p_2, we have

$$\begin{aligned} p_2 &= P(\{\|\nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{2,\infty} \geq a_1\lambda\}) \\ &\leq P(\{\|\nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*)\|_{2,\infty} \geq a_1\lambda/2\}) \\ &\quad + P(\{\|\nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*) - \nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{2,\infty} \geq a_1\lambda/2\}). \end{aligned} \tag{43}$$

By the same reasoning as in (38), we deduce

$$P(\{\|\nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*)\|_{2,\infty} \geq a_1\lambda/2\}) \leq 2(p - s_{\mathcal{A}}) \ \exp\big(-\frac{Cn\lambda^2 a_1^2}{4K_0^2 M_0^2 \overline{p}_m}\big). \tag{44}$$

Developing the second term of (43), we have

$$\begin{aligned} &P\Big(\{\|\nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*) - \nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{2,\infty} \geq a_1\lambda/2\}\Big) \\ &\qquad \leq P\Big(\{\max_{k\in\mathcal{A}^c} p_k^{1/2}\|\nabla_k\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*) - \nabla_k\boldsymbol{\Psi}_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_\infty \geq a_1\lambda/2\}\Big) \\ &\qquad \leq P\Big(\|\nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*) - \nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_\infty \geq \frac{a_1\lambda}{2\overline{p}_m^{1/2}}\Big). \end{aligned} \tag{45}$$

Let $\mathbf{d} = (d_i, i = 1\ldots, n)^\top$ with $d_i = \boldsymbol{\Psi}'_\tau(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}^{oracle}) - \boldsymbol{\Psi}'_\tau(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}^*)$. Using the Cauchy-Schwarz inequality and Lemma 2 of Gu and Zou (2016), we have

$$\begin{aligned} &\|\nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*) - \nabla_{\mathcal{A}^c}\boldsymbol{\Psi}_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_\infty \\ &\qquad = n^{-1}\max_{k\in\mathcal{A}^c}|\sum_i^n d_k x_{ik}| \\ &\qquad \leq n^{-1}\max_{k\in\mathcal{A}^c}\|\mathbf{d}\|_2\|X_k\|_2 \\ &\qquad \leq (2\overline{c}M_0)[(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oracle} - \boldsymbol{\beta}_{\mathcal{A}}^*)^\top(n^{-1}\mathbf{X}_{\mathcal{A}}^\top\mathbf{X}_{\mathcal{A}})(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oracle} - \boldsymbol{\beta}_{\mathcal{A}}^*)]^{1/2} \\ &\qquad \leq 2\overline{c}M_0\rho_{\max}^{1/2}\|\hat{\boldsymbol{\beta}}^{oracle} - \boldsymbol{\beta}^*\|_2. \end{aligned} \tag{46}$$

Combining (45) and (46), it follows from Lemma 3 and Lemma 4 of Gu and Zou (2016) that

$$\begin{aligned} P\Big(\{\|\nabla_{\mathcal{A}^c}\Psi_\tau(\boldsymbol{\beta}^*) - \nabla_{\mathcal{A}^c}\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{2,\infty} \geq a_1\lambda/2\}\Big) \\ \leq P\Big(\|\hat{\boldsymbol{\beta}}^{oracle} - \boldsymbol{\beta}^*\|_2 \geq \frac{a_1\lambda}{4\overline{c}M_0\rho_{\max}^{1/2}\overline{p}_m^{1/2}}\Big) \\ \overset{(a)}{\leq} P\Big(\|n^{-1}\mathbf{X}_{\mathcal{A}}^\top\xi\|_2 \geq \frac{a_1\underline{c}\rho_{\min}}{2\overline{c}M_0\rho_{\max}^{1/2}\overline{p}_m^{1/2}}\lambda\Big) \\ = P\Big(\|n^{-1}\mathbf{X}_{\mathcal{A}}^\top\xi\|_2 \geq Q_1\lambda\Big) \\ \overset{(b)}{\leq} \Gamma(Q_1\lambda, n, s_{\mathcal{A}}, K_0, M_0, \rho_{\max}, \nu_0). \end{aligned} \tag{47}$$

The inequalities (a) and (b) are due to the Lemmas 4(2) and 3(3) of Gu and Zou (2016) respectively.

Combining (43), (44) and (47), we immediately get the upper bound for p_2

$$p_2 = 2(p - s_{\mathcal{A}})\ \exp\Big(-\frac{Cn\lambda^2a_1^2}{4K_0^2M_0^2\overline{p}_m}\Big) + \Gamma(Q_1\lambda, n, s_{\mathcal{A}}, K_0, M_0, \rho_{\max}, \nu_0). \tag{48}$$

To derive the upper bound for p_3, let $R = \min_{k\in\mathcal{A}}\|\boldsymbol{\beta}_k^*\|_2 - a\lambda > 0$. Then, we have

$$\begin{aligned} P(\mathcal{E}_3^c) &\leq P(\{\min_{k\in\mathcal{A}}\|\hat{\boldsymbol{\beta}}_k^{oracle}\|_2 \leq a\lambda\}) \\ &\leq P(\max_{k\in\mathcal{A}}\|\hat{\boldsymbol{\beta}}_k^{oracle} - \boldsymbol{\beta}_k^*\|_2 > R) \\ &\leq P(\max_{k\in\mathcal{A}}p_k^{1/2}\|\hat{\boldsymbol{\beta}}_k^{oracle} - \boldsymbol{\beta}_k^*\|_\infty > R) \\ &\leq P(\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oracle} - \boldsymbol{\beta}_{\mathcal{A}}^*\|_\infty > \frac{R}{\overline{p}_{\mathcal{A}}}) \\ &\leq P\Big(\|n^{-1}\mathbf{X}_{\mathcal{A}}^\top\xi\|_2 \geq 2\underline{c}\rho_{\min}R\overline{p}_{\mathcal{A}}^{-1}\Big) \\ &\overset{(a)}{\leq} \Gamma(2\underline{c}\rho_{\min}R\overline{p}_{\mathcal{A}}^{-1}, n, s_{\mathcal{A}}, K_0, M_0, \rho_{\max}, \nu_0), \end{aligned} \tag{49}$$

where the inequality (a) is due to Lemma 3(3) of Gu and Zou (2016).

This ends the proof of Theorem 4. □

Proof of Theorem 6 Let $\boldsymbol{\delta}_1 = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, $\boldsymbol{\delta}_2 = \hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*$, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\phi}}^\top)^\top$, $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\delta}}_2^\top, \hat{\boldsymbol{\delta}}_2^\top)^\top$, $z_{1\infty}^* = \|\partial S_\tau(\boldsymbol{\theta}^*)/\partial \boldsymbol{\beta}_k\|_{2,\infty}$ and $z_{2\infty}^* = \|\partial S_\tau(\boldsymbol{\theta}^*)/\partial \boldsymbol{\phi}_k\|_{2,\infty}$. By Lemma 1 and similar arguments in the proof of Theorem 3, we get

$$
\begin{aligned}
0 &\le n^{-1} c_0 \|(\boldsymbol{I}_2 \otimes \mathbf{X})\hat{\boldsymbol{\delta}}\|_2^2 / n \\
&\le \langle \nabla S_\tau(\hat{\boldsymbol{\theta}}) - \nabla S_\tau(\boldsymbol{\theta}^*), \hat{\boldsymbol{\delta}} \rangle \\
&= \sum_{k=1}^{K} \langle \nabla S_\tau(\hat{\boldsymbol{\theta}}_k) - \nabla S_\tau(\boldsymbol{\theta}_k^*), \hat{\boldsymbol{\delta}}_k \rangle \\
&= \sum_{k \in \mathcal{A}_1} \langle \nabla S_\tau(\hat{\boldsymbol{\theta}}_k) - \nabla S_\tau(\boldsymbol{\theta}_k^*), \hat{\boldsymbol{\delta}}_k \rangle + \sum_{k \in \mathcal{A}_1^c} \langle \nabla S_\tau(\hat{\boldsymbol{\theta}}_k) - \nabla S_\tau(\boldsymbol{\theta}_k^*), \hat{\boldsymbol{\delta}}_k \rangle \\
&\quad + \sum_{k \in \mathcal{A}_2} \langle \nabla S_\tau(\hat{\boldsymbol{\theta}}_k) - \nabla S_\tau(\boldsymbol{\theta}_k^*), \hat{\boldsymbol{\delta}}_k \rangle + \sum_{k \in \mathcal{A}_2^c} \langle \nabla S_\tau(\hat{\boldsymbol{\theta}}_k) - \nabla S_\tau(\boldsymbol{\theta}_k^*), \hat{\boldsymbol{\delta}}_k \rangle \\
&\le (z_{1\infty}^* + \lambda_1^{\mathrm{GLasso}}) \|(\hat{\boldsymbol{\delta}}_1)_{\mathcal{A}_1}\|_{2,1} + (z_{1\infty}^* - \lambda_1^{\mathrm{GLasso}}) \|(\hat{\boldsymbol{\delta}}_1)_{\mathcal{A}_1^c}\|_{2,1} \\
&\quad + (z_{2\infty}^* + \lambda_2^{\mathrm{GLasso}}) \|(\hat{\boldsymbol{\delta}}_2)_{\mathcal{A}_2}\|_{2,1} + (z_{2\infty}^* - \lambda_2^{\mathrm{GLasso}}) \|(\hat{\boldsymbol{\delta}}_2)_{\mathcal{A}_2^c}\|_{2,1}.
\end{aligned} \tag{50}
$$

Under the event $\xi_1 = \{z_{1\infty}^* \le \lambda_1^{\mathrm{GLasso}}/2\}$ and $\xi_2 = \{z_{2\infty}^* \le \lambda_2^{\mathrm{GLasso}}/2\}$, it follows from the later inequality that

$$
\begin{aligned}
&(-z_{1\infty}^* + \lambda_1^{\mathrm{GLasso}}) \|(\hat{\boldsymbol{\delta}}_1)_{\mathcal{A}_1^c}\|_{2,1} + (-z_{2\infty}^* + \lambda_2^{\mathrm{GLasso}}) \|(\hat{\boldsymbol{\delta}}_2)_{\mathcal{A}_2^c}\|_{2,1} \\
&\quad \le (z_{1\infty}^* + \lambda_1^{\mathrm{GLasso}}) \|(\hat{\boldsymbol{\delta}}_1)_{\mathcal{A}_1}\|_{2,1} \\
&\qquad + (z_{2\infty}^* + \lambda_2^{\mathrm{GLasso}}) \|(\hat{\boldsymbol{\delta}}_2)_{\mathcal{A}_2}\|_{2,1},
\end{aligned}
$$

which implies that

$$
\begin{aligned}
2^{-1}\lambda_1^{\mathrm{GLasso}} \|(\hat{\boldsymbol{\delta}}_1)_{\mathcal{A}_1^c}\|_{2,1} + 2^{-1}\lambda_2^{\mathrm{GLasso}} \|(\hat{\boldsymbol{\delta}}_2)_{\mathcal{A}_2^c}\|_{2,1} &\le (3/2)\lambda_1^{\mathrm{GLasso}} \|(\hat{\boldsymbol{\delta}}_1)_{\mathcal{A}_1}\|_{2,1} \\
&\quad + (3/2)\lambda_2^{\mathrm{GLasso}} \|(\hat{\boldsymbol{\delta}}_2)_{\mathcal{A}_2}\|_{2,1}.
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
2^{-1}\underline{\lambda}^{\mathrm{GLasso}} \|\hat{\boldsymbol{\delta}}_{\mathcal{A}_0^c}\|_{2,1} &\le 2^{-1}\lambda_1^{\mathrm{GLasso}} \|(\hat{\boldsymbol{\delta}}_1)_{\mathcal{A}_1^c}\|_{2,1} + 2^{-1}\lambda_2^{\mathrm{GLasso}} \|(\hat{\boldsymbol{\delta}}_2)_{\mathcal{A}_2^c}\|_{2,1} \\
&\le (3/2)\lambda_1^{\mathrm{GLasso}} \|(\hat{\boldsymbol{\delta}}_1)_{\mathcal{A}_1}\|_{2,1} + (3/2)\lambda_2^{\mathrm{GLasso}} \|(\hat{\boldsymbol{\delta}}_2)_{\mathcal{A}_2}\|_{2,1} \\
&\le (3/2)\overline{\lambda}^{\mathrm{GLasso}} \|\hat{\boldsymbol{\delta}}_{\mathcal{A}_0}\|_{2,1}.
\end{aligned}
$$

Then, we have $\hat{\boldsymbol{\delta}} \in \boldsymbol{\xi}_{3\tilde{N}}$. Now under conditions $(C3)' - (C4)'$ we have from (50)

$$
\begin{aligned}
c_0 \overline{\kappa} \|\hat{\boldsymbol{\delta}}\|_{2,1}^2 &\le n^{-1} c_0 \|(\boldsymbol{I}_2 \otimes \mathbf{X})\hat{\boldsymbol{\delta}}\|_2^2 \le (3/2)\overline{\lambda}^{\mathrm{GLasso}} \|\hat{\boldsymbol{\delta}}_{\mathcal{A}_0}\|_{2,1} \\
&\le (3/2)\overline{\lambda}^{\mathrm{GLasso}} \|\hat{\boldsymbol{\delta}}\|_{2,1};
\end{aligned}
$$

then, one has

$$
\|\hat{\boldsymbol{\delta}}\|_{2,1} \le (3/2)\overline{\lambda}^{\mathrm{GLasso}} (c_0 \overline{\kappa})^{-1}.
$$

Similarly, by condition $(C5)'$

$$c_0\varrho\|\hat{\boldsymbol{\delta}}_{\mathcal{A}_0}\|_{2,1}\|\hat{\boldsymbol{\delta}}\|_{2,\infty} \leq n^{-1}c_0\|(\boldsymbol{I}_2 \otimes \mathbf{X})\hat{\boldsymbol{\delta}}\|_2^2 \leq (3/2)\overline{\lambda}^{\text{GLasso}}\|\hat{\boldsymbol{\delta}}_{\mathcal{A}_0}\|_{2,1};$$

thus,

$$\|\hat{\boldsymbol{\delta}}\|_{2,\infty} \leq (3/2)\overline{\lambda}^{\text{GLasso}}(c_0\varrho)^{-1}.$$

It follows that under event ξ_1 and ξ_2, we have $\|\hat{\boldsymbol{\delta}}\|_{2,\infty} \leq (3/2)(c_0\varrho)^{-1}\overline{\lambda}^{\text{GLasso}}$ and $\|\hat{\boldsymbol{\delta}}\|_{2,1} \leq (3/2)(c_0\kappa)^{-1}\overline{\lambda}^{\text{GLasso}}$. By Lemma 6 of Gu and Zou (2016), ϵ_i and $\eta = S'_\tau(\epsilon_i - e_\tau)$ are both mean zero sub-Gaussian random variables with $K_1 = \|\epsilon_i\|_{SG}$ and $K_2 = \|\eta_i\|_{SG}$. It follows that $\epsilon_i + \eta_i$ is also sub-Gaussian and we have $\|\epsilon_i + \eta_i\|_{SG} \leq K_1 + K_2$. Since $M_1 = \|\mathbf{X}\boldsymbol{\theta}^*\|_\infty$, we get

$$\begin{aligned}
&P\left(\left[\|\hat{\boldsymbol{\delta}}\|_{2,1} \leq (3/2)(c_0\kappa)^{-1}\overline{\lambda}^{\text{GLasso}}\right] \cap \left[\|\hat{\boldsymbol{\delta}}\|_{2,\infty} \leq (3/2)(c_0\varrho)^{-1}\overline{\lambda}^{\text{GLasso}}\right]\right) \\
&\quad \geq P(\xi_1 \cap \xi_2) \\
&\quad \geq 1 - P(\xi_1^c) - P(\xi_2^c) \\
&\quad = 1 - P\left(\|n^{-1}\mathbf{X}^\top \boldsymbol{W}(\boldsymbol{\epsilon} + \boldsymbol{\eta})\|_{2,\infty} \geq \lambda_1^{\text{GLasso}}/2\right) \\
&\qquad - P\left(\|n^{-1}\mathbf{X}^\top \boldsymbol{W}\boldsymbol{\eta}\|_{2,\infty} \geq \lambda_2^{\text{GLasso}}/2\right) \\
&\quad \geq 1 - \sum_{k=1}^{K} P\left(\|n^{-1}\mathbf{x}_k^\top \boldsymbol{W}(\boldsymbol{\epsilon} + \boldsymbol{\eta})\|_2 \geq \lambda_1^{\text{GLasso}}/2\right) \\
&- \sum_{k=1}^{K} P\left(\|n^{-1}\mathbf{x}_k^\top \boldsymbol{W}\boldsymbol{\eta}\| \geq \lambda_2^{\text{GLasso}}/2\right) \\
&\quad \geq 1 - \sum_{k=1}^{K} P\left(\|n^{-1}\mathbf{x}_k^\top \boldsymbol{W}(\boldsymbol{\epsilon} + \boldsymbol{\eta})\|_\infty \geq \lambda_1^{\text{GLasso}}/(2\sqrt{p_k})\right) \\
&\qquad - \sum_{k=1}^{K} P\left(\|n^{-1}\mathbf{x}_k^\top \boldsymbol{W}\boldsymbol{\eta}\|_\infty \geq \lambda_2^{\text{GLasso}}/(2\sqrt{p_k})\right) \\
&\quad \geq 1 - \sum_{k=1}^{K} 2p_k \ \exp\left(-\frac{Cn(\lambda_1^{\text{GLasso}})^2}{4(K_1+K_2)^2 M_0^2 M_1^2 p_k}\right) - \\
&\qquad\qquad \sum_{k=1}^{K} 2p_k \ \exp\left(-\frac{Cn(\lambda_2^{\text{GLasso}})^2}{4K_2^2 M_0^2 M_1^2 p_k}\right) \\
&\quad \geq 1 - 2p \ \exp\left(-\frac{Cn(\lambda_1^{\text{GLasso}})^2}{4(K_1+K_2)^2 M_0^2 M_1^2 \overline{p}_m}\right) - 2p \ \exp\left(-\frac{Cn(\lambda_2^{\text{GLasso}})^2}{4K_2^2 M_0^2 M_1^2 \overline{p}_m}\right).
\end{aligned}$$

This ends the proof of Theorem 6. □

Proof of Theorem 7 From Lemma 7 of Gu and Zou (2016), the restriction of $S_\tau(\boldsymbol{\beta}, \boldsymbol{\phi})$ to the set $\mathcal{S} = \{\boldsymbol{\beta}, \boldsymbol{\phi} \in \mathbb{R}^{2p} : \boldsymbol{\beta}_{\mathcal{A}_1^c} = \mathbf{0}, \boldsymbol{\phi}_{\mathcal{A}_2^c} = \mathbf{0}\}$ is strongly convex. Hence, the oracle estimators $(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})$ are the unique solutions of problem (22).

Under the event

$$\mathcal{E}_1 = \{\|\hat{\boldsymbol{\beta}}^{\text{GLasso}} - \boldsymbol{\beta}^*\|_{2,\infty} \le a_0\lambda_1; \|\hat{\boldsymbol{\phi}}^{\text{GLasso}} - \boldsymbol{\phi}^*\|_{2,\infty} \le a_0\lambda_2\}$$

and assumption (A2), we have

$$\min_{k\in\mathcal{A}_1} \|\hat{\boldsymbol{\beta}}_k^{\text{GLasso}}\|_2 \ge \min_{k\in\mathcal{A}_1} \|\boldsymbol{\beta}_k^*\|_2 - \|\boldsymbol{\beta}^{\text{GLasso}} - \boldsymbol{\beta}^*\|_{2,\infty} > a\lambda_1,$$

which implies that $P'_\lambda(\|\hat{\boldsymbol{\beta}}_k^{\text{GLasso}}\|_2) = 0$ for $k \in \mathcal{A}_1$.

We have also

$$\|\hat{\boldsymbol{\beta}}_{\mathcal{A}_1^c}^{\text{GLasso}}\|_{2,\infty} \le \|\boldsymbol{\beta}^{\text{GLasso}} - \boldsymbol{\beta}^*\|_{2,\infty} \le a_2\lambda_1,$$

implying that

$$P'_\lambda(\|\hat{\boldsymbol{\beta}}_k^{\text{GLasso}}\|_2) \ge a_1\lambda_1 \text{ for } k \in \mathcal{A}_1^c.$$

A similar argument is used to show that $P'_\lambda(\|\hat{\boldsymbol{\phi}}_k^{\text{GLasso}}\|_2) = 0$ for $k \in \mathcal{A}_2$ and $P'_\lambda(\|\hat{\boldsymbol{\phi}}_k^{\text{GLasso}}\|_2) \ge a_1\lambda_2$ for $k \in \mathcal{A}_2^c$.

Now, let $\hat{\boldsymbol{\beta}}^1$ and $\hat{\boldsymbol{\phi}}^1$ be the update after the first iteration of the GLLA algorithm, then under $\mathcal{E}_1$, $(\hat{\boldsymbol{\beta}}^1\ \hat{\boldsymbol{\phi}}^1)$ is minimizers of

$$L_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}) := S_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}) + \sum_{k\in\mathcal{A}_1^c} P'_\lambda(\|\hat{\boldsymbol{\beta}}_k^{\text{GLasso}}\|_2)\|\boldsymbol{\beta}_k\|_2 + \sum_{k\in\mathcal{A}_2^c} P'_\lambda(\|\hat{\boldsymbol{\phi}}_k^{\text{GLasso}}\|_2)\|\boldsymbol{\phi}_k\|_2. \tag{51}$$

By definition of the oracle estimators, $\partial S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})/\partial\boldsymbol{\beta}_k = 0$ for $k \in \mathcal{A}_1$ and $\partial S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})/\partial\boldsymbol{\phi}_k = 0$ for $k \in \mathcal{A}_2$. Also $\hat{\boldsymbol{\beta}}_k^{oracle} = 0$ for $k \in \mathcal{A}_1^c$ and $\hat{\boldsymbol{\phi}}_k^{oracle} = 0$ for $k \in \mathcal{A}_2^c$.

It follows from convexity of $S_\tau(\boldsymbol{\beta}, \boldsymbol{\phi})$ that

$$\begin{aligned} S_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}) &\ge S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle}) + \sum_{k=1}^{K} \langle \nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle}), \boldsymbol{\delta}_k - \hat{\boldsymbol{\delta}}_k^{oracle} \rangle; \\ &= S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle}) + \sum_{k\in\mathcal{A}_1^c} \langle \nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle}), \boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k^{oracle} \rangle \\ &\quad + \sum_{k\in\mathcal{A}_2^c} \langle \nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle}), \boldsymbol{\phi}_k - \hat{\boldsymbol{\phi}}_k^{oracle} \rangle \end{aligned} \tag{52}$$

Combining (51) and (52), we have

$$
\begin{aligned}
& L_\tau(\boldsymbol{\beta}, \boldsymbol{\phi}) - L_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle}) \\
& \overset{(a)}{\geq} \sum_{k \in \mathcal{A}_1^c} \left(P'_{\lambda_1}(\|\boldsymbol{\beta}_k^{(0)}\|_2 - \|\nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_2 \right) \|\boldsymbol{\beta}_k\|_2 \\
& \quad + \sum_{k \in \mathcal{A}_2^c} \left(P'_{\lambda}(\|\boldsymbol{\phi}_k^{(0)}\|_2 - \|\nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_2 \right) \|\boldsymbol{\phi}_k\|_2 \\
& \geq \sum_{k \in \mathcal{A}_1^c} \left(a_1\lambda_1 - \|\nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_2 \right) \|\boldsymbol{\beta}_k\|_2 \\
& \quad + \sum_{k \in \mathcal{A}_2^c} \left(a_1\lambda_2 - \|\nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_2 \right) \|\boldsymbol{\phi}_k\|_2 \\
& \overset{(b)}{\geq} 0.
\end{aligned}
$$

Inequality (a) is due to

$$\langle \nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle}), \hat{\boldsymbol{\beta}}_k^{oracle} \rangle \geq -\|\nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_2 \|\boldsymbol{\beta}_k\|_2$$

and

$$\langle \nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle}), \hat{\boldsymbol{\phi}}_k^{oracle} \rangle \geq -\|\nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_2 \|\boldsymbol{\phi}_k\|_2.$$

Inequality (b) is true under the conditions $\mathcal{E}_2 = \{\|\nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_2 < a_1\lambda_1, \forall k \in \mathcal{A}_1^c; \|\nabla_k S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_2 < a_1\lambda_2, \forall k \in \mathcal{A}_2^c\}$.

Combining the last inequality with the uniqueness of the solution of problem (22), we conclude that $(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})$ is the unique solution to (25).

Hence $(\hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\phi}}^{(1)}) = (\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})$. We start the second iteration of GLLA algorithm with the initial value $(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})$ solution of the problem (25) at the first iteration. Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}})$ be the solution to the convex optimization problem in the second iteration of the GLLA algorithm. Under the condition

$$\mathcal{E}_2 = \{\min_{k \in \mathcal{A}_1} \|\hat{\boldsymbol{\beta}}_k^{oracle}\|_2 > a\lambda_1, \min_{k \in \mathcal{A}_2} \|\hat{\boldsymbol{\phi}}_k^{oracle}\|_2 > a\lambda_2\},$$

we obtain

$$P'_{\lambda_1}(\|\hat{\boldsymbol{\beta}}_k^{oracle}\|_2) = 0, \ \forall k \in \mathcal{A}_1 \text{ and } P'_{\lambda_2}(\|\hat{\boldsymbol{\phi}}_k^{oracle}\|_2) = 0, \ \forall k \in \mathcal{A}_2.$$

We have also

$$(\hat{\boldsymbol{\beta}}_k^{oracle}, \hat{\boldsymbol{\phi}}_{k'}^{oracle}) = \mathbf{0}, \ \forall (k, k') \in \mathcal{A}_1^c \times \mathcal{A}_2^c.$$

Then, by property (*P*5), we have

$$P'_{\lambda_1}(\hat{\boldsymbol{\beta}}_k^{oracle}) = P'_{\lambda_1}(0) \geq a_1\lambda_1 \text{ and } P'_{\lambda_2}(\hat{\boldsymbol{\phi}}_k^{oracle}) = P'_{\lambda_2}(0) \geq a_1\lambda_2.$$

Hence, optimization problem in the second iteration becomes

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta},\boldsymbol{\phi}} \left(S_\tau(\boldsymbol{\beta},\boldsymbol{\phi}) + \sum_{k\in\mathcal{A}_1^c} P'_{\lambda_1}(\hat{\boldsymbol{\beta}}_k^{oracle})\|\boldsymbol{\beta}_k\|_2 + \sum_{k\in\mathcal{A}_2^c} P'_{\lambda_2}(\hat{\boldsymbol{\phi}}_k^{oracle})\|\boldsymbol{\phi}_k\|_2 \right). \tag{53}$$

The problem (53) is very similar to (51), thus, we deduce that $(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})$ is the unique solution to (53) under the event

$$\mathcal{E}_3 = \{\|\nabla_{\mathcal{A}_1} S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_{2,\infty} < a_1\lambda_1; \|\nabla_{\mathcal{A}_2} S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_{2,\infty} < a_1\lambda_2\}.$$

Then, under the assumption of Theorem 7, the probability that Algorithm 4 initialized by $(\hat{\boldsymbol{\beta}}^{\text{GLasso}}, \hat{\boldsymbol{\phi}}^{\text{GLasso}})$ given by Theorem 6 converges to $(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})$ after two iterations is at least $1 - P(\mathcal{E}_1^c) - P(\mathcal{E}_2^c) - P(\mathcal{E}_3^c)$.

By the assumption of Theorem 7, we immediately get

$$\begin{aligned}
P(\mathcal{E}_1^c) &\leq P(\|\hat{\boldsymbol{\delta}}\|_{2,\infty} > a_0\lambda) \\
&\leq P\left(\|\hat{\boldsymbol{\delta}}\|_{2,\infty} > (3/2)\overline{\lambda}^{\text{GLasso}} \left((c_0\overline{\kappa})^{-1} \wedge (c_0\overline{\varrho})^{-1} \right) \right) \\
&\leq P\left(\|\hat{\boldsymbol{\delta}}\|_{2,1} > (3/2)\overline{\lambda}^{\text{GLasso}} (c_0\overline{\kappa})^{-1} \right) \\
&\vee P\left(\|\hat{\boldsymbol{\delta}}\|_{2,\infty} > (3/2)\overline{\lambda}^{\text{GLasso}} (c_0\overline{\varrho})^{-1} \right) \\
&\leq \pi_1.
\end{aligned} \tag{54}$$

To establish the bound for $P(\mathcal{E}_2^c)$, we have

$$\begin{aligned}
P(\mathcal{E}_2^c) &\leq P(\{\|\nabla_{\mathcal{A}_1^c} S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_{2,\infty} \geq a_1\lambda_1\} \cup \\
&\qquad \{\|\nabla_{\mathcal{A}_2^c} S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_{2,\infty} \geq a_1\lambda_2\}) \\
&\leq P(\{\|\nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c} S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle})\|_{2,\infty} \geq a_1\lambda\}) \\
&\leq P(\{\|\nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c} S_\tau(\boldsymbol{\beta}^*, \boldsymbol{\phi}^*)\|_{2,\infty} \geq a_1\lambda/2\}) \\
&\quad + P(\{\|\nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c} S_\tau(\hat{\boldsymbol{\beta}}^{oracle}, \hat{\boldsymbol{\phi}}^{oracle}) \\
&\quad - \nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c} S_\tau(\boldsymbol{\beta}^*, \boldsymbol{\phi}^*)\|_{2,\infty} \geq a_1\lambda/2\}).
\end{aligned} \tag{55}$$

Using (38), we deduce that

$$
\begin{aligned}
&P(\{\|\nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c} S_\tau(\boldsymbol{\beta}^*, \boldsymbol{\phi}^*)\|_{2,\infty} \geq a_1\lambda/2\}) \\
&\quad \leq P(\{\|n^{-1}\mathbf{X}^\top_{\mathcal{A}_1^c} W(\boldsymbol{\epsilon}+\boldsymbol{\eta})\|_{2,\infty} \geq a_1\lambda/2\}) \\
&+ P(\{\|n^{-1}\mathbf{X}^\top_{\mathcal{A}_2^c} W\boldsymbol{\eta}\|_{2,\infty} \geq a_1\lambda/2\}) \\
&\quad \leq P(\{\|n^{-1}\mathbf{X}^\top_{\mathcal{A}_1^c} W(\boldsymbol{\epsilon}+\boldsymbol{\eta})\|_{\infty} \geq a_1\lambda/(2\overline{p}_{\mathcal{A}_1^c})\}) \\
&\qquad + P(\{\|n^{-1}\mathbf{X}^\top_{\mathcal{A}_2^c} W\boldsymbol{\eta}\|_{\infty} \geq a_1\lambda/(2\overline{p}_{\mathcal{A}_2^c})\}) \\
&\quad \leq 2(p-s_{\mathcal{A}_1})\ \exp\left(-\frac{Cn\lambda^2 a_1^2}{4M_0^2M_1^2(K_1+K_2)^2\overline{p}^2_{\mathcal{A}_1^c}}\right) \\
&\qquad + 2(p-s_{\mathcal{A}_2})\ \exp\left(-\frac{Cn\lambda^2 a_1^2}{4M_0^2M_1^2K_2^2\overline{p}^2_{\mathcal{A}_2^c}}\right),
\end{aligned}
\tag{56}
$$

where $s_{\mathcal{A}_1} = \sum_{k\in\mathcal{A}_1} p_k$ and $s_{\mathcal{A}_2} = \sum_{k\in\mathcal{A}_2} p_k$.

Let $\mathbf{d} = (d_i, i = 1\ldots, n)^\top$ with $d_i = \rho'_\tau(y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}^{oracle} - \mathbf{x}_i^\top\hat{\boldsymbol{\phi}}^{oracle}) - \rho'_\tau(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}^* - \mathbf{x}_i^\top\boldsymbol{\phi}^*)$. It follows that

$$
\begin{aligned}
&P\left(\{\|\nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c}\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*) - \nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c}\boldsymbol{\Psi}_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{2,\infty} \geq a_1\lambda/2\}\right) \\
&\quad \leq P\left(\{\max_{k\in(\mathcal{A}_1\cup\mathcal{A}_2)^c} p_k\|\nabla_k\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*) - \nabla_k\boldsymbol{\Psi}_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{\infty} \geq a_1\lambda/2\}\right) \\
&\quad \leq P\left(\|\nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c}\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*) - \nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c}\boldsymbol{\Psi}_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{\infty} \geq \frac{a_1\lambda}{2\overline{p}_{(\mathcal{A}_1\cup\mathcal{A}_2)^c}}\right).
\end{aligned}
\tag{57}
$$

We have then

$$
\begin{aligned}
&\|\nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c}\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}^*) - \nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c}\boldsymbol{\Psi}_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{\infty} \\
&\qquad \leq M_0(\|\mathbf{X}(\hat{\boldsymbol{\beta}}^{oracle} - \boldsymbol{\beta}^*)\|_2 + \|\mathbf{b}\|_2)/\sqrt{n} \\
&\qquad \leq M_0\Big[(1+2\overline{c})\|\mathbf{X}_{\mathcal{A}_1}(\hat{\boldsymbol{\beta}}^{oracle}_{\mathcal{A}_1} - \boldsymbol{\beta}^*_{\mathcal{A}_1})\|_2 \\
&\qquad\quad + (2\overline{c})\|\mathbf{X}_{\mathcal{A}_2}(\hat{\boldsymbol{\phi}}^{oracle}_{\mathcal{A}_2} - \boldsymbol{\phi}^*_{\mathcal{A}_2})\|_2\Big]/\sqrt{n} \\
&\qquad \leq (1+2\overline{c})M_0\phi^{1/2}_{\max}\|\hat{\boldsymbol{\theta}}^{oracle} - \boldsymbol{\theta}^*\|_2.
\end{aligned}
\tag{58}
$$

By Lemma 3 and Lemma 6 of Gu and Zou (2016), we get

$$
\begin{aligned}
&P\left(\{\|\nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c}\Psi_\tau(\boldsymbol{\beta}^*)-\nabla_{(\mathcal{A}_1\cup\mathcal{A}_2)^c}\Psi_\tau(\hat{\boldsymbol{\beta}}^{oracle})\|_{2,\infty}\geq a_1\lambda/2\}\right)\\
&\leq P\left(\|\hat{\boldsymbol{\theta}}^{oracle}-\boldsymbol{\theta}^*\|_2\geq\frac{a_1\lambda}{(1+2\overline{c})M_0\phi_{\max}^{1/2}}\right)\\
&\leq P\left(\left\|\frac{1}{n}\begin{pmatrix}\mathbf{X}_{\mathcal{A}_1}^\top W(\boldsymbol{\epsilon}+\boldsymbol{\eta})\\ \mathbf{X}_{\mathcal{A}_2}^\top W\boldsymbol{\eta}\end{pmatrix}\right\|_2\geq Q_2\lambda\right)\\
&\leq P\left(\left\|\frac{1}{n}\mathbf{X}_{\mathcal{A}_1}^\top W(\boldsymbol{\epsilon}+\boldsymbol{\eta})\right\|_2\geq Q_2\lambda/2\right)\\
&+P\left(\left\|\frac{1}{n}\mathbf{X}_{\mathcal{A}_2}^\top W\boldsymbol{\eta}\right\|_2\geq Q_2\lambda/2\right)\\
&\leq \Gamma(Q_2\lambda/2,n,s_{\mathcal{A}_1},K_1+K_2,M_0,M_1,M_1^2\rho_{1,max},\nu_1)\\
&\quad+\Gamma(Q_2\lambda/2,n,s_{\mathcal{A}_2},K_2,M_0M_1,M_1^2\rho_{2,max},\nu_2).
\end{aligned}
\tag{59}
$$

Combining (57), (58), and (59), it follows from Lemma 3 and Lemma 4 of Gu and Zou (2016) that

$$
\begin{aligned}
\pi_2=\;&2(p-s_{\mathcal{A}_1})\ \exp\\
&\left(-\frac{Cn\lambda^2a_1^2}{4M_0^2M_1^2(K_1+K_2)^2\overline{p}_{\mathcal{A}_1^c}^2}\right)\\
&+2(p-s_{\mathcal{A}_2})\ \exp\left(-\frac{Cn\lambda^2a_1^2}{4M_0^2M_1^2K_2^2\overline{p}_{\mathcal{A}_2^c}^2}\right)\\
&+\Gamma(Q_2\lambda/2,n,s_{\mathcal{A}_1},K_1\\
&+K_2,M_0,M_1,M_1^2\rho_{1,max},\nu_1)\\
&+\Gamma(Q_2\lambda/2,n,s_{\mathcal{A}_2},K_2,M_0M_1,M_1^2\rho_{2,max},\nu_2).
\end{aligned}
\tag{60}
$$

To derive the upper bound for $P(\mathcal{E}_3^c)$, we use assumption (A2) to get

$$
\min_{k\in\mathcal{A}_1}\|\hat{\boldsymbol{\beta}}^{oracle}\|\geq\min_{k\in\mathcal{A}_1}\|\boldsymbol{\beta}^*\|-\|\hat{\boldsymbol{\beta}}^{oracle}-\boldsymbol{\beta}^*\|_{2,\infty}
$$

and

$$
\min_{k\in\mathcal{A}_2}\|\hat{\boldsymbol{\phi}}^{oracle}\|\geq\min_{k\in\mathcal{A}_2}\|\boldsymbol{\phi}^*\|-\|\hat{\boldsymbol{\phi}}^{oracle}-\boldsymbol{\phi}^*\|_{2,\infty}.
$$

It follows that

$$
\begin{aligned}
P(\mathcal{E}_3^c) &\leq P(\|\hat{\theta}^{oracle} - \theta^*\|_{2,\infty} > \overline{R}) \leq P(\|\hat{\theta}^{oracle} - \theta^*\|_{\infty} > \frac{\overline{R}}{\overline{p}_k}) \\
&\leq P\left(\left\|\frac{1}{n}\begin{pmatrix}\mathbf{X}_{\mathcal{A}_1}^\top W(\epsilon+\eta)\\ \mathbf{X}_{\mathcal{A}_2}^\top W\eta\end{pmatrix}\right\|_2 \geq c_0\phi_{\min}\frac{\overline{R}}{\overline{p}_k}\right) \\
&\leq P\left(\left\|\frac{1}{n}\mathbf{X}_{\mathcal{A}_1}^\top W(\epsilon+\eta)\right\|_2 \geq c_0\phi_{\min}\frac{\overline{R}}{2\overline{p}_k}\right) \\
&\quad + P\left(\left\|\frac{1}{n}\mathbf{X}_{\mathcal{A}_2}^\top W\eta\right\|_2 \geq c_0\phi_{\min}\frac{\overline{R}}{2\overline{p}_k}\right) \\
&\leq \Gamma(c_0\phi_{\min}\frac{\overline{R}}{2\overline{p}_k}, n, s_{\mathcal{A}_1}, K_1+K_2, M_0, M_1, M_1^2\rho_{1,max}, \nu_1) \\
&\quad + \Gamma(c_0\phi_{\min}\frac{\overline{R}}{2\overline{p}_k}, n, s_{\mathcal{A}_2}, K_2, M_0M_1, M_1^2\rho_{2,max}, \nu_2).
\end{aligned}
\tag{61}
$$

This ends the proof of Theorem 7. □

Appendix 5: Checking KKT condition

We have been proved that the GPER and COGPER algorithms hold in the descent property. In the following section, we show that those algorithms converge to a stationary point by checking KKT conditions. Theorically, the solutions in (6) and (20) are established based on KKT conditions, then, they must always verify exactly KKT conditions. But, the numerical solution may fail the KKT conditions. more details are given (Yang and Zou 2015; Ouhourane et al. 2021). We define numerical KKT conditions for the GPER approach with the penalties GLasso, GMCP, GSCAD and GLLA, respectively, as follow

$$
\begin{cases}
\|\nabla_k\Psi_\tau(\boldsymbol{\beta}) + \lambda\omega_k.\frac{\beta_k}{\|\beta_k\|_2}\|_2 \leqslant \epsilon, & if\ \ \boldsymbol{\beta}_k \neq 0\\
\|\nabla_k\Psi_\tau(\boldsymbol{\beta})\|_2 \leqslant \lambda\omega_k + \epsilon, & if\ \ \boldsymbol{\beta}_k = 0,
\end{cases}
$$

$$
\begin{cases}
\|\nabla_k\Psi_\tau(\boldsymbol{\beta}) + \lambda\omega_k.\frac{\beta_k}{\|\beta_k\|_2} - \frac{\beta_k}{\theta}\|_2 \leqslant \epsilon, & if\ \ \boldsymbol{\beta}_k \neq 0 \ \text{ and } \ \|\boldsymbol{\beta}_k\|_2 \leqslant \theta\lambda\\
\|\nabla_k\Psi_\tau(\boldsymbol{\beta})\|_2 \leqslant \lambda\omega_k + \epsilon, & if\ \ \boldsymbol{\beta}_k = 0 \ \text{ and } \ \|\boldsymbol{\beta}_k\|_2 \leqslant \theta\lambda\\
\|\nabla_k\Psi_\tau(\boldsymbol{\beta})\|_2 \leqslant \epsilon = 0, & if\ \ \|\boldsymbol{\beta}_k\|_2 > \theta\lambda,
\end{cases}
$$

$$
\begin{cases}
\|\nabla_k\Psi_\tau(\boldsymbol{\beta}) + \lambda\omega_k.\frac{\beta_k}{\|\beta_k\|_2}\|_2 \leqslant \epsilon, & if\ \ \boldsymbol{\beta}_k \neq 0 \ \text{ and } \ \|\boldsymbol{\beta}_k\|_2 \leqslant \lambda\\
\|\nabla_k\Psi_\tau(\boldsymbol{\beta})\|_2 \leqslant \lambda\omega_k + \epsilon, & if\ \ \boldsymbol{\beta}_k = 0 \ \text{ and } \ \|\boldsymbol{\beta}_k\|_2 \leqslant \lambda\\
\|\nabla_k\Psi_\tau(\boldsymbol{\beta}) + \frac{\theta}{\theta-1}\lambda\omega_k.\frac{\beta_k}{\|\beta_k\|_2} - \frac{\beta_k}{(\theta-1)}\|_2 \leqslant \epsilon, & if\ \ \lambda < \|\boldsymbol{\beta}_k\|_2 \leqslant \theta\lambda\\
\|\nabla_k\Psi_\tau(\boldsymbol{\beta})\|_2 \leqslant \epsilon, & if\ \ \|\boldsymbol{\beta}_k\|_2 > \theta\lambda,
\end{cases}
$$

$$
\begin{cases}
\|\nabla_k\Psi_\tau(\boldsymbol{\beta}) + \lambda\omega'_k.\frac{\beta_k}{\|\beta_k\|_2}\|_2 \leqslant \epsilon, & if\ \ \boldsymbol{\beta}_k \neq 0\\
\|\nabla_k\Psi_\tau(\boldsymbol{\beta})\|_2 \leqslant \lambda\omega'_k + \epsilon, & if\ \ \boldsymbol{\beta}_k = 0.
\end{cases}
$$

To obtain the KKT conditions for COGPER approach, we replace $\Psi_\tau(.)$, $\boldsymbol{\beta}_k$ and (w_k, w_k') in the above KKT conditions by $S_\tau(.)$, $\boldsymbol{\beta}_k$ or $\boldsymbol{\phi}_k$ and (u_k, u_k'), respectively.

Funding This work is supported by the Natural Sciences and Engineering Research Council of Canada through individual discovery research grant to Karim Oualkacha and by the Fonds de recherche du Québec-Santé through individual Grant # 31110 to Karim Oualkacha.

Declaration

Conflict of interest We declare that there are no conflict of interest regarding the publication of this paper. Additionally, the data utilized in this study is publicly available.

References

Bickel PJ, Ritov Y, Tsybakov AB et al (2009) Simultaneous analysis of Lasso and Dantzig selector. Ann Stat 37(4):1705–1732

Bottai M, Frongillo EA, Sui X, O'Neill JR, McKeown RE, Burns TL, Liese AD, Blair SN, Pate RR (2014) Use of quantile regression to investigate the longitudinal association between physical activity and body mass index. Obesity 22(5):149–156

Bühlmann P, Van De Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, Berlin

Candes E, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n. Ann Stat 35(6):2313–2351

Chiolero A, Bovet P, Paccaud F (2005) Association between maternal smoking and low birth weight in switzerland: the eden study. Swiss Med Wkly 135(35–36):525–530

Daouia A, Gijbels I, Stupfler G (2019) Extremiles: A new perspective on asymmetric least squares. J Am Stat Assoc 114(527):1366–1381

Daouia A, Gijbels I, Stupfler G (2021) Extremile regression. Journal of the American Statistical Association, 1–8

Efron B (1991) Regression percentiles using asymmetric squared error loss. Stat Sin 1:93–125

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360

Fan J, Peng H (2004) Nonconcave penalized likelihood with a diverging number of parameters. Ann Stat 32(3):928–961

Fan J, Xue L, Zou H (2014) Strong oracle optimality of folded concave penalized estimation. Ann Stat 42(3):819

Gu Y, Zou H et al (2016) High-dimensional generalizations of asymmetric least squares regression and their applications. Ann Stat 44(6):2661–2694

Hashem H, Vinciotti V, Alhamzawi R, Yu K (2016) Quantile regression with group lasso for classification. Adv Data Anal Classif 10(3):375–390

Hertz JM, Schell G, Doerfler W (1999) Factors affecting de novo methylation of foreign dna in mouse embryonic stem cells. J Biol Chem 274(34):24232–24240

Hosmer Jr DW, Lemeshow S, Sturdivant RX (2013) Applied Logistic Regression vol. 398. John Wiley & Sons, ???

Huang J, Zhang C-H (2012) Estimation and selection via absolute penalized convex minimization and its multistage adaptive applications. J Mach Learn Res 13:1839–1864

Jiang C, Jiang M, Xu Q, Huang X (2017) Expectile regression neural network model with applications. Neurocomputing 247:73–86

Koenker R, Bassett G Jr (1978) Regression quantiles. Econom J Econom Soc 46:33–50

Koenker R, Zhao Q (1994) L-estimatton for linear heteroscedastic models. Journaltitle of Nonparametric Statistics 3(3–4):223–235

Lakhal-Chaieb L, Greenwood CM, Ouhourane M, Zhao K, Abdous B, Oualkacha K (2017) A smoothed em-algorithm for dna methylation profiles from sequencing-based methods in cell lines or for a single cell type. Statistical applications in genetics and molecular biology 16(5–6):333–347

Liao L, Park C, Choi H (2019) Penalized expectile regression: an alternative to penalized quantile regression. Ann Inst Stat Math 71(2):409–438

McGregor K, Bernatsky S, Colmegna I, Hudson M, Pastinen T, Labbe A, Greenwood C (2016) An evaluation of methods correcting for cell-type heterogeneity in dna methylation studies. Genome Biology **17**(84)

Meier L, Van De Geer S, Bühlmann P (2008) The group Lasso for logistic regression. J R Stat Soc Ser B (Methodol) 70(1):53–71

Meier L, Geer S, Bühlmann P et al (2009) High-dimensional additive modeling. Ann Stat 37(6B):3779–3821

Mitchell JA, Hakonarson H, Rebbeck TR, Grant SF (2013) Obesity-susceptibility loci and the tails of the pediatric BMI distribution. Obesity 21(6):1256–1260

Mkhadri A, Ouhourane M (2015) A group visa algorithm for variable selection. Statistical Methods & Applications 24(1):41–60

Mkhadri A, Ouhourane M, Oualkacha K (2017) A coordinate descent algorithm for computing penalized smooth quantile regression. Stat Comput 27(4):865–883

Newey WK, Powell JL (1987) Asymmetric least squares estimation and testing. Econom J Econom Soc 55:819–847

Ogutu JO, Piepho H-P (2014) Regularized group regression methods for genomic prediction: bridge, MCP, SCAD, group bridge, group Lasso, sparse group Lasso, group MCP and group SCAD. In: BMC proceedings. BioMed Central, p 7

Ouhourane M, Yang Y, Benedet AL, Oualkacha K (2021) Group penalized quantile regression. Statistical Methods & Applications, 1–35

Rudelson M, Vershynin R, et al (2013) Hanson-wright inequality and sub-gaussian concentration. Electronic Communications in Probability **18**

Sobotka F, Kauermann G, Waltrup LS, Kneib T (2013) On confidence intervals for semiparametric expectile regression. Stat Comput 23(2):135–148

Spady DW, Atrens MA, Szymanski WA (1986) Effects of mother's smoking on their infants' body composition as determined by total body potassium. Pediatr Res 20(8):716–719

Tang S, Cai Z, Fang Y, Lin M (2021) A new quantile treatment effect model for studying smoking effect on birth weight during mother's pregnancy. Journal of Management Science and Engineering 6(3):336–343

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B (Methodol) 58:267–288

Turgeon M, Oualkacha K, Ciampi A, Miftah H, Dehghan G, Zanke BW, Benedet AL, Rosa-Neto P, Greenwood CM, Labbe A, et al (2016) Principal component of explained variance: an efficient and optimal data dimension reduction framework for association studies. Statistical methods in medical research, 0962280216660128

Venables WN, Ripley BD (2013) Modern Applied Statistics with S-PLUS. Springer, ???

Vershynin R (2010) Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027

Wang L, Wu Y, Li R (2012) Quantile regression for analyzing heterogeneity in ultra-high dimension. J Am Stat Assoc 107(497):214–222

Wei F, Zhu H (2012) Group coordinate descent algorithms for nonconvex penalized regression. Comput Stat Data Anal 56(2):316–326

Wilcox AJ (1993) Birth weight and perinatal mortality: the effect of maternal smoking. Am J Epidemiol 137(10):1098–1104

Yang Y, Zou H (2015) Nonparametric multiple expectile regression via ER-boost. J Stat Comput Simul 85(7):1442–1458

Yang Y, Zou H (2015) A fast unified algorithm for solving group-lasso penalize learning problems. Stat Comput 25(6):1129–1141

Yang Y, Zhang T, Zou H (2018) Flexible expectile regression in reproducing kernel Hilbert spaces. Technometrics 60(1):26–35

Ye F, Zhang C-H (2010) Rate minimaxity of the Lasso and Dantzig selector for the l_q loss in l_r balls. J Mach Learn Res 11:3519–3540

Yousefi PD, Suderman M, Langdon R, Whitehurst O, Davey Smith G, Relton CL (2022) Dna methylation-based predictors of health: applications and statistical considerations. Nat Rev Genet 23:369–383

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B (Methodol) 68(1):49–67

Zhang C-H et al (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38(2):894–942

Zhao J, Zhang Y (2018) Variable selection in expectile regression. Commun Stat Theory Methods 47(7):1731–1746

Zhao J, Yan G, Zhang Y (2022) Robust estimation and shrinkage in ultrahigh dimensional expectile regression with heavy tails and variance heterogeneity. Stat Pap 63(1):1–28

Zou H, Li R (2008) One-step sparse estimates in nonconcave penalized likelihood models. Ann Stat 36(4):1509