Annals of the Institute of Statistical Mathematics (2021) 73:787–819
https://doi.org/10.1007/s10463-020-00766-z

High-dimensional sign-constrained feature selection and grouping

Shanshan Qin[1] · Hao Ding[1] · Yuehua Wu[1] · Feng Liu[2]

Received: 10 January 2020 / Revised: 4 September 2020 / Accepted: 8 September 2020 /
Published online: 12 October 2020

Abstract

In this paper, we propose a non-negative feature selection/feature grouping (nnFSG) method for general sign-constrained high-dimensional regression problems that allows regression coefficients to be disjointly homogeneous, with sparsity as a special case. To solve the resulting non-convex optimization problem, we provide an algorithm that incorporates the difference of convex programming, augmented Lagrange and coordinate descent methods. Furthermore, we show that the aforementioned nnFSG method recovers the oracle estimate consistently, and that the mean-squared errors are bounded. Additionally, we examine the performance of our method using finite sample simulations and applying it to a real protein mass spectrum dataset.

Keywords Difference convex programming · Feature grouping · Feature selection · High-dimensional · Non-negative

✉ Hao Ding
dinghaostat@gmail.com

Shanshan Qin
ssqin267@yorku.ca

Yuehua Wu
wuyh@yorku.ca

Feng Liu
feng.liu@uts.edu.au

[1] Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada

[2] Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, NSW 2007, Australia

1 Introduction

In recent decades, high-dimensional problems appear in many fields due to the increasing prevalence of big data. A classical model for data analysis is the linear regression model,

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i \ \ (i = 1, \dots , n), \tag{1}$$

where y_i are response observations, $\boldsymbol{x}_i = (x_{i1}, \dots , x_{ip})^\top$ are p-dimensional vectors of predictors, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of unknown regression coefficients, ϵ_i are random errors, and $\boldsymbol{x}_i$ are independent of ϵ_i. Regression analysis aims at identifying the relevant explanatory variables of the response and achieving high prediction accuracy (Rekabdarkolaee et al. 2017). In the high-dimensional setting, p is at least of the same order of magnitude as n, say $p = O(n)$ (p is not fixed), or $p >> n$, in which case $\boldsymbol{\beta}$ is usually assumed to be sparse, i.e., only a small set of elements are nonzero (Slawski and Hein 2013). For high-dimensional regression problems, regularization methods are of critical importance in a broad sense, and much work has been devoted to exploring sparseness of regression vectors. Examples include Bridge regression (Frank and Friedman 1993), Lasso (Tibshirani 1996), SCAD (Fan and Li 2001), elastic net (Zou and Hastie 2005), adaptive Lasso (Zou 2006), and MCP (Zhang 2010). Moreover, extracting one kind of lower-dimensional structure defined by groups has received increasing attention. One can turn to Tibshirani et al. (2005), Yuan and Lin (2006), Huang et al. (2009), She (2010), Jang et al. (2011), Tibshirani and Taylor (2011), Shen et al. (2012a), Yang et al. (2012), Zhu et al. (2013), Xiang et al. (2015), Arnold and Tibshirani (2016), and among others, for overviews of the literature. Methods introduced in the above articles intend to solve the problems where the regression vectors may contain some kind of structure, in which the vectors can be partitioned into disjoint, homogeneous subgroups.

Statistical modeling can entail many challenges stemming from the complexity of data. In high-dimensional regression models, there are some commonly stated constraints that should be imposed on the regression coefficients in order to avoid physically impossible or uninterpretable results. For instance, non-negativity is a common constraint when modeling non-negative data, e.g., time measurements, count data, chemical concentrations, intensity values of an image and economical quantities such as prices, incomes and growth rates (Slawski and Hein 2013).

The non-negativity constraint on the regression coefficients is an effective regularization technique for a certain class of high-dimensional regression problems. Slawski et al. (2012) proposed non-negative least squares (NNLS)/non-negative least absolute deviation (NNLAD) regression to extract patterns from a raw spectrum. Slawski and Hein (2013) showed that the performance of NNLS is comparable to that of Lasso in terms of prediction and estimation. Similarly, Meinshausen (2013) confirmed the effectiveness of the sign constraint for sparse recovery if explanatory variables are strongly correlated. Koike and Tanoue (2019) extended the results of Slawski and Hein (2013) and Meinshausen (2013)) to a more general setup, allowing for general convex loss functions and nonlinearity relationships between response and explanatory variables. Wen et al. (2015)

proposed a projection-based gradient descent method for solving NNLS problems and then applied it to the inverse problem of constructing a probabilistic Boolean network. Shadmi et al. (2019) investigated NNLS for recovering sparse non-negative vectors from noisy linear and biased measurements, as good as l_1 regularized estimations but without tuning parameters.

Other methods for dealing with such non-negative and sparse structures of regression coefficients combine the regularization techniques with non-negativity constraints, like, non-negative Lasso (Wu et al. 2014; Itoh et al. 2016), non-negative elastic net (Wu and Yang 2014) and non-negative adaptive Lasso (Yang and Wu 2016). Esser et al. (2013) added sparsity penalties, which are related to the ratio of l_1 and l_2 norms, to the objective function in an NNLS-type model to solve linear unmixing problems. Hu et al. (2015) applied a non-negative Lasso-based variable selection to identify the important amino acid sites and to evaluate their importance. Mandal and Ma (2016) proposed an efficient regularization path algorithm for generalized linear models with non-negative regression coefficients. Those methods are based on convex optimization with non-negative constraints.

To the best of our knowledge, many aspects of the sign-constrained feature selection/grouping remain unknown, including its theoretical properties and computational result. Note that, throughout this paper, we only consider non-negative constraints on the regression coefficients since one can replace the predictors that are imposed to be negative coefficients by their negative counterparts (Meinshausen 2013). In this paper, we propose a novel regularization scheme-based method to deal with the non-negative feature selection problem, i.e., the regression coefficients have sparse structures and non-negative constraints. In addition, our method is applicable to the feature grouping cases where those regression coefficients may contain homogeneous subgroups within which the elements are similar or identical. In light of Shen et al. (2012a), we initially introduce the nnFSG in its constrained form, followed by its regularized form. To solve the resulting non-convex regularization problem, we provide a hybrid algorithm that combines with the difference convex programming, augmented Lagrange and coordinate descent.

Our study makes four contributions. First, we propose a regularization scheme-based method that deals with a series of sign-constrained high-dimensional regression problems. It permits the regression coefficients to contain a structure of disjoint homogeneity, including sparsity as a special case. Second, to obtain a non-negative estimate, we initially adopt a penalty method instead of convex optimization with non-negative constraints. Although our method is in light of Shen et al. (2012a) by imposing non-negative constraints, one significant difference is the penalty function that we adopted shrinking these negative regression coefficients. Third, we leverage the associated proofs of Shen et al. (2012a), but we make the following theoretical improvements: (1) we introduce an oracle estimate defined by (5) that is based on the underlying true groupings with non-negative constraints, which is different from the one defined in Shen et al. (2012a). The probability of the event that the oracle estimate is not equal to the least squares estimate defined by (6) converges to 0 at a rate of $O(n^{-1}(\log n)^{1/2})$ under some mild conditions; (2) we further show that both the constrained estimate and the nnFSG estimate recover the oracle estimate consistently, and the mean-squared errors are bounded. Last but foremost, the proposed nnFSG outperforms other existing methods in terms of prediction accuracy, identifying informative variables and subgroups.

The rest of this paper is organized as follows. In Sect. 2, we introduce the constrained optimization problem. We develop the regularized nnFSG and an algorithm in Sect. 3, along with the convergence of the algorithm and the theoretical properties of the estimator. We present the numerical studies in Sect. 4. We conclude this paper in Sect. 5. The proofs of these lemmas and theorems are given in "Appendix".

Throughout the rest of this paper, the following notations and definitions will be used. We denote the design matrix $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\top = (\boldsymbol{x}_{(1)}, \dots, \boldsymbol{x}_{(p)})$. Define P_X as the projection matrix onto X. For any $\mathcal{A} \subset \{1, \dots, p\}$, $|\mathcal{A}|$ and $\mathcal{A}^c$ denote the size and the complement of $\mathcal{A}$, respectively. For any $\mathcal{B} \subset \mathcal{A}$, $\mathcal{A} \backslash \mathcal{B} = \mathcal{A} \cap \mathcal{B}^c$. We define $X_{\mathcal{A}} = (\boldsymbol{x}_{(j_1)}, \dots, \boldsymbol{x}_{(j_{|\mathcal{A}|})})$, a $n \times |\mathcal{A}|$ matrix indexed by $\mathcal{A} = \{j_1, \dots, j_{|\mathcal{A}|}\}$. Let $\mathbf{1}_d$ be a $d \times 1$ vector having elements 1. For a square matrix A, we define its smallest and largest eigenvalues by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. We use $I_{\{\cdot\}}$ and I to denote an indicator function and an identity matrix, respectively. We denote the l_2-norm, l_1-norm and l_∞-norm of a vector $\boldsymbol{a}$ by $\|\boldsymbol{a}\|$, $\|\boldsymbol{a}\|_1$, $\|\boldsymbol{a}\|_\infty$, respectively. We define a vector $\boldsymbol{a} \geq \mathbf{0}$, and thus has all components larger or equal to 0. For any $a, b \in \mathbb{R}$, $\min\{a, b\}$ and $\max\{a, b\}$ return the minimum and maximum of a and b, respectively. $\operatorname{sign}(a)$ is the sign of a. $a_+ = a$ if $a \geq 0$, otherwise $a_+ = 0$. We use $\Phi(\cdot)$ to denote the cumulative distribution function of the standard normal distribution.

2 Constrained optimization problem

Consider the linear regression model (1), where the regression coefficients are assumed to be sparse with non-negative constraints. Suppose that $\boldsymbol{\beta}^0$ is the true regression vector. Define the support of $\boldsymbol{\beta}^0$, $\mathcal{S} = \{j : \beta_j^0 > 0\}, j = 1, \dots, p$. The non-negative feature selection (nnFS) problem is formulated by the constrained least squares criterion

$$\min_{\boldsymbol{\beta} \geq \mathbf{0}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2, \tag{2}$$

subject to

$$\sum_{j=1}^{p} \min\left\{ \frac{|\beta_j|}{\tau}, 1 \right\} \leq s_1, \tag{3}$$

where $s_1 (> 0)$ is a tuning parameter that controls feature selection. $\tau > 0$, a threshold parameter, determines when a small regression coefficient should be penalized.

In particular, the unknown vector $\boldsymbol{\beta}$ may contain a structure with disjoint homogeneous subgroups within which the coordinates are identical or similar. Let the number of disjoint subgroups be $K + 1$ $(K \leq p - 1)$, and denote the coefficients index of k-th group by $\mathcal{G}_k$ satisfying $\cup_k \mathcal{G}_k = \{1, 2, \dots, p\}$ and $\cap_k \mathcal{G}_k = \emptyset$. Denote $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_1^\top, \dots, \boldsymbol{\beta}_K^\top)^\top$, where $\boldsymbol{\beta}_k = \alpha_k \mathbf{1}_{|\mathcal{G}_k|}$, $\alpha_0 = 0$ and $\alpha_k > 0$ for $k = 1, \dots, K$. In light of Shen et al. (2012a), the nnFSG problem is formulated by solving the problem (2) subjecting to

$$\sum_{j=1}^{p} \min\left\{\frac{|\beta_j|}{\tau}, 1\right\} \leq s_1, \text{and} \sum_{(j,j')\in\varepsilon} \min\left\{\frac{|\beta_j - \beta_{j'}|}{\tau}, 1\right\} \leq s_2, \tag{4}$$

where $\varepsilon = \{(j,j') : j < j', j, j' = 1, \ldots, p\}$, an arbitrary undirected graph. The tuning parameter, $s_2 (> 0)$, controls feature grouping. $\tau (> 0)$ also determines when a small difference between two coefficients should be penalized. More details on the constraints of (4) can be referred to Shen et al. (2012b, 2013). Note that the nnFSG problem is reduced to nnFS problem if $K = p - 1$. Throughout this paper, we thus only consider the nnFSG problem. Our goal is to estimate $\boldsymbol{\beta}$ or equivalently, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^\top$ and $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_K)^\top$. A solution to (2) subjecting to (4) will be referred to as a constrained nnFSG estimator, denoted by $\hat{\boldsymbol{\beta}}^{\text{cons}}$.

We denote the true grouping by $\mathcal{G}^0 = (\mathcal{G}_0^0, \mathcal{G}_1^0, \ldots, \mathcal{G}_{K^0}^0) = (\mathcal{G}_0^0, \mathcal{G}_0^{0^c})$, and the true regression parameter for the group k by α_k^0 for $k = 1, \ldots, K^0$, where $K^0 + 1$ is the true grouping number. Then $\boldsymbol{\beta}^0$ can be written as

$$\boldsymbol{\beta}^0 = \left(0\mathbf{1}_{|\mathcal{G}_0^0|}^\top, \alpha_1^0 \mathbf{1}_{|\mathcal{G}_1^0|}^\top, \ldots, \alpha_{K^0}^0 \mathbf{1}_{|\mathcal{G}_{K^0}^0|}^\top\right)^\top.$$

Denote $\boldsymbol{\alpha}^0 = (\alpha_1^0, \ldots, \alpha_{K^0}^0)^\top$, and $Z_{\mathcal{G}_0^{0c}} = (X_{\mathcal{G}_1^0}\mathbf{1}_{|\mathcal{G}_1^0|}, \ldots, X_{\mathcal{G}_{K^0}^0}\mathbf{1}_{|\mathcal{G}_{K^0}^0|})$, where $X_{\mathcal{G}_k^0}$ is the design matrix spanned by the predictors of $\mathcal{G}_k^0$.

Now, we define the oracle estimator,

$$\hat{\boldsymbol{\beta}}^{\text{ora}} = \left(\hat{\beta}_1^{ora}, \ldots, \hat{\beta}_p^{ora}\right)^\top = \left(0\mathbf{1}_{|\mathcal{G}_0^0|}^\top, \hat{\alpha}_1^{ora}\mathbf{1}_{|\mathcal{G}_1^0|}^\top, \ldots, \hat{\alpha}_{K^0}^{ora}\mathbf{1}_{|\mathcal{G}_{K^0}^0|}^\top\right)^\top, \tag{5}$$

where $\hat{\boldsymbol{\alpha}}^{\text{ora}} = (\hat{\alpha}_1^{ora}, \ldots, \hat{\alpha}_{K^0}^{ora})^\top$, satisfying that

$$\hat{\boldsymbol{\alpha}}^{\text{ora}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2n}\|\boldsymbol{y} - Z_{\mathcal{G}_0^{0c}}\boldsymbol{\alpha}\|^2, \ \ \alpha_k > 0, \ k = 1, \ldots, K^0.$$

We denote the least squares estimator by

$$\hat{\boldsymbol{\beta}}^{\text{ols}} = \left(\hat{\beta}_1^{\text{ols}}, \ldots, \hat{\beta}_p^{\text{ols}}\right)^\top = \left(0\mathbf{1}_{|\mathcal{G}_0^0|}^\top, \hat{\alpha}_1^{\text{ols}}\mathbf{1}_{|\mathcal{G}_1^0|}^\top, \ldots, \hat{\alpha}_{K^0}^{\text{ols}}\mathbf{1}_{|\mathcal{G}_{K^0}^0|}^\top\right)^\top, \tag{6}$$

where $\hat{\boldsymbol{\alpha}}^{\text{ols}} = (\hat{\alpha}_1^{\text{ols}}, \ldots, \hat{\alpha}_{K^0}^{\text{ols}})^\top = (Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})^{-1} Z_{\mathcal{G}_0^{0c}}^\top \boldsymbol{y}$ with $Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}}$ invertible. Note that both $\hat{\boldsymbol{\beta}}^{\text{ora}}$ and $\hat{\boldsymbol{\beta}}^{\text{ols}}$ are defined based on the true grouping $\mathcal{G}^0$.

Before proceeding, we provide two metrics proposed by Shen et al. (2012a) and Zhu et al. (2013), which reflect the model's difficulty, i.e.,

$$C_{\min} = \min_{\mathcal{G}\in\mathcal{T}} \frac{\|(I - P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\|^2}{n\max\{|\mathcal{G}_0 \backslash \mathcal{G}_0^0|, 1\}},$$

and

$$\gamma_{\min} = \min_{\{j,j' \in \mathcal{S}, (j,j') \in \varepsilon\}} \left\{ \beta_j^0, |\beta_j^0 - \beta_{j'}^0| \right\},$$

where $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_0^c) = (\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_K)$, $Z_{\mathcal{G}_0^c} = (X_{\mathcal{G}_1} \mathbf{1}_{|\mathcal{G}_1|}, \ldots, X_{\mathcal{G}_K} \mathbf{1}_{|\mathcal{G}_K|})$, $P_{Z_{\mathcal{G}_0^c}} = Z_{\mathcal{G}_0^c}(Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c})^{-1} Z_{\mathcal{G}_0^c}^\top$, $\mathcal{T} = \{\mathcal{G} \neq \mathcal{G}^0 : \sum_{j=1}^p I_{\{\beta_j > 0\}} \leq s_1^0, \sum_{(j,j') \in \varepsilon} I_{\{\beta_j \neq \beta_{j'}\}} \leq s_2^0\}$, a constrained set corresponding to (4), with $s_1^0 = |\mathcal{S}| = p - |\mathcal{G}_0^0|$, $s_2^0 = \sum_{(j,j') \in \varepsilon} I_{\{\beta_j^0 \neq \beta_{j'}^0\}}$. We remark that $C_{\min}$ defines the degree of separation between $\mathcal{G}_0^0$ and a least favorable candidate model for feature grouping and selection in the l_2-norm, while $\gamma_{\min}$ represents the resolution level of true regression coefficients. The smaller the values of $C_{\min}$ and $\gamma_{\min}$, the more difficult the situation. Denote

$$\bar{K} = \max_{1 \leq i \leq s_1^0} K_i^* / i,$$

where $K_i^* = \max_{\{\mathcal{G} \in \mathcal{T}, |\mathcal{G}_0 \backslash \mathcal{G}_0^0| = i\}} K(\mathcal{G}_0^c)$, and $K(\mathcal{G}_0^c)$ is the grouping number of $\mathcal{G}_0^c$. Let

$$\bar{T} = \max_{1 \leq i \leq s_1^0} \log T_i / i,$$

where $T_i = \max_{\{\mathcal{G} \in \mathcal{T}, |\mathcal{G}_0 \backslash \mathcal{G}_0^0| = i\}} |T_{\mathcal{G}_0^c}|$ and $T_{\mathcal{G}_0^c} = \{\mathcal{G} = (\mathcal{G}_0^*, \mathcal{G}_1, \ldots, \mathcal{G}_K) \in \mathcal{T} : \mathcal{G}_0^* = \mathcal{G}_0\}$, a set of groupings indexed by the sets of positive coefficients. More details on $C_{\min}$, $\gamma_{\min}$, $\bar{T}$ and $\bar{K}$ can be referred to Shen et al. (2012a) and Zhu et al. (2013).

Now, we make the following assumptions.

(A1) $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n.$

(A2) There exists a constant c_0 such that $\lambda_{\min}\left(n^{-1} Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}}\right) \geq c_0 > 0.$

(A3) For the same constant c_0 as in (A2), $\gamma_{\min} > [2\sigma^2 \log\{2nK^0/(2\pi)^{1/2}\}/(nc_0)]^{1/2}.$

Lemma 1 *Under the assumptions (A1)–(A3), it holds that*

$$\mathrm{pr}(\hat{\boldsymbol{\beta}}^{\mathrm{ora}} \neq \hat{\boldsymbol{\beta}}^{\mathrm{ols}}) = O\left(\frac{1}{n(\log n)^{1/2}}\right).$$

In Lemma 1, we show that $\min_{1 \leq k \leq K^0} \hat{\alpha}_k^{\mathrm{ols}} > 0$ with probability at least $1 - 2K^0\left\{1 - \Phi\left([2\log\{2nK^0/(2\pi)^{1/2}\}]^{1/2}\right)\right\}$, which implies that with probability at least $1 - 2K^0\left\{1 - \Phi\left([2\log\{2nK^0/(2\pi)^{1/2}\}]^{1/2}\right)\right\}$, $\hat{\boldsymbol{\beta}}^{\mathrm{ora}} = \hat{\boldsymbol{\beta}}^{\mathrm{ols}}$. More details can be found in the proof of Lemma 1.

Theorem 1 *Under the assumptions (A1)–(A3), it follows that, for any* $0 < \tau \leq \sigma[\log p / \{2np\lambda_{\max}(X^\top X)\}]^{1/2}$,

$$\mathrm{pr}\left(\hat{\boldsymbol{\beta}}^{\mathrm{cons}} \neq \hat{\boldsymbol{\beta}}^{\mathrm{ora}}\right) \leq \{\exp(1) + 1\} \exp(c^*) + \frac{c}{n(\log n)^{1/2}},$$

where $c^* = -10^{-1}\sigma^{-2}n\{C_{\min} - 10\sigma^2 n^{-1}(3\log p + \bar{T} + \bar{K}/2)\}$. *If, additionally,* $C_{\min} \geq 10\sigma^2 n^{-1}\left(\log n + 2^{-1}\log\log n + 3\log p + \bar{T} + 2^{-1}\bar{K}\right)$, *then*

1. $\mathrm{pr}\left(\hat{\boldsymbol{\beta}}^{\mathrm{cons}} \neq \hat{\boldsymbol{\beta}}^{\mathrm{ora}}\right) = O\left(n^{-1}(\log n)^{-1/2}\right)$;
2. $n^{-1}E\left\|X\hat{\boldsymbol{\beta}}^{\mathrm{cons}} - X\boldsymbol{\beta}^0\right\|^2 = n^{-1}K^0\sigma^2(1+o(1))$.

Note that $\hat{\boldsymbol{\beta}}^{\mathrm{cons}}$ yields a consistent recovery of $\hat{\boldsymbol{\beta}}^{\mathrm{ora}}$, and also generates a bounded mean-squared error.

3 Regularized optimization method

3.1 Penalty method

Before proceeding, we briefly describe the penalty method. The idea of a penalty method is to replace a constrained problem by an unconstrained problem. Consider the constrained problem

$$\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}) \text{ subject to } \beta_j \geq 0, j = 1, \dots, p, \tag{7}$$

where g is a continuous function on $\mathbb{R}^p$. Applying the idea of penalty method, problem (7) can be replaced by

$$\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}) + \lambda_3 p_3(\boldsymbol{\beta}), \tag{8}$$

where λ_3 is a positive constant and $p_3(\boldsymbol{\beta}) = \sum_{j=1}^{p}(\min\{\beta_j, 0\})^2$, satisfying: (1) p_3 is continuous, (2) $p_3(\boldsymbol{\beta}) \geq 0$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$, and (3) $p_3(\boldsymbol{\beta}) = 0$ if and only if $\beta_j \geq 0$, $j = 1, \dots, p$.

By the penalty method, the procedure for solving problem (7) are as follows. Let $\{\lambda_{3,k}\}$, $k = 1, 2, \dots$, be a sequence tending to infinity such that for each k, $\lambda_{3,k} \geq 0, \lambda_{3,k+1} > \lambda_{3,k}$. For each $\lambda_{3,k}$, problem (8) has a solution, denoted by $\boldsymbol{\beta}_k$. Luenberger and Ye (2015) showed the global convergence of the penalty method.

Theorem 2 *Let* $\{\boldsymbol{\beta}_k\}$ *be a sequence of solution to problem* (8) *for each* $\{\lambda_{3,k}\}$, $k = 1, 2, \dots$. *Then, any limit point of the sequence is a solution to problem* (7).

The proof of this theorem is provided on page 412 of Luenberger and Ye (2015). This theorem implies that there exists a large value M such that a solution to problem (8) is a solution to problem (7) if $\lambda_3 > M$. One can refer to Chapter 13 of Luenberger and Ye (2015) for more details. We also perform simulations to illustrate the effects of λ_3 in Sect. 4.2.

3.2 Formulation on the regularized nnFSG problem

By Lemma 1 of Shen et al. (2012a), the minimizer of

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

subject to

$$\sum_{j=1}^{p} \min\left\{\frac{|\beta_j|}{\tau}, 1\right\} \le s_1, \sum_{(j,j')\in\varepsilon} \min\left\{\frac{|\beta_j - \beta_{j'}|}{\tau}, 1\right\} \le s_2,$$

is a local minimizer of

$$f(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1 p_1(\boldsymbol{\beta}) + \lambda_2 p_2(\boldsymbol{\beta}),$$

where $p_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} \min\left\{|\beta_j|/\tau, 1\right\}$, and $p_2(\boldsymbol{\beta}) = \sum_{(j,j')\in\varepsilon} \min\left\{|\beta_j - \beta_{j'}|/\tau, 1\right\}$. We impose non-negative constraints on $\boldsymbol{\beta}$, i.e.,

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \text{ subject to } \beta_j \ge 0, j = 1, \dots, p. \tag{9}$$

Using the penalty method above, the regularized form of (9) is thus given by

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1 p_1(\boldsymbol{\beta}) + \lambda_2 p_2(\boldsymbol{\beta}) + \lambda_3 p_3(\boldsymbol{\beta}), \tag{10}$$

where $p_3(\boldsymbol{\beta}) = \sum_{j=1}^{p}(\min\{\beta_j, 0\})^2$. $\lambda_1(> 0), \lambda_2(\ge 0)$ correspond to s_1, s_2 in (4), respectively. $\lambda_3(> 0)$ controls the shrinkage speed of negative regression coefficients. Obviously, by setting $\lambda_2 = 0$, problem (10) reduces to the regularized nnFS problem, which solves the feature selection problems with non-negative constraints on the regression coefficients. A solution to (10), denoted by $\hat{\boldsymbol{\beta}}$, will be referred to as a nnFSG estimator.

Denote $S(\boldsymbol{\beta}) = (2n)^{-1}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1 p_1(\boldsymbol{\beta}) + \lambda_2 p_2(\boldsymbol{\beta}) + \lambda_3 p_3(\boldsymbol{\beta})$. Since $S(\boldsymbol{\beta})$ is non-convex, the difference of convex programming is thus applied to solve (10). Our main technical contribution is to extend the algorithm in Shen et al. (2012a) to a more general one by adding another penalty term $p_3(\boldsymbol{\beta})$, which, together with $p_1(\boldsymbol{\beta})$, controls the non-negativity of the regression coefficients.

Firstly, decompose the objective function $S(\boldsymbol{\beta})$ in (10) into the difference of two convex functions as follows,

$$S(\boldsymbol{\beta}) = S_1(\boldsymbol{\beta}) - S_2(\boldsymbol{\beta}), \tag{11}$$

where the convex functions $S_1(\boldsymbol{\beta})$ and $S_2(\boldsymbol{\beta})$ are given, respectively, by

$$S_1(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \frac{\lambda_1}{\tau}\sum_{j=1}^{p}|\beta_j| + \frac{\lambda_2}{\tau}\sum_{(j,j')\in\varepsilon}|\beta_j - \beta_{j'}| + \lambda_3\sum_{j=1}^{p}\beta_j^2,$$

$$S_2(\boldsymbol{\beta}) = \frac{\lambda_1}{\tau}\sum_{j=1}^{p}(|\beta_j| - \tau)_+ + \frac{\lambda_2}{\tau}\sum_{(j,j')\in\varepsilon}(|\beta_j - \beta_{j'}| - \tau)_+ + \lambda_3\sum_{j=1}^{p}((\beta_j)_+)^2.$$

Define $\boldsymbol{\eta} = (|\beta_1|, \dots, |\beta_p|, |\beta_{12}|, \dots, |\beta_{1p}|, \dots, |\beta_{(p-1)p}|, \beta_1^2, \dots, \beta_p^2)^\top$, where $\beta_{jj'} = \beta_j - \beta_j', (j,j') \in \varepsilon$. Then, $S_2(\boldsymbol{\beta})$ can be expressed to

$$\tilde{S}_2(\boldsymbol{\eta}) = \frac{\lambda_1}{\tau}\sum_{j=1}^{p}(|\beta_j| - \tau)_+ + \frac{\lambda_2}{\tau}\sum_{(j,j')\in\varepsilon}(|\beta_{jj'}| - \tau)_+ + \lambda_3\sum_{j=1}^{p}\beta_j^2 I_{\{\beta_j\geq 0\}}.$$

Approximate $\tilde{S}_2(\boldsymbol{\eta})$ by its affine minorization $\tilde{S}_2(\boldsymbol{\eta}^*) + \langle\boldsymbol{\eta} - \boldsymbol{\eta}^*, \partial\tilde{S}_2(\boldsymbol{\eta}^*)\rangle$ at a neighborhood of $\boldsymbol{\eta}^* \in \mathbb{R}^{(p^2+3p)/2}$, where $\partial\tilde{S}_2(\boldsymbol{\eta})$ is the first derivative of $\tilde{S}_2(\boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$; $\langle\cdot,\cdot\rangle$ is the inner product. Now we construct a sequence of approximations of $S_2(\boldsymbol{\beta})$ iteratively. At the m-th iteration, we replace $S_2(\boldsymbol{\beta})$ by $S_2^{(m)}(\boldsymbol{\beta}) = \tilde{S}_2^{(m)}(\boldsymbol{\eta}) = \tilde{S}_2(\hat{\boldsymbol{\eta}}^{(m-1)}) + \langle\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}^{(m-1)}, \partial\tilde{S}_2(\hat{\boldsymbol{\eta}}^{(m-1)})\rangle$. Specifically,

$$\begin{aligned}S_2^{(m)}(\boldsymbol{\beta}) = {} & S_2\left(\hat{\boldsymbol{\beta}}^{(m-1)}\right) + \frac{\lambda_1}{\tau}\sum_{j=1}^{p} I_{\{|\hat{\beta}_j^{(m-1)}|\geq\tau\}}\left(|\beta_j| - |\hat{\beta}_j^{(m-1)}|\right)\\ & + \frac{\lambda_2}{\tau}\sum_{(j,j')\in\varepsilon} I_{\{|\hat{\beta}_j^{(m-1)} - \hat{\beta}_{j'}^{(m-1)}|\geq\tau\}}\left(|\beta_j - \beta_{j'}| - |\hat{\beta}_j^{(m-1)} - \hat{\beta}_{j'}^{(m-1)}|\right)\\ & + \lambda_3\sum_{j=1}^{p} I_{\{\hat{\beta}_j^{(m-1)}\geq 0\}}\left(\beta_j^2 - \left(\hat{\beta}_j^{(m-1)}\right)^2\right).\end{aligned}$$

Finally, an approximation function to $S(\boldsymbol{\beta})$ in (11) at the m-th iteration can be obtained by $S^{(m)}(\boldsymbol{\beta}) = S_1(\boldsymbol{\beta}) - S_2^{(m)}(\boldsymbol{\beta})$, which formulates the following subproblem,

$$\min_{\beta} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \frac{\lambda_1}{\tau}\sum_{j\in\mathcal{F}^{(m-1)}}|\beta_j| + \frac{\lambda_2}{\tau}\sum_{(j,j')\in\varepsilon^{(m-1)}}|\beta_j - \beta_{j'}| + \lambda_3\sum_{j\in\mathcal{N}^{(m-1)}}\beta_j^2, \tag{12}$$

where

$$\begin{aligned}\mathcal{F}^{(m-1)} &= \left\{j : |\hat{\beta}_j^{(m-1)}| < \tau\right\},\\ \varepsilon^{(m-1)} &= \left\{(j,j') : j < j', |\hat{\beta}_j^{(m-1)} - \hat{\beta}_{j'}^{(m-1)}| < \tau\right\},\\ \mathcal{N}^{(m-1)} &= \left\{j : \hat{\beta}_j^{(m-1)} < 0\right\}.\end{aligned} \tag{13}$$

How to efficiently solve the subproblem (12) plays a key role in solving the problem (10). Though we can apply quadratic programming to solve the subproblem (12), it is inefficient for large-scale problems.

3.3 Algorithm

For the subproblem (12), it is necessary to develop an effective computational strategy. In light of Shen et al. (2012a), an algorithm integrated with augmented Lagrange and coordinate descent methods is developed to solve the subproblem (12).

We convert the subproblem (12) with linear constraints to its unconstrained version through slack variables $\beta_{jj'} = \beta_j - \beta_{j'}$. Define

$$\boldsymbol{\xi} = (\beta_1, \dots, \beta_p, \beta_{12}, \dots, \beta_{1p}, \dots, \beta_{(p-1)p})^\top.$$

Then an augmented equivalent problem of (12) is given, i.e.,

$$\min_{\xi} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 + \frac{\lambda_1}{\tau} \sum_{j \in \mathcal{F}^{(m-1)}} |\beta_j| + \frac{\lambda_2}{\tau} \sum_{(j,j') \in \varepsilon^{(m-1)}} |\beta_{jj'}| + \lambda_3 \sum_{j \in \mathcal{N}^{(m-1)}} \beta_j^2. \tag{14}$$

For (14), the augmented Lagrange is employed to solve its equivalent unconstrained problem iteratively with respect to t at the m-th iteration. Denote $\tilde{S}^{(m)}(\boldsymbol{\xi}) = (2n)^{-1} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1 \tau^{-1} \sum_{j \in \mathcal{F}^{(m-1)}} |\beta_j| + \lambda_2 \tau^{-1} \sum_{(j,j') \in \varepsilon^{(m-1)}} |\beta_{jj'}| + \lambda_3 \sum_{j \in \mathcal{N}^{(m-1)}} \beta_j^2$. In the t-th iteration, we minimize

$$\begin{aligned} \bar{S}^{(m)}(\boldsymbol{\xi}) = \tilde{S}^{(m)}(\boldsymbol{\xi}) &+ \sum_{(j,j') \in \varepsilon^{(m-1)}} \tau_{jj'}^{(t)} \left(\beta_j - \beta_{j'} - \beta_{jj'}\right) \\ &+ \frac{1}{2} \nu^{(t)} \sum_{(j,j') \in \varepsilon^{(m-1)}} \left(\beta_j - \beta_{j'} - \beta_{jj'}\right)^2, \end{aligned} \tag{15}$$

where $\tau_{jj'}^{(t)}$, $\nu^{(t)}$ are Lagrange multipliers. Update $\tau_{jj'}$ and ν by

$$\tau_{jj'}^{(t+1)} = \tau_{jj'}^{(t)} + \nu^{(t)} \left(\hat{\beta}_j^{(m,t)} - \hat{\beta}_{j'}^{(m,t)} - \hat{\beta}_{jj'}^{(m,t)}\right) \quad \text{and} \quad \nu^{(t+1)} = \rho \nu^{(t)}, \tag{16}$$

where ρ controls the speed of convergence. To speed convergence, ρ is chosen to be larger than 1.

We use the coordinate descent method to compute $\hat{\boldsymbol{\xi}}^{(m,t)}$ in terms of (15). For each component of $\boldsymbol{\xi}$, we fix the other components at their current values. Set an initial value $\hat{\boldsymbol{\xi}}^{(m,0)} = \hat{\boldsymbol{\xi}}^{(m-1)}$, where $\hat{\boldsymbol{\xi}}^{(m-1)}$ is the solution of the subproblem (12). Then update $\hat{\boldsymbol{\xi}}^{(m,t)}$ by the following formulas, $t = 1, 2, \dots$.

1. Given $\hat{\beta}_j^{(m,t-1)}$, updating $\hat{\beta}_j^{(m,t)}$ $(j = 1, 2, \dots, p)$ by:

$$\hat{\beta}_j^{(m,t)} = \alpha^{-1} \gamma, \tag{17}$$

where

$$\alpha = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 + 2\lambda_3 I_{\{\hat{\beta}_j^{(m-1)} < 0\}} + \nu^{(t)} \left| j' : (j,j') \in \varepsilon^{(m-1)} \text{ or } (j',j) \in \varepsilon^{(m-1)} \right|.$$

Let

$$\gamma^* = \frac{1}{n}\sum_{i=1}^{n} x_{ij} b_{i,j}^{(m,t)} - \sum_{(j,j')\in\varepsilon^{(m-1)}} \tau_{jj'}^{(t)} + v^{(t)} \sum_{(j,j')\in\varepsilon^{(m-1)}} \left(\hat{\beta}_{j'}^{(m,t)} + \hat{\beta}_{jj'}^{(m,t)}\right).$$

Then $\gamma = \gamma^*$ if $|\hat{\beta}_j^{(m-1)}| \geq \tau$, i.e., $j \in \mathscr{F}^{(m-1)^c}$. Otherwise, $\gamma = ST\left(\gamma^*, \lambda_1/\tau\right)$. Herein, $b_{i,j}^{(m,t)} = y_i - \boldsymbol{x}_{i(j)}^{\top}\hat{\boldsymbol{\beta}}_{(j)}^{(m,t)}$; $\boldsymbol{x}_{i(j)}$ is the vector $\boldsymbol{x}_i$ after deleting the j-th element; x_{ij} is the j-th element of vector $\boldsymbol{x}_i$; $\tau_{jj'} = -\tau_{j'j}$ if $j > j'$; $\beta_{jj'} = -\beta_{j'j}$ if $j > j'$. $ST(b,\delta) = \text{sign}(b)(|b| - \delta)_+$ is the soft-thresholding operator.

2. Given $\hat{\beta}_{jj'}^{(m,t-1)}$, updating $\hat{\beta}_{jj'}^{(m,t)}, (1 \leq j < j' \leq p)$ by:

$$\hat{\beta}_{jj'}^{(m,t)} = \begin{cases} (v^{(t)})^{-1} ST\left(\tau_{jj'}^{(t)} + v^{(t)}(\hat{\beta}_j^{(m,t)} - \hat{\beta}_{j'}^{(m,t)}), \frac{\lambda_2}{\tau}\right) & (j,j') \in \varepsilon^{(m-1)}, \\ \hat{\beta}_{jj'}^{(m-1)} & (j,j') \in \varepsilon^{(m-1)^c}. \end{cases} \tag{18}$$

The process of coordinate descent iterates until convergence, which satisfies the terminate condition $\|\hat{\boldsymbol{\beta}}^{(m,t)} - \hat{\boldsymbol{\beta}}^{(m,t-1)}\|_\infty \leq \delta^*$, where δ^* is a given small positive value, say, 10^{-5}. Hence, $\hat{\boldsymbol{\beta}}^{(m)} = \hat{\boldsymbol{\beta}}^{(m,t^*)}$, where t^* denotes the iteration at termination. The pseudo codes of the developed algorithm are summarized in Algorithm 1. The convergence of the algorithm is given in Theorem 3. When solving the problem (10), the proposed method could potentially lead to a local optimum as the objective function in (10) is non-convex. Hence, it is critical to assign a suitable initial value $\hat{\boldsymbol{\beta}}^{(0)}$. Possible candidate initial values are ones estimated by the R package `glmnet` (Friedman et al. 2016) or `nnls` (Mullen and van Stokkum 2012). We remark that our numerical studies indicate that the algorithm still converges if we only update non-zero $\hat{\beta}_j^{(m,t)}$ in (17).

Algorithm 1 A hybrid algorithm integrated with augmented Lagrange and coordinate descent

```
Input: design matrix $X \in \mathbb{R}^{n\times p}$, response vector $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$, parameters $\tau, \lambda_1, \lambda_2, \lambda_3, \rho, v, \delta^*$.
Output: $\hat{\boldsymbol{\beta}}^{(m)}$
 Initialization: $\hat{\boldsymbol{\beta}}^{(0)}, m = 0$
 do
     $m \leftarrow m + 1$.
     Update $\mathcal{F}^{(m)}, \varepsilon^{(m)}, \mathcal{N}^{(m)}$ according to (13).
     Initialization: $\hat{\boldsymbol{\beta}}^{(m,0)} \leftarrow \hat{\boldsymbol{\beta}}^{(m-1)}, t = 0$
     do
         $t \leftarrow t + 1$.
         Update $\hat{\beta}_l^{(m,t)}$ according to updating formulas (17).
         Update $\hat{\beta}_{jj'}^{(m,t)}$ according to updating formula (18).
     while $\|\hat{\boldsymbol{\beta}}^{(m,t)} - \hat{\boldsymbol{\beta}}^{(m,t-1)}\|_\infty \geq \delta^*$
 while $S(\hat{\boldsymbol{\beta}}^{(m)}) - S(\hat{\boldsymbol{\beta}}^{(m+1)}) > 0$
```

Theorem 3 *The proposed algorithm 1 converges. That is*

$$S(\hat{\boldsymbol{\beta}}^{(m)}) \to c, \text{ as } m \to +\infty, \tag{19}$$

where c is a non-negative constant.

We derive the convergence of the proposed algorithm that is analogous to Shen et al. (2012a). Next, we show some properties of the proposed nnFSG estimator $\hat{\boldsymbol{\beta}}$. Before proceeding, we make the following assumption.

(A4) $4\tau^{-2}(\lambda_1 s^* + \lambda_2|\mathcal{N}|) < \min_{K(\mathcal{G}_0^c)\le K^*} \lambda_{\min}(n^{-1}Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c})$, where s^* and K^* are the upper bounds of maximal number of non-zero coefficients and non-zero groupings, respectively. $|\mathcal{N}|$ is the maximal number of direct connections of variable x_j to variable $x_{j'}$, where $(j,j') \in \varepsilon$ and $j,j' \in \mathcal{G}_k, k = 1,\dots,K$.

We remark that for a full connection $\varepsilon = \{(j,j') : j < j', j,j' = 1,\dots,p\}$, $|\mathcal{N}| = s^*(s^*-1)/2$. s^* and K^* are different from the s_1^0 and K^0, respectively. Specifically, $s_1^0 \le s^* \le p$, $K^0 \le K^* \le s^*$.

Theorem 4 *Under the assumptions (A1)- (A4), if* $\gamma_{\min} > 2\tau$,

$$\left\{(\gamma_{\min} - 2\tau)n^{1/2}\lambda_{\min}^{1/2}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})\sigma^{-1}\right\}^2 \ge \max\left\{8\log\frac{nK^0(K^0-1)}{(2\pi)^{1/2}}, 2\log\frac{2n(p-|\mathcal{G}_0^0|)}{(2\pi)^{1/2}}\right\},$$

$$\left(\frac{n\lambda_1/\tau}{\sigma \max\limits_{1\le j\le p}\|\boldsymbol{x}_{(j)}\|}\right)^2 \ge 2\log\frac{2n|\mathcal{G}_0^0|}{(2\pi)^{1/2}}, \quad \left(\frac{n\lambda_2/\tau}{2\sigma\mathcal{D}}\right)^2 \ge 2\log\frac{2n|\mathcal{N}|}{(2\pi)^{1/2}},$$

where $\mathcal{D} = \max_{k,A\subset\mathcal{G}_k^0}\|X_A\mathbf{1}\|/|\varepsilon\cap\{A\times(\mathcal{G}_k^0\backslash A)\}|$, *and* $'\times'$ *denotes the Cartesian product, then*

$$\mathrm{pr}\left(\hat{\boldsymbol{\beta}} \ne \hat{\boldsymbol{\beta}}^{\mathrm{ora}}\right) = O\left(\frac{1}{n(\log n)^{1/2}}\right).$$

Furthermore, if

$$\frac{1}{n}\|X\boldsymbol{\beta}^0\|^2 + \frac{\tau^2}{16}\min_{K(\mathcal{G}_0^c)\le K^*}\lambda_{\min}\left(\frac{1}{n}Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c}\right) = o(K^0(\log n)^{1/2}),$$

then we have

$$\frac{1}{n}E\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 = \frac{K^0\sigma^2}{n}(1+o(1)).$$

Note that the results of Theorem 4 are parallel to that of Theorem 1. We remark that the feature selection problem, which only contains a zero group, can be regarded as a special case of our problem.

4 Numerical studies

4.1 Evaluation measures

The criteria used for measuring the prediction accuracy of the estimate $\hat{\boldsymbol{\beta}}$ are the mean-squared error (MSE), $\mathrm{MSE} = n^{-1}\|X\boldsymbol{\beta}^0 - X\hat{\boldsymbol{\beta}}\|^2$, and mean absolute error (MAE), $\mathrm{MAE} = n^{-1}\|X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\|_1$. Since the regression vector is sparse and may also contain a structure with disjoint subgroups, in light of (Yang et al. 2012), we thus provide another two metrics, feature true positive rate (FTP),

$$\mathrm{FTP} = \frac{\sum_{j\in\mathcal{G}_0^0} I_{\{\hat{\beta}_j=0\}} + \sum_{j\notin\mathcal{G}_0^0} I_{\{\hat{\beta}_j\neq 0\}}}{p},$$

and group true positive rate (GTP),

$$\mathrm{GTP} = \frac{\sum_{k=1}^{K^0}\mathrm{GTP}_k + \mathrm{FTP}}{K^0+1},$$

where

$$\mathrm{GTP}_k = \frac{\sum_{i\neq j, i,j\in\mathcal{G}_k^0} I_{\{\hat{\beta}_i=\hat{\beta}_j\}} + \sum_{i\neq j, i\in\mathcal{G}_k^0, j\notin\mathcal{G}_k^0} I_{\{\hat{\beta}_i\neq\hat{\beta}_j\}}}{|\mathcal{G}_k^0|(p-1)}, \quad k = 1,\dots,K^0.$$

FTP and GTP measure the accuracy of method's performance in terms of feature selection and feature grouping. It is clear that FTP, $\mathrm{GTP}_k (k = 1,\dots,K^0)$ and GTP $\in [0,1]$. Ideally, they should be close to 1.

4.2 Tuning free parameter: λ_3

λ_3 shrinks the negative coordinates of $\boldsymbol{\beta}$, which, together with λ_1, controls the non-negativity. We perform 500 simulations to illustrate the effects of λ_3 by fixing $\tau, \lambda_1, \lambda_2$. We generate the samples $(\boldsymbol{x}_i, y_i), i = 1,\dots,n$, from the linear model $y_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} + \epsilon_i$, where $\boldsymbol{x}_i \overset{iid}{\sim} N_p(\mathbf{0}, \Sigma)$ with $\Sigma = (\sigma_{\ell j})$ and $\sigma_{\ell j} = 0.5^{|\ell - j|}, \ell, j = 1,\dots,p$; the random error $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$. We set the true regression coefficient $\boldsymbol{\beta}^0 = (\underbrace{1,\dots,1}_{4}, \underbrace{2,\dots,2}_{4}, \underbrace{3,\dots,3}_{4}, \underbrace{4,\dots,4}_{4}, \underbrace{0,\dots,0}_{p-16})^\top \in \mathbb{R}^p$. Let $\tau = 0.1, \lambda_1 = \lambda_2 = 10^{-3}\bar{\lambda}$, where $\bar{\lambda} = \|X^\top\boldsymbol{y}\|_\infty$, and $\lambda_3 \in \{0, 1, 5, 10, 15\}$. Herein, we take $\sigma = 1, n = 100$, $p = 500, 1000, 2000$.

Define STP as the proportion of non-negative coordinates of $\hat{\boldsymbol{\beta}}$, i.e., $\mathrm{STP} = \sum_{j=1}^p I_{\{\hat{\beta}_j \geq 0\}}/p$. Figure 1 displays the values of STP, MAE, FTP and GTP, averaged over 500 simulations for the post samples. In an instance where p is fixed, as λ_3 increases, the values of STP, FTP and GTP increase slightly, while the values of MAE decrease. Nonetheless, all the measures tend to be stable as λ_3 exceeds a critical value. For example, when $p = 500$, the FTPs, GTPs and MAEs no longer

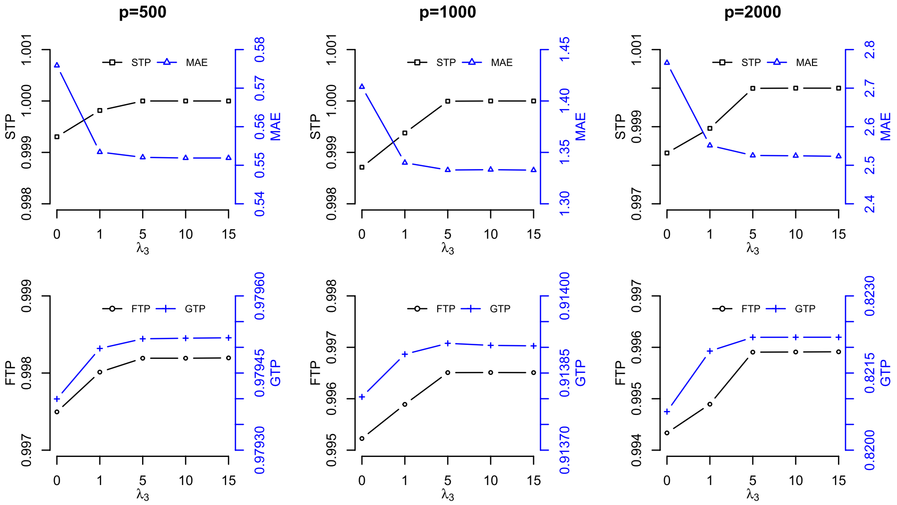

Fig. 1 The values of STP, MAE, FTP and GTP with different λ_3 by fixing $\lambda_1, \lambda_2, \tau$, averaged over 500 simulations. The upper three panels show STP (left black axis) and MAE (right blue axis) when $p = 500$, 1000, and 2000, while the lower three panels present the corresponding FTP (left black axis) and GTP (right blue axis)

change with λ_3 and the STP values are exactly 1 as $\lambda_3 \geq 5$. We arrive at the same results for the other two instances, say, $p = 1000, 2000$. It is noted that when $\lambda_3 = 0$, nnFSG is reduced to the model proposed by Shen et al. (2012a). When the underlying true regression coefficients are non-negative, the penalty $p_3(\boldsymbol{\beta})$ involved in nnFSG helps to increase the capacity of prediction accuracy of $\hat{\boldsymbol{\beta}}$ as well as its feature selection and grouping slightly. Although the regularization nnFSG method contains four parameters, say, $\tau, \lambda_1, \lambda_2$ and λ_3, the amount of work to select tuning parameters is parallel to that of Shen et al. (2012a). The introduction of $p_3(\boldsymbol{\beta})$ achieves non-negative estimates with the associated λ_3 free tuning. We remark that the critical value may be different under different model settings. In a real application, we thus take a large number of λ_3, say, 10, or even larger.

A natural question that one may ask is regarding the computing time of nnFSG. To estimate $\hat{\boldsymbol{\beta}}$, the algorithm involves the difference of convex programming, Lagrange, and coordinate descent methods. Though it appears to be complex, the updating formulas of (17) and (18) are explicit. Under the above settings, and let $\lambda_3 = 10$, the average time (seconds) for $p = 500, 1000, 2000$ are 13.87 s, 84.96 s and 371.29 s, respectively. We conduct simulation studies in the R programming environment. The machine we used equips Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00 GHZ.

4.3 Model comparisons of non-negative feature selection

In our simulation study, we are interested in the performance of our proposed method in feature selection. We carry out simulations via the linear model

$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i$, $i = 1, \ldots, n$. The positive elements of the true coefficient vector $\boldsymbol{\beta}^0$ are randomly generated from a uniform distribution [0.5, 5], $s_1^0 = 10$. The setting of $\boldsymbol{x}_i$ is the same as in Sect. 4.2. And the noise term $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Put $\lambda_2 = \lambda_1, \lambda_3 = 10$. λ_1 and τ are selected from candidate sets using fivefold cross-validation. The method's performance is assessed using the simulation settings with $n = 100, p = 100, 500, 1000, 2000, \sigma = 0.5, 1, 2$.

We compare our method with others that also achieve non-negative estimators and are available in the R packages, such as `nnls`, `glmnet`, `penalized` (Goeman 2010), `CVXR` (Fu et al. 2017). The comparisons are based on how well they estimate the true underlying parameters, measured using MSE and MAE; and how well they perform in terms of feature selection, measured using FTP. The larger the values of FTP, the better the performance of feature selection. Table 1 reports the average and standard deviations of MSE, MAE and FTP, which are obtained based on 500 simulations. As Table 1 illustrates, nnFSG outperforms the other methods in terms of MSE, MAE and FTP uniformly.

4.4 Synthetic malaria vaccine data

Not all sites in amino acid (AA) sequence have equal importance due to the structures of protein. In vaccine design study, it is thus very crucial to locate the important AA sites. A vaccine that is designed to match those important AA sites can improve induced immunity. Furthermore, the sites associated with immune response with negative coefficients should be excluded from the model (Hu et al. 2015). A non-negative lasso method was thus applied. For some confidential reasons, we are not allowed to access the original data. We thus assess the performance of our proposed method using the synthetic, but realistic, data under the simulation benchmarks, similar to what was done in Hu et al. (2015). In this paper, three cases are considered under the settings of $n = 500, p = 3000, s_1^0 = 24$, which is more challenging than that in their article where $n = 100$ and $p = 90$. Suppose that the explanatory variables are all independent. We randomly generate the i-th sample $(\boldsymbol{x}_i, y_i)$ via the linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i$. Denote $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$, where $x_{ij} \sim$ Bernoulli (p_j), $p_j \sim$ Beta(2, 5) for $j = 1, \ldots, p$. And $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$. Consider three cases for the underlying true $\boldsymbol{\beta}^0$,

- Case I: $|\mathcal{G}^0| = 2$, $\boldsymbol{\beta}^0 = (\underbrace{10, \ldots, 10}_{s_1^0}, \underbrace{0, \ldots, 0}_{p-s_1^0})^\top$.
- Case II: $|\mathcal{G}^0| = 3$, $\boldsymbol{\beta}^0 = (\underbrace{2, \ldots, 2}_{s_1^0/2}, \underbrace{1, \ldots, 1}_{s_1^0/2}, \underbrace{0, \ldots, 0}_{p-s_1^0})^\top$.
- Case III: $|\mathcal{G}^0| = 4$, $\boldsymbol{\beta}^0 = (\underbrace{1, \ldots, 1}_{s_1^0/3}, \underbrace{0.5, \ldots, 0.5}_{s_1^0/3}, \underbrace{0.3, \ldots, 0.3}_{s_1^0/3}, \underbrace{0, \ldots, 0}_{p-s_1^0})^\top$.

β_j $(j = 1, \ldots, p)$ within a subgroup implies that the associated AA sites have equal importance. We remark that the order of the elements of $\boldsymbol{\beta}^0$ is randomly given. Note that y_i are the immune response observations that are usually measured by the growth inhibition assay. The source of measurement error may be systematic.

Table 1 Comparison of glmnet, penalized, CVXR, nnls and nnFSG under the settings with $n = 100, p = 100, 500, 1000, 2000; \sigma = 0.5, 1, 2$

p	Methods	$\sigma = 0.5$			$\sigma = 1$			$\sigma = 2$		
		MSE	MAE	FTP	MSE	MAE	FTP	MSE	MAE	FTP
100	glmnet	0.0745 (0.0339)	0.2128 (0.0478)	0.9693 (0.0211)	0.2402 (0.1029)	0.3832 (0.0820)	0.9187 (0.0382)	1.0019 (0.4436)	0.7814 (0.1713)	0.9150 (0.0454)
	Penalized	0.0655 (0.0309)	0.1991 (0.0465)	0.9421 (0.0270)	0.2572 (0.1188)	0.3951 (0.0894)	0.9416 (0.0257)	1.0477 (0.4942)	0.7961 (0.1858)	0.9420 (0.0271)
	CVXR	0.0648 (0.0307)	0.1979 (0.0463)	0.7697 (0.1436)	0.6138 (0.2301)	0.6164 (0.1171)	0.6961 (0.1362)	1.0358 (0.4910)	0.7914 (0.1851)	0.7020 (0.0946)
	nnls	0.1554 (0.0586)	0.3096 (0.0575)	0.7224 (0.0427)	0.6138 (0.2301)	0.6164 (0.1171)	0.7256 (0.0441)	2.4786 (0.9224)	1.2368 (0.2278)	0.7220 (0.0427)
	nnFSG	0.0392 (0.0335)	0.1469 (0.0589)	1.0000 (0.0000)	0.1765 (0.1121)	0.3206 (0.1013)	0.9976 (0.0057)	0.6383 (0.3524)	0.6179 (0.1645)	0.9954 (0.0065)
500	glmnet	0.1032 (0.0406)	0.2516 (0.0503)	0.9826 (0.0082)	0.4008 (0.1587)	0.4960 (0.0992)	0.9665 (0.0168)	1.6459 (0.6945)	1.0044 (0.2104)	0.9646 (0.0205)
	Penalized	0.1128 (0.0490)	0.2621 (0.0565)	0.9836 (0.0058)	0.4513 (0.1961)	0.5242 (0.1130)	0.9836 (0.0059)	1.7973 (0.7788)	1.0461 (0.2248)	0.9835 (0.0059)
	CVXR	0.1109 (0.0476)	0.2600 (0.0555)	0.6585 (0.1499)	0.4438 (0.1903)	0.5202 (0.1111)	0.6364 (0.0882)	1.7680 (0.7576)	1.0382 (0.2213)	0.6918 (0.0917)
	nnls	1.3167 (0.5088)	0.9040 (0.1648)	0.8239 (0.0021)	5.2825 (1.9996)	1.8102 (0.3254)	0.8223 (0.0022)	21.531 (8.7338)	3.6487 (0.6863)	0.8189 (0.0030)
	nnFSG	0.0418 (0.0326)	0.1522 (0.0604)	1.0000 (0.0000)	0.1564 (0.0974)	0.3027 (0.0928)	0.9999 (0.0005)	0.7428 (0.4988)	0.6581 (0.2031)	0.9988 (0.0013)
1000	glmnet	0.1140 (0.0445)	0.2654 (0.0517)	0.9874 (0.0055)	0.4609 (0.1785)	0.5334 (0.1029)	0.9782 (0.0116)	1.8752 (0.7348)	1.0749 (0.2103)	0.9771 (0.0140)
	Penalized	0.1325 (0.0567)	0.2855 (0.0599)	0.9915 (0.0029)	0.5304 (0.2276)	0.5710 (0.1199)	0.9915 (0.0029)	2.1115 (0.8735)	1.1401 (0.2344)	0.9916 (0.0030)
	CVXR	0.1300 (0.0562)	0.2827 (0.0596)	0.6754 (0.0931)	0.5204 (0.2256)	0.5655 (0.1195)	0.1457 (0.0500)	2.0714 (0.8636)	1.1290 (0.2331)	0.8871 (0.0926)
	nnls	0.6416 (0.1951)	0.6328 (0.0944)	0.9158 (0.0014)	2.6942 (0.8139)	1.2960 (0.1936)	0.9113 (0.0013)	11.607 (4.6156)	2.6798 (0.4663)	0.9113 (0.0018)
	nnFSG	0.0415 (0.0340)	0.1509 (0.0625)	1.0000 (0.0000)	0.1562 (0.0973)	0.3026 (0.0944)	0.9999 (0.0004)	0.7837 (0.4986)	0.6807 (0.2032)	0.9993 (0.0007)
2000	glmnet	0.1326 (0.0518)	0.2859 (0.0556)	0.9917 (0.0036)	0.5379 (0.2094)	0.5766 (0.1108)	0.9873 (0.0065)	2.1857 (0.8518)	1.1620 (0.2234)	0.9863 (0.0083)
	Penalized	0.1641 (0.0725)	0.3167 (0.0693)	0.9954 (0.0015)	0.6570 (0.2914)	0.6338 (0.1389)	0.9954 (0.0015)	2.6186 (1.1566)	1.2657 (0.2761)	0.9955 (0.0016)
	CVXR	0.1604 (0.0714)	0.3132 (0.0687)	0.8553 (0.1115)	0.6425 (0.2866)	0.6266 (0.1376)	0.9018 (0.1509)	2.5624 (1.1402)	1.2518 (0.2738)	0.9068 (0.1731)
	nnls	0.4959 (0.1366)	0.5582 (0.0789)	0.9600 (0.0007)	2.2363 (1.1091)	1.1768 (0.2152)	0.9587 (0.0009)	10.977 (6.3035)	2.5831 (0.6098)	0.9569 (0.0011)
	nnFSG	0.0450 (0.0349)	0.1576 (0.0631)	1.0000 (0.0000)	0.1854 (0.1173)	0.3282 (0.1026)	0.9999 (0.0002)	0.8336 (0.5771)	0.6972 (0.2169)	0.9997 (0.0003)

The average values of MSE, MAE, FTP as well as their standard deviations (in parenthesis) are based on 500 simulations

We thus assume that the variability of measurement error is small. Let $\sigma = 0.2, 0.3$. Again, put $\lambda_2 = \lambda_1, \lambda_3 = 10$. λ_1 and τ are selected via fivefold cross-validation. We compare our proposed method with the others that are estimable by using the R package `glmnet`, `penalized`, `CVXR`, `nnls`.

Hu et al. (2015) computed sensitivity (Sen) to measure the probability of an important variable associated with a non-zero coefficient being selected, and specificity (Spe) to measure the probability of an unimportant variable associated with a zero coefficient not being selected. The larger the values of both Spe and Sen, the better the performance of the method. A perfect situation would be described as 100% sensitivity, meaning all important sites were correctly identified, and 100% specificity, meaning all unimportant sites were not selected.

The simulation results are shown in Tables 2 and 3. From Table 2, we observe that our proposed method performs best for all scenarios in terms of MSE and MAE. With the purpose of locating important AA sites where the associated coefficients are non-zero, we are more interested in Spe, Sen and FTP (see Table 3), measuring effectiveness of identifying main sites or features. Moreover, one may be interested in identifying those important AA sites that have equal importance. Considering the simulation settings of $\boldsymbol{\beta}^0$, we thus provide the GTP and estimated number of grouping $|\mathcal{G}|$ in Table 3 as well. In terms of Spe, Sen, FTP and GTP, nnFSG achieves the largest values almost all cases except Case III when $\sigma = 0.2$, where the largest values of Spe, FTP and GTP are obtained by our method, albeit the Sen is slightly smaller than others. In reality, there is usually a trade-off between Spe

Table 2 Comparison of glmnet, penalized, CVXR, nnls and nnFSG assessed on synthetic malaria vaccine data under the settings with $n = 500, p = 3000, \sigma = 0.2, 0.3$

Case	Methods	$\sigma = 0.2$		$\sigma = 0.3$	
		MSE	MAE	MSE	MAE
I	glmnet	0.2661 (0.0317)	0.4099 (0.0246)	0.2457 (0.0317)	0.3939 (0.0255)
	Penalized	0.0482 (0.0205)	0.1817 (0.0441)	0.1085 (0.0462)	0.2726 (0.0662)
	CVXR	0.0166 (0.0034)	0.1027 (0.0103)	0.0375 (0.0076)	0.1542 (0.0155)
	nnls	0.0196 (0.0038)	0.1117 (0.0106)	0.0441 (0.0085)	0.1675 (0.0158)
	nnFSG	0.0001 (0.0001)	0.0069 (0.0052)	0.0002 (0.0002)	0.0099 (0.0074)
II	glmnet	0.0137 (0.0028)	0.0932 (0.0093)	0.0316 (0.0065)	0.1417 (0.0142)
	Penalized	0.0482 (0.0206)	0.1817 (0.0442)	0.1085 (0.0462)	0.2726 (0.0663)
	CVXR	0.0168 (0.0034)	0.1031 (0.0104)	0.0379 (0.0078)	0.1549 (0.0157)
	nnls	0.0196 (0.0038)	0.1117 (0.0106)	0.0441 (0.0085)	0.1675 (0.0158)
	nnFSG	0.0002 (0.0002)	0.0092 (0.0049)	0.0004 (0.0004)	0.0139 (0.0076)
III	glmnet	0.0147 (0.0029)	0.0967 (0.0095)	0.0331 (0.0065)	0.1450 (0.0142)
	Penalized	0.0485 (0.0206)	0.1823 (0.0441)	0.1100 (0.0450)	0.2754 (0.0645)
	CVXR	0.0174 (0.0035)	0.1046 (0.0105)	0.0385 (0.0078)	0.1556 (0.0156)
	nnls	0.0198 (0.0037)	0.1121 (0.0104)	0.0446 (0.0083)	0.1684 (0.0156)
	nnFSG	0.0004 (0.0008)	0.0131 (0.0089)	0.0022 (0.0044)	0.0291 (0.0234)

The averaged MSE and MAE as well as their standard deviations (in parenthesis) are based on 500 simulations

Table 3 Comparison of glmnet, penalized, CVXR, nnls, and nnFSG assessed on synthetic malaria vaccine data under the settings with $n = 500, p = 3000, \sigma = 0.2, 0.3$

Case	Methods	$\sigma = 0.2$					$\sigma = 0.3$				
		Sen	Spe	FTP	GTP	$\lvert\mathcal{G}\rvert$	Sen	Spe	FTP	GTP	$\lvert\mathcal{G}\rvert$
I	glmnet	1.0000 (0.0000)	0.9826 (0.0020)	0.9827 (0.0020)	0.9875 (0.0010)	74.55 (5.5122)	1.0000 (0.0000)	0.9822 (0.0022)	0.9824 (0.0022)	0.9873 (0.0011)	75.40 (5.9169)
	Penalized	1.0000 (0.0000)	0.9802 0.0024)	0.9804 (0.0024)	0.9864 (0.0012)	75.13 (5.8575)	1.0000 (0.0000)	0.9802 (0.0024)	0.9803 (0.0024)	0.9863 (0.0012)	77.84 (6.4350)
	CVXR	1.0000 (0.0000)	0.2319 (0.0688)	0.2380 (0.0683)	0.6152 (0.0341)	82.45 (5.3307)	1.0000 (0.0000)	0.1963 (0.0660)	0.2027 (0.0655)	0.5975 (0.0327)	89.26 (6.2682)
	nnls	1.0000 (0.0000)	0.9528 (0.0042)	0.9531 (0.0042)	0.9727 (0.0021)	133.68 (8.2237)	1.0000 (0.0000)	0.9527 (0.0042)	0.9531 (0.0042)	0.9727 (0.0021)	142.88 (9.2281)
	nnFSG	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	2.01 (0.0996)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	2.01 (0.0996)
II	glmnet	1.0000 (0.0000)	0.9649 (0.0042)	0.9652 (0.0042)	0.9860 (0.0014)	103.55 (7.8472)	1.0000 (0.0000)	0.9653 (0.0045)	0.9656 (0.0044)	0.9861 (0.0015)	110.53 (9.7439)
	Penalized	1.0000 (0.0000)	0.9802 (0.0024)	0.9804 (0.0024)	0.9910 (0.0008)	75.26 (5.9022)	1.0000 (0.0000)	0.9802 (0.0024)	0.9803 (0.0024)	0.9910 (0.0008)	77.96 (6.3804)
	CVXR	1.0000 (0.0000)	0.4865 (0.0975)	0.4906 (0.0968)	0.8278 (0.0323)	86.72 (6.3307)	1.0000 (0.0000)	0.4341 (0.0796)	0.4386 (0.0789)	0.8104 (0.0263)	93.51 (7.2051)
	nnls	1.0000 (0.0000)	0.9528 (0.0042)	0.9531 (0.0042)	0.9819 (0.0014)	133.82 (8.2232)	1.0000 (0.0000)	0.9527 (0.0042)	0.9531 (0.0042)	0.9819 (0.0014)	142.97 (9.2042)
	nnFSG	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	$0.999\dot{9}^{a}$ (0.00003)	3.0200 (0.1401)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	$0.999\dot{9}^{a}$ (0.00003)	3.04 (0.1962)

Table 3 (continued)

Case	Methods	$\sigma = 0.2$					$\sigma = 0.3$				
		Sen	Spe	FTP	GTP	$\lvert\mathcal{G}\rvert$	Sen	Spe	FTP	GTP	$\lvert\mathcal{G}\rvert$
III	glmnet	1.0000 (0.0000)	0.9660 (0.0044)	0.9663 (0.0043)	0.9898 (0.0011)	103.64 (8.5684)	1.0000 (0.0000)	0.9660 (0.0042)	0.9663 (0.0042)	0.9898 (0.0010)	110.35 (9.5269)
	Penalized	1.0000 (0.0000)	0.9802 (0.0024)	0.9804 (0.0024)	0.9934 (0.0006)	75.30 (5.8480)	1.0000 (0.0000)	0.9802 (0.0024)	0.9804 (0.0023)	0.9934 (0.0006)	77.85 (6.1707)
	CVXR	1.0000 (0.0000)	0.8928 (0.0422)	0.8936 (0.0419)	0.9717 (0.0105)	91.24 (6.4550)	1.0000 (0.0000)	0.8887 (0.0466)	0.8896 (0.0463)	0.9707 (0.0116)	97.52 (6.9624)
	nnls	1.0000 (0.0000)	0.9527 (0.0042)	0.9531 (0.0042)	0.9865 (0.0011)	133.99 (8.3155)	1.0000 (0.0000)	0.9525 (0.0041)	0.9529 (0.0040)	0.9865 (0.0010)	143.70 (9.0355)
	nnFSG	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	0.999$\dot{9}$[a] (0.00005)	4.05 (0.2182)	0.9992 (0.0058)	0.999$\dot{9}$[a] (0.00003)	0.999$\dot{9}$[a] (0.0001)	0.9993 (0.0044)	4.25 (0.5040)

The averaged Sen, Spe, FTP, GTP and $\lvert\mathcal{G}\rvert$ as well as their standard deviations (in parenthesis) are based on 500 simulations

[a] $\dot{9}$ means that the fifth decimal is 9

and Sen, informedness (Spe + Sen − 1), the magnitude of which measures the probability of an informed decision. Considering informedness, nnFSG outperforms the other methods. The outperformance of nnFSG is also demonstrated by the fact that the estimated grouping number $|\mathcal{G}|$ is similar to $|\mathcal{G}^0|$ for each scenario.

4.5 Protein mass spectrometry data

Mass spectrometry (MS) analysis has become a key tool for extracting reliable proteomic features (peptides) from complex biological mixtures (Renard et al. 2008), which is a fundamental step in the automated analysis of proteomic MS experiments. A peptide produces a signal at multiple mass positions, which manifests as a series of regularly spaced peaks. For more details on MS analysis, one can refer to Renard et al. (2008), Slawski and Hein (2010), and Slawski et al. (2012). Figure 2 shows a protein mass spectrum of Myoglobine in the m/z 800–2500 range, 118,464 (m/z, intensity) pairs in total. The m/z range of 800–834 is shown in greater detail. The peptides whose intensities differ drastically occur in different m/z-regions. The data set was kindly provided by B. Gregorius and A. Tholey, Department of Experimental Medicine, Working Group for Systematic Proteomics, Christian-Albrechts-Universitaet zu Kiel, and is avaialable in the R package `IPPD` (Slawski et al. 2012).

The peptides extraction problem is to identify those m/z-positions where a peptide is located. This can be recast as a sparse recovery problem. Renard et al. (2008), Slawski and Hein (2010), and Slawski et al. (2012) proposed template matching-based methods

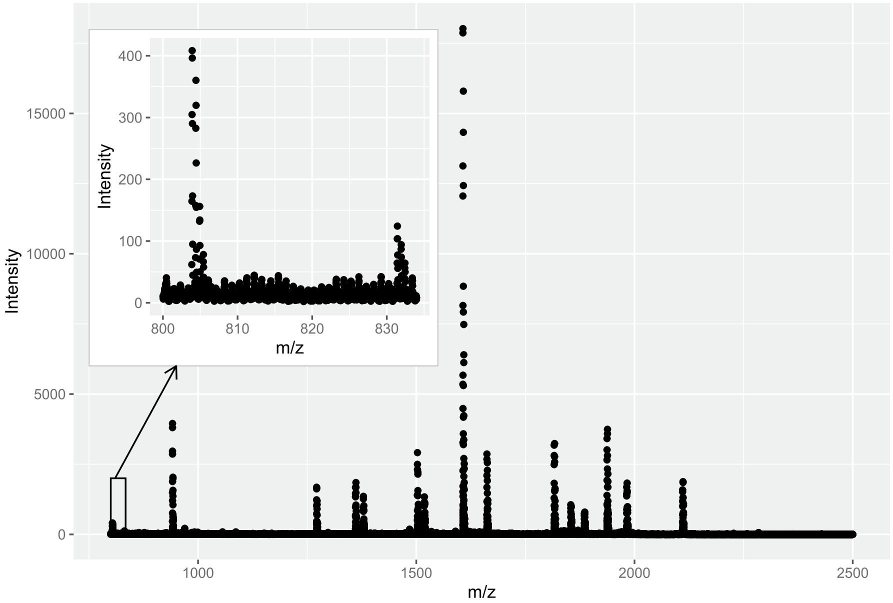

Fig. 2 Raw protein mass spectrum of Myoglobine in the m/z 800–2500 range. The left upper panel zooms at the m/z range of 800–834

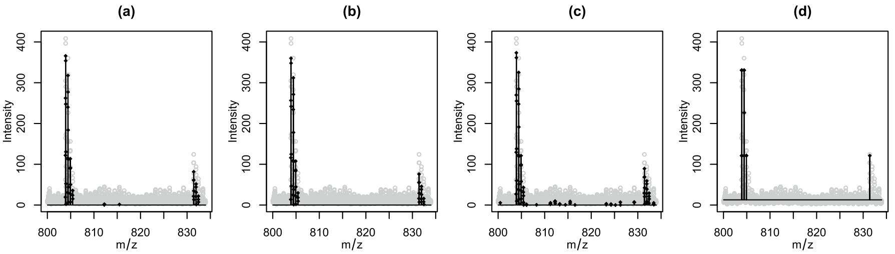

Fig. 3 The estimates of β_i $(y_i, i = 1, \ldots, n)$ at the m/z range of 800–834 via **a** glmnet, **b** penalized **c** CVXR, **d** nnFSG. The grey points represent the MS data, and the solid black line represents the estimated weights $\hat{\boldsymbol{\beta}}$ from the proposed method. The black points describe the weights of these m/z sites that are extracted

Table 4 Prediction Performance: MSE and MAE for the methods of glmnet, penalized, CVXR, nnls and nnFSG

Measures	glmnet	Penalized	CVXR	nnFSG
MAE	13.2411	13.3579	13.0176	6.4912
MSE	245.3436	255.9443	228.152	110.0166

to solve the problem. Motivated by Tibshirani and Wang (2008), we can also regard the peptides extraction to be a 'hot spot' detection problem. The model's setup for the protein MS data is $p = n = 118{,}464$, and $X = I_p$, that is, $y_i = \beta_i + \epsilon_i, i = 1, \ldots, n$. Given the non-negativity of y_i (intensity), it is reasonable to impose non-negative constraints on β_i, the weight of the i-th m/z-site. Since the design matrix is identity and $n = p$, nnls thus doesn't work for this case. The estimate of β_i is exactly equal to y_i, $i = 1, \ldots, n$. We compare the performance of the methods: glmnet, penalized, CVXR, and nnFSG.

Simultaneously estimating the weights of all m/z sites for the MS data in Fig. 2 is difficult since $p = 118{,}464$ is ultra-high. We thus divide the data into consecutive blocks, which has no effect on estimation. Herein, we choose m/z-sites in the range of 800–834 for analysis, giving a total of 2009 points. The performance of those methods on the MS data is illustrated in Fig. 3 and Table 4. The proposed method nnFSG puts the same weights at those sites where the amplifications are not significant, rather than zeros obtained via the other three methods. We consider the sites with identical weights as one base group (see the horizontal black solid line in Fig. 3d). Sites that are not in the base group can be regarded as peptides, which are extracted and marked with black points in Fig. 3. Those methods successfully identify the amplification. The R package `glmnet`, `penalized` and nnFSG perform well on peptides extraction, while in terms of prediction errors, nnFSG outperforms the others.

5 Conclusions

We have proposed a method for high-dimensional regression problems where the regression coefficients are sign-constrained, sparse, or even containing a structure with homogeneous subgroups. We aim to identify the underlying optimal grouping and obtain the optimal estimator that satisfies the sign constraints. Specifically, we formulate a regularized minimization problem with a non-convex, but difference of convex, objective function. Using the difference of convex programming, a subproblem at each iteration is reformulated as a constrained minimization problem with a convex objective, which is solved applying augmented Lagrange and coordinated decent methods. The theoretical results show that the developed nnFSG method recovers the oracle estimate consistently, and the MSE are also bounded. In addition, the numerical studies show that the proposed nnFSG outperforms some existing methods in terms of prediction accuracy, feature selection and grouping.

Acknowledgements This work is supported by Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-05720). Qin also gratefully acknowledges the financial support from the China Scholarship Council (Grant No.201506180073).

Appendix

Proof of Lemma 1 Since $\hat{\boldsymbol{\alpha}}^{\text{ols}} = (\hat{\alpha}_1^{\text{ols}}, \dots, \hat{\alpha}_{K^0}^{\text{ols}})^\top = (Z^\top_{\mathcal{G}_0^{0c}} Z_{\mathcal{G}_0^{0c}})^{-1} Z^\top_{\mathcal{G}_0^{0c}} \boldsymbol{y} = \boldsymbol{\alpha}^0 + (Z^\top_{\mathcal{G}_0^{0c}} Z_{\mathcal{G}_0^{0c}})^{-1} Z^\top_{\mathcal{G}_0^{0c}} \boldsymbol{\epsilon}$, $\hat{\boldsymbol{\alpha}}^{\text{ols}} \sim N\left(\boldsymbol{\alpha}^0, \sigma^2 (Z^\top_{\mathcal{G}_0^{0c}} Z_{\mathcal{G}_0^{0c}})^{-1}\right)$, namely,

$$\hat{\alpha}_k^{\text{ols}} - \alpha_k^0 \sim N\left(0, \sigma^2 (Z^\top_{\mathcal{G}_0^{0c}} Z_{\mathcal{G}_0^{0c}})^{-1}_{kk}\right), k = 1, \dots, K^0,$$

where $(Z^\top_{\mathcal{G}_0^{0c}} Z_{\mathcal{G}_0^{0c}})^{-1}_{kk}$ denotes the k-th diagonal element of matrix $(Z^\top_{\mathcal{G}_0^{0c}} Z_{\mathcal{G}_0^{0c}})^{-1}$.

By the assumption (A2), it yields that the variance of $\hat{\alpha}_k^{\text{ols}}$ is bounded from above by $\sigma^2/(nc_0)$ for all $k = 1, \dots, K^0$. In view of the assumption (A3), $\min_{1 \le k \le K^0} \alpha_k^0 = \min_{j \in \mathcal{G}_0^{0c}} \beta_j^0 > c_n$, where $c_n = [2\sigma^2 \log\{2nK^0/(2\pi)^{1/2}\}/(nc_0)]^{1/2}$. Similar to Meinshausen (2013), by Bonferroni's inequality, we thus have

$$\|\hat{\boldsymbol{\alpha}}^{\text{ols}} - \boldsymbol{\alpha}^0\|_\infty \le c_n,$$

with probability at least

$$1 - 2K^0\left\{1 - \Phi\left(c_n (nc_0)^{1/2}/\sigma\right)\right\} = 1 - 2K^0\left\{1 - \Phi\left([2\log\{2nK^0/(2\pi)^{1/2}\}]^{1/2}\right)\right\}.$$

It implies that with probability at least $1 - 2K^0\left\{1 - \Phi\left([2\log\{2nK^0/(2\pi)^{1/2}\}]^{1/2}\right)\right\}$, $\min_{1 \le k \le K^0} \hat{\alpha}_k^{\text{ols}} > 0$, and thus $\hat{\boldsymbol{\alpha}}^{ora} = \hat{\boldsymbol{\alpha}}^{\text{ols}}$, $\hat{\boldsymbol{\beta}}^{\text{ora}} = \hat{\boldsymbol{\beta}}^{\text{ols}}$. That is,

$$\text{pr}\left(\hat{\boldsymbol{\beta}}^{\text{ora}} \ne \hat{\boldsymbol{\beta}}^{\text{ols}}\right) \le 2K^0\left\{1 - \Phi\left([2\log\{2nK^0/(2\pi)^{1/2}\}]^{1/2}\right)\right\}.$$

Since $1 - \Phi(x) \le (2\pi)^{-1/2} x^{-1} \exp(-x^2/2)$ for any $x > 0$, it follows that

$$\text{pr}\left(\hat{\boldsymbol{\beta}}^{\text{ora}} \ne \hat{\boldsymbol{\beta}}^{\text{ols}}\right) \le \frac{1}{n} \frac{1}{[2\log\{2nK^0/(2\pi)^{1/2}\}]^{1/2}} = O\left(\frac{1}{n(\log n)^{1/2}}\right).$$

□

Proof of Theorem 1 Let $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_K)$ be a grouping of the constrained problem in Sect. 2, satisfying that $0 \leq \hat{\beta}_j^{\text{cons}} \leq \tau$ if $j \in \mathcal{G}_0$, $|\hat{\beta}_j^{\text{cons}} - \hat{\beta}_{j'}^{\text{cons}}| > \tau$ if $j \in \mathcal{G}_k$, $j' \in \mathcal{G}_{k'}$, $j = 1, \ldots, p; 1 \leq k \neq k' \leq K$.

If $\mathcal{G} = \mathcal{G}^0$, then $|\mathcal{G}_0^c| = s_1^0$. By the first constraint $\sum_{j=1}^p \min\left\{\frac{|\beta_j|}{\tau}, 1\right\} \leq s_1$, $\sum_{j \in \mathcal{G}_0} \hat{\beta}_j^{\text{cons}}/\tau + s_1^0 \leq s_1^0$, which implies that $\hat{\beta}_j^{\text{cons}} = 0$, $j \in \mathcal{G}_0$. By the second constraint $\sum_{(j,j') \in \varepsilon} \min\left\{\frac{|\beta_j - \beta_{j'}|}{\tau}, 1\right\} \leq s_2$, similarly, we obtain that $\hat{\beta}_j^{\text{cons}} = \hat{\beta}_{j'}^{\text{cons}}$, $j, j' \in \mathcal{G}_k = \mathcal{G}_k^0$, $(j, j') \in \varepsilon$, $k = 1, \ldots, K$. Thus, $\hat{\boldsymbol{\beta}}^{\text{cons}} = \hat{\boldsymbol{\beta}}^{\text{ora}}$ if $\mathcal{G} = \mathcal{G}^0$, which, together with the fact that $\text{pr}(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ora}}) = \text{pr}(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ora}}, \mathcal{G} \neq \mathcal{G}^0) + \text{pr}(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ora}}, \mathcal{G} = \mathcal{G}^0)$, yields that

$$\text{pr}\left(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ora}}\right) = \text{pr}\left(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ora}}, \mathcal{G} \neq \mathcal{G}^0\right). \tag{20}$$

Denote $\bar{S}(\boldsymbol{\beta}) = 2^{-1}\|Y - X\boldsymbol{\beta}\|^2$. In view that $\text{pr}(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ora}}, \mathcal{G} \neq \mathcal{G}^0) = \text{pr}(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ora}}, \hat{\boldsymbol{\beta}}^{\text{ora}} = \hat{\boldsymbol{\beta}}^{\text{ols}}, \mathcal{G} \neq \mathcal{G}^0) + \text{pr}(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ora}}, \hat{\boldsymbol{\beta}}^{\text{ora}} \neq \hat{\boldsymbol{\beta}}^{\text{ols}}, \mathcal{G} \neq \mathcal{G}^0)$, (20) thus becomes

$$\begin{aligned}
&\text{pr}\left(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ora}}\right) \\
&\leq \text{pr}\left(\bar{S}\left(\hat{\boldsymbol{\beta}}^{\text{cons}}\right) - \bar{S}(\hat{\boldsymbol{\beta}}^{\text{ora}}) \leq 0, \hat{\boldsymbol{\beta}}^{\text{ora}} = \hat{\boldsymbol{\beta}}^{\text{ols}}, \mathcal{G} \neq \mathcal{G}^0\right) + \text{pr}\left(\hat{\boldsymbol{\beta}}^{\text{ora}} \neq \hat{\boldsymbol{\beta}}^{\text{ols}}\right).
\end{aligned} \tag{21}$$

The second term in (21) has already provided in Lemma 2.1. Next, we work on the first term in (21), and denote it by $\boldsymbol{\Gamma}$.

Consider the case where $\hat{\boldsymbol{\beta}}^{\text{ora}} = \hat{\boldsymbol{\beta}}^{\text{ols}}$ and $\mathcal{G} \neq \mathcal{G}^0$. Define $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_p)^\top$, satisfying

$$\bar{\beta}_j = \begin{cases} \frac{\sum_{j' \in \mathcal{G}_k} \hat{\beta}_{j'}^{\text{cons}}}{|\mathcal{G}_k|}, & \text{if } j \in \mathcal{G}_k, k = 1, \ldots, K, \\ 0, & \text{if } j \in \mathcal{G}_0. \end{cases}$$

It follows that $|\bar{\beta}_j - \hat{\beta}_j^{\text{cons}}| \leq \tau$, $\|\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{\text{cons}}\|^2 \leq \tau^2 p$, and thus

$$\|X(\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{\text{cons}})\|^2 \leq \lambda_{\max}(X^\top X)\tau^2 p. \tag{22}$$

Note that

$$\|Y - X\bar{\boldsymbol{\beta}}\|^2 \geq \|Y - P_{Z_{\mathcal{G}_0^c}} Y\|^2 = \|(I - P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0 + (I - P_{Z_{\mathcal{G}_0^c}})\boldsymbol{\epsilon}\|^2. \tag{23}$$

For any vector $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$ and $a > 0$, it holds that $\|\boldsymbol{u} + \boldsymbol{v}\|^2 \geq a^{-1}(a-1)\|\boldsymbol{u}\|^2 - (a-1)\|\boldsymbol{v}\|^2$ (Shen et al. 2012a). We thus have

$$\bar{S}\left(\hat{\boldsymbol{\beta}}^{\text{cons}}\right) = \frac{1}{2}\left\|Y - X\hat{\boldsymbol{\beta}}^{\text{cons}}\right\|^2 \geq \frac{a-1}{2a}\|Y - X\bar{\boldsymbol{\beta}}\|^2 - \frac{a-1}{2}\left\|X(\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{\text{cons}})\right\|^2. \tag{24}$$

By substituting (22)–(23) into (24) and together with $\bar{S}\left(\hat{\boldsymbol{\beta}}^{\text{ols}}\right) = 2^{-1}\|(I - P_{Z_{\mathcal{G}_0^{0c}}})\boldsymbol{\epsilon}\|^2 \leq 2^{-1}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$, we obtain that, for any $a > 1$,

$$2a\left\{\bar{S}\left(\hat{\boldsymbol{\beta}}^{\text{cons}}\right)-\bar{S}(\hat{\boldsymbol{\beta}}^{\text{ora}})\right\}=2a\left\{\bar{S}\left(\hat{\boldsymbol{\beta}}^{\text{cons}}\right)-\bar{S}\left(\hat{\boldsymbol{\beta}}^{\text{ols}}\right)\right\}\geq -L_1-L_2+L_3,$$

where $L_1=\{\boldsymbol{\epsilon}-(a-1)(I-P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\}^\top(I-P_{Z_{\mathcal{G}_0^c}})\{\boldsymbol{\epsilon}-(a-1)(I-P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\}$, and $L_1\sigma^{-2}$ follows noncentral Chi-squared distribution $\chi^2_{k,\varLambda}$ with degrees of freedom $k=\max\{n-K(\mathcal{G}_0^c),0\}$, and noncentral parameter $\varLambda=(a-1)^2\|(I-P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\|^2/\sigma^2$; $L_2=a\boldsymbol{\epsilon}^\top P_{Z_{\mathcal{G}_0^c}}\boldsymbol{\epsilon}$ is independent of L_1, and $a^{-1}\sigma^{-2}L_2$ follows Chi-squared distribution χ^2_κ with degrees of freedom $\kappa=K(\mathcal{G}_0^c)$; $L_3=a(a-1)\|(I-P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\|^2-a(a-1)$ $\lambda_{\max}(X^\top X)\tau^2p\leq 2^{-1}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon},.$ Note that, by the definition of $C_{\min}$, $\|(I-P_{Z_{\mathcal{G}_0^c}})X\boldsymbol{\beta}^0\|^2\geq nC_{\min}$.

For $\varGamma$, by Markov inequality and moment-generating function of Chi-squared distribution, it holds that, for any $0<t<1/(2a)$ and $1-2at<1-2t<1$ $(a>1)$, by Shen et al. (2012a),

$$\varGamma\leq\sum_{i=1}^{s_1^0}\sum_{j=0}^{i}\binom{p-s_1^0}{j}\binom{s_1^0}{s_1^0-i}T_il_1^*l_2^{*K_i^*/2}\frac{1}{(1-2t)^{n/2}},$$

where $l_1^*=\exp\left\{\frac{(a-1)\log p}{4n}-n\frac{t(a-1)iC_{\min}}{\sigma^2}\frac{1-2at}{1-2t}\right\}$, $l_2^*=(1-2t)/(1-2at)$, $K_i^*=\max_{\{\mathcal{G}\in\mathcal{T},|\mathcal{G}_0\setminus\mathcal{G}_0^0|=i\}}K(\mathcal{G}_0^c)$. Note that the last inequality holds true because

$$\frac{t}{\sigma^2}a(a-1)\lambda_{\max}(X^\top X)p\tau^2\leq\frac{2ta(a-1)\log p}{4n}\leq\frac{(a-1)\log p}{4n}$$

for any $\tau\leq\sigma[\log p/\{2np\lambda_{\max}(X^\top X)\}]^{1/2}$. We choose $a=4+n/4$, $t=4^{-1}(a-1)^{-1}$, and define $b=(1-2t)/(1-2at)$. Then $b=(2a-3)/(a-2)<5/2$, and $(a-1)/(4n)\leq 1$. Since $-\log(1-x)\leq x(1-x)^{-1}$ for $0<x<1$, and $0<2t=2^{-1}(a-1)^{-1}<1$, it follows that

$$-\frac{n}{2}\log(1-2t)\leq\frac{n}{2}\frac{1/\{2(a-1)\}}{1-1/\{2(a-1)\}}\leq\frac{n}{2}\frac{1}{2(4+n/4)-3}\leq 1,$$

which jointly with the facts

$$\binom{s_1^0}{s_1^0-i}\leq(s_1^0)^i,\sum_{j=0}^{i}\binom{p-s_1^0}{j}\leq(p-s_i^0)^i\quad\text{and}\quad(p-s_1^0)s_1^0\leq p^2/4$$

yields that

$$\begin{aligned}\varGamma&\leq\sum_{i=1}^{s_1^0}\left(\frac{p^2}{4}\right)^iT_i\exp\left\{\frac{(a-1)\log p}{4n}-n\frac{iC_{\min}}{4b\sigma^2}\right\}b^{K_i^*/2}\frac{1}{(1-2t)^{n/2}}\\&\leq\exp(1)\sum_{i=1}^{s_1^0}\exp\left\{-i\frac{n}{10\sigma^2}\left(C_{\min}-\frac{10\sigma^2}{n}(3\log p+\bar{T}+\bar{K}/2)\right)\right\}.\end{aligned}\tag{25}$$

Since $(1-z)^{-1}=\sum_{i=0}^{\infty}z^i$ for $|z|<1$, we thus obtain that, for $x<0$,

$$\sum_{i=1}^{s_1^0} \exp(ix) \le -1 + \frac{1}{1-\exp(x)} = \frac{\exp(x)}{1-\exp(x)}.$$

We take $x = -10^{-1}\sigma^{-2}n\{C_{\min} - 10\sigma^2 n^{-1}(3\log p + \bar{T} + \bar{K}/2)\}$ if $C_{\min} > 10\sigma^2 n^{-1}(3\log p + \bar{T} + \bar{K}/2)$. Together with $\Gamma \le 1$, (25) becomes

$$\Gamma \le \{\exp(1)+1\} \exp\left[-\frac{n}{10\sigma^2}\left\{C_{\min} - \frac{10\sigma^2}{n}(3\log p + \bar{T} + \bar{K}/2)\right\}\right]. \tag{26}$$

Similarly, we can show that (26) still holds for $C_{\min} \le 10\sigma^2 n^{-1}(3\log p + \bar{T} + \bar{K}/2)$. By Lemma 2.1 and (26), (21) becomes

$$\begin{aligned} &\mathrm{pr}\left(\hat{\boldsymbol{\beta}}^{\mathrm{cons}} \ne \hat{\boldsymbol{\beta}}^{\mathrm{ora}}\right) \\ &\le \{\exp(1)+1\} \exp\left[-\frac{n}{10\sigma^2}\left\{C_{\min} - \frac{10\sigma^2}{n}(3\log p + \bar{T} + \bar{K}/2)\right\}\right] + \frac{c}{n(\log n)^{1/2}}. \end{aligned} \tag{27}$$

1. If $C_{\min} \ge 10\sigma^2 n^{-1}\left(\log n + 2^{-1}\log\log n + 3\log p + \bar{T} + \bar{K}/2\right)$, by (27),

$$\mathrm{pr}\left(\hat{\boldsymbol{\beta}}^{\mathrm{cons}} \ne \hat{\boldsymbol{\beta}}^{\mathrm{ora}}\right) = O\left(\frac{1}{n(\log n)^{1/2}}\right).$$

2. We denote $T_1 = n^{-1}E(\|X\hat{\boldsymbol{\beta}}^{\mathrm{cons}} - X\boldsymbol{\beta}^0\|^2 I_{\{G\}})$, and $T_2 = n^{-1}E(\|X\hat{\boldsymbol{\beta}}^{\mathrm{cons}} - X\boldsymbol{\beta}^0\|^2 I_{\{G^c\}})$, where $G = \{n^{-1}\|X\hat{\boldsymbol{\beta}}^{\mathrm{cons}} - X\boldsymbol{\beta}^0\|^2 \ge 25\sigma^2\}$. It is easy to see that

$$\frac{1}{n}E\left\|X\hat{\boldsymbol{\beta}}^{\mathrm{cons}} - X\boldsymbol{\beta}^0\right\|^2 = T_1 + T_2.$$

Now, we work on T_1. By the definition, $T_1 = \int_{25\sigma^2}^{\infty} \mathrm{pr}(n^{-1}\|X\hat{\boldsymbol{\beta}}^{\mathrm{cons}} - X\boldsymbol{\beta}^0\|^2 \ge x)\mathrm{d}x$ $+25\sigma^2\mathrm{pr}(n^{-1}\|X\hat{\boldsymbol{\beta}}^{\mathrm{cons}} - X\boldsymbol{\beta}^0\|^2 \ge 25\sigma^2)$. For the first term of T_1,

$$\begin{aligned} &\int_{25\sigma^2}^{\infty} \mathrm{pr}\left(n^{-1}\|X\hat{\boldsymbol{\beta}}^{\mathrm{cons}} - X\boldsymbol{\beta}^0\|^2 \ge x\right)\mathrm{d}x \\ &\quad \le \int_{25\sigma^2}^{\infty} \mathrm{pr}\left(4n^{-1}\|\boldsymbol{\epsilon}\|^2 \ge x\right)\mathrm{d}x \\ &\quad \le \int_{25\sigma^2}^{\infty} E\left\{\exp\left(\frac{\|\boldsymbol{\epsilon}\|^2}{3\sigma^2}\right)\right\}\exp\left(-\frac{nx}{12\sigma^2}\right)\mathrm{d}x \\ &\quad = \int_{25\sigma^2}^{\infty} \exp\left[-\frac{n}{12\sigma^2}\{x - 6(\log 3)\sigma^2\}\right]\mathrm{d}x \\ &\quad < \int_{25\sigma^2}^{\infty} \exp\left\{-\frac{n}{12\sigma^2}(x - 24\sigma^2)\right\}\mathrm{d}x \\ &\quad = \frac{12\sigma^2}{n}\exp\left(-\frac{n}{12}\right) = o\left(\frac{K^0\sigma^2}{n}\right). \end{aligned} \tag{28}$$

Since $\|X\hat{\boldsymbol{\beta}}^{\text{cons}} - X\boldsymbol{\beta}^0\|^2 \le 2(\|Y - X\hat{\boldsymbol{\beta}}^{\text{cons}}\|^2 + \|Y - X\boldsymbol{\beta}^0\|^2) \le 4\|Y - X\boldsymbol{\beta}^0\|^2 = 4\|\boldsymbol{\epsilon}\|^2$, the first '≤' follows. The second '≤' is obtained by the Markov inequality. In view of the moment generating function for Chi-squared distribution, the first '=' holds. For the second term of T_1,

$$25\sigma^2 \text{pr}(n^{-1}\|X\hat{\boldsymbol{\beta}}^{\text{cons}} - X\boldsymbol{\beta}^0\|^2 \ge 25\sigma^2) \le 25\sigma^2 \exp(-n/12) = o\left(\frac{K^0\sigma^2}{n}\right). \tag{29}$$

By (28) and (29), we thus have $T_1 = o(K^0\sigma^2/n)$.
On the other hand,

$$\begin{aligned} T_2 &= E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{\text{cons}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G^c\}} I_{\{\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ols}}\}}\right) \\ &\quad + E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{\text{cons}} - X\boldsymbol{\beta}^0\right\|^2 \left(1 - I_{\{G\}}\right) I_{\{\hat{\boldsymbol{\beta}}^{\text{cons}} = \hat{\boldsymbol{\beta}}^{\text{ols}}\}}\right). \end{aligned} \tag{30}$$

For the first term in (30), it follows that

$$\begin{aligned} &E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{\text{cons}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G^c\}} I_{\{\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ols}}\}}\right) \\ &\le 25\sigma^2 \text{pr}\left(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ols}}\right) \\ &\le 25\sigma^2 \text{pr}\left(\hat{\boldsymbol{\beta}}^{\text{cons}} \neq \hat{\boldsymbol{\beta}}^{\text{ora}}\right) + 25\sigma^2 \text{pr}\left(\hat{\boldsymbol{\beta}}^{\text{ora}} \neq \hat{\boldsymbol{\beta}}^{\text{ols}}\right) \\ &\le \frac{100\sigma^2}{n(\log n)^{1/2}} + \frac{50\sigma^2 c}{n(\log n)^{1/2}} = o\left(\frac{K^0\sigma^2}{n}\right). \end{aligned} \tag{31}$$

For the second term in (30),

$$\begin{aligned} &E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{\text{cons}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G\}} I_{\{\hat{\boldsymbol{\beta}}^{\text{cons}} = \hat{\boldsymbol{\beta}}^{\text{ols}}\}}\right) \\ &\quad \le E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{\text{cons}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G\}}\right) = o\left(\frac{K^0\sigma^2}{n}\right), \end{aligned} \tag{32}$$

and

$$\begin{aligned} &E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{\text{cons}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{\hat{\boldsymbol{\beta}}^{\text{cons}} = \hat{\boldsymbol{\beta}}^{\text{ols}}\}}\right) \\ &= \frac{1}{n}E\left(\left\|X\hat{\boldsymbol{\beta}}^{\text{ols}} - X\boldsymbol{\beta}^0\right\|^2\right) = \frac{1}{n}E\left(\left\|P_{Z_{\mathcal{G}_0^{0c}}}\boldsymbol{\epsilon}\right\|^2\right) = \frac{K^0\sigma^2}{n}. \end{aligned} \tag{33}$$

By (30)–(33), $T_2 = n^{-1}K^0\sigma^2(1 + o(1))$. Therefore,

$$\frac{1}{n}E\left(\left\|X\hat{\boldsymbol{\beta}}^{\text{cons}} - X\boldsymbol{\beta}^0\right\|^2\right) = T_1 + T_2 = \frac{K^0\sigma^2}{n}(1 + o(1)).$$

□

Proof of Theorem 3 This proof mimics the proof of Theorem 1 in (Shen et al. 2012a). We thus omit the details. □

Proof of Theorem 4 By Sect. 3, there exists a finite m^* such that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(m^*)}$. Denote the grouping of $\hat{\boldsymbol{\beta}}$ by $\mathcal{G} = (\mathcal{G}_0, \mathcal{G}_1, \dots, \mathcal{G}_K)$ with $K < K^*$. Then $\hat{\boldsymbol{\beta}}$ satisfies that, for grouping $\mathcal{G}$,

$$\begin{cases} -(X_{\mathcal{G}_k}\mathbf{1})^\top(\boldsymbol{y} - X\boldsymbol{\beta}) + n\sum_{j\in\mathcal{G}_k} \Delta_j(\boldsymbol{\beta}) = 0 & k = 1, \dots, K \\ |(X_A\mathbf{1})^\top(\boldsymbol{y} - X\boldsymbol{\beta}) - n\sum_{j\in A} \Delta_j(\boldsymbol{\beta})| \le n\frac{\lambda_2}{\tau}|\varepsilon \cap \{A \times (\mathcal{G}_k \backslash A)\}| & A \subset \mathcal{G}_k, |\mathcal{G}_k| > 1, \\ |\boldsymbol{x}_{(j)}^\top(\boldsymbol{y} - X\boldsymbol{\beta}) - n\Delta_j(\boldsymbol{\beta})| \le n\frac{\lambda_1}{\tau} & j \in \mathcal{G}_0, \end{cases} \tag{34}$$

where

$$\Delta_j(\boldsymbol{\beta}) = \lambda_1\tau^{-1}\text{sign}(\beta_j)I_{\{|\beta_j|\le\tau\}} + \lambda_2\tau^{-1}\sum_{j':(j',j)\in\varepsilon} \text{sign}(\beta_j - \beta_{j'})I_{\{|\beta_j-\beta_{j'}|\le\tau\}} + 2\lambda_3\beta_j I_{\{\beta_j<0\}}.$$

Denote $\mathcal{J} = \mathcal{J}_{11} \cap \mathcal{J}_{12} \cap \mathcal{J}_{21} \cap \mathcal{J}_{22}$, where $\mathcal{J}_{11} = \{\min_{j\notin\mathcal{G}_0^0} \hat{\beta}_j^{\text{ols}} > 2\tau\}$, $\mathcal{J}_{12} = \{\max_{j\in\mathcal{G}_0^0} |\boldsymbol{x}_{(j)}^\top(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{ols}})| \le n\lambda_1\tau^{-1}\}$, $\mathcal{J}_{21} = \{\min_{1\le k<l\le K^0} |\hat{\alpha}_k^{\text{ols}} - \hat{\alpha}_l^{\text{ols}}| > 2\tau\}$, $\mathcal{J}_{22} = \cap_{k=1,\dots,K^0:|\mathcal{G}_k^0|>1}\{\max_{A\subset\mathcal{G}_k^0} |(X_A\mathbf{1})^\top(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{ols}})| \le n\lambda_2\tau^{-1}|\varepsilon \cap \{A \times (\mathcal{G}_k^0\backslash A)\}|\}$. First, we show that $\hat{\boldsymbol{\beta}}^{\text{ols}}$ is a solution to (34) on $\mathcal{J}$. Note that, $\sum_{j\in\mathcal{G}_k^0} \Delta_j\left(\hat{\boldsymbol{\beta}}^{\text{ols}}\right) = 0$ on the set $\mathcal{J}_{11} \cap \mathcal{J}_{21}$. By the definition of $\hat{\boldsymbol{\beta}}^{\text{ols}}$, $(X_{\mathcal{G}_k^0}\mathbf{1})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{ols}}) = 0$. Thus, the first equation in (34) holds for $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{\text{ols}}$. Since $\sum_{j\in\mathcal{G}_k^0} \Delta_j\left(\hat{\boldsymbol{\beta}}^{\text{ols}}\right) = 0$ on $\mathcal{J}$, one can easily see that the second and third inequalities also hold for $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{\text{ols}}$.

Next, we show that (34) has a unique solution on $\mathcal{J}$, and thus $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{ols}}$. We provide the proof by contradiction. Assume that $\hat{\boldsymbol{\beta}} \ne \hat{\boldsymbol{\beta}}^{\text{ols}}$. Let $\mathcal{H} = (\mathcal{H}_1, \dots, \mathcal{H}_L) = \mathcal{G}_0^c \vee \mathcal{G}_0^{0c}$. Herein, we give an example to explain the sign '$\vee$'. Define two sets $A_1 = \{\{1,2,3,4\},\{5,6\}\}$, and $A_2 = \{\{1,2\},\{3,4,5,6\},\{7\}\}$. Then $A_1 \vee A_2 = \{\{1,2\},\{3,4\},\{5,6\},\{7\}\}$. Denote $\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}} = (\hat{\alpha}_{\mathcal{H}_1}^{\text{ols}}, \dots, \hat{\alpha}_{\mathcal{H}_L}^{\text{ols}})^\top$, $\hat{\boldsymbol{\alpha}}_{\mathcal{H}} = (\hat{\alpha}_{\mathcal{H}_1}, \dots, \hat{\alpha}_{\mathcal{H}_L})^\top$ the coefficients estimated by OLS and the algorithm 1, respectively. Then $S(\boldsymbol{\alpha}_{\mathcal{H}}) = (2n)^{-1}\|\boldsymbol{y} - Z_{\mathcal{H}}\boldsymbol{\alpha}_{\mathcal{H}}\|^2 + J(\boldsymbol{\alpha}_{\mathcal{H}})$, where

$$J(\boldsymbol{\alpha}_{\mathcal{H}}) = \lambda_1\sum_{k=1}^{L} |\mathcal{H}_k| \min\left\{\frac{|\alpha_{\mathcal{H}_k}|}{\tau}, 1\right\} + \lambda_2\sum_{1\le k<l\le L} |\varepsilon_{kl}| \min\left\{\frac{|\alpha_{\mathcal{H}_k} - \alpha_{\mathcal{H}_l}|}{\tau}, 1\right\}$$
$$+ \lambda_3\sum_{k=1}^{L} |\mathcal{H}_k|(\min\{\alpha_{\mathcal{H}_k}, 0\})^2$$

for $\boldsymbol{\alpha}_{\mathcal{H}} = (\alpha_{\mathcal{H}_1}, \dots, \alpha_{\mathcal{H}_L})^\top$, where ε_{kl} is the set of undirected edge between $\mathcal{H}_k$ and $\mathcal{H}_l$. We thus have

$$\frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} - \frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} = \frac{1}{n} Z_{\mathcal{H}}^{\top} Z_{\mathcal{H}} (\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}) + \boldsymbol{\varphi},$$

where $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_L)^{\top} = \boldsymbol{\varphi}_1 + \boldsymbol{\varphi}_2$, $\boldsymbol{\varphi}_1 = (\varphi_{11}, \dots, \varphi_{L1})^{\top}$, $\boldsymbol{\varphi}_2 = (\varphi_{12}, \dots, \varphi_{L2})^{\top}$, $\varphi_{k1} = \lambda_1 \tau^{-1} |\mathcal{H}_k| (a_k I_{\{|\hat{\alpha}_{\mathcal{H}_k}| \le \tau\}} - a_k^{\text{ols}} I_{\{|\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}}| \le \tau\}}) + \lambda_2 \tau^{-1} \sum_{l \ne k} |\varepsilon_{kl}| (b_{kl} I_{\{|\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_l}| \le \tau\}} - b_{kl}^{\text{ols}} I_{\{|\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} - \hat{\alpha}_{\mathcal{H}_l}^{\text{ols}}| \le \tau\}})$, $\varphi_{k2} = 2\lambda_3 (|\mathcal{H}_k| \hat{\alpha}_{\mathcal{H}_k} I_{\{\hat{\alpha}_{\mathcal{H}_k} < 0\}} - |\mathcal{H}_k| \hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} I_{\{\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} < 0\}})$, where $k = 1, \dots, L$, $a_k = \text{sign}(\hat{\alpha}_{\mathcal{H}_k})$, if $\hat{\alpha}_{\mathcal{H}_k} \ne 0$, $a_k \in [-1, 1]$ otherwise; $b_{kl} = \text{sign}(\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_l})$ if $\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_l} \ne 0$, $b_{kl} \in [-1, 1]$ otherwise. Similarly, we have a_k^{ols} and b_{kl}^{ols}. Note that $\|\boldsymbol{\varphi}_1\|^2 \le 4\tau^{-2}(\lambda_1 s^* + \lambda_2 |\mathcal{N}|)^2$.

Now, we consider two cases: (1) $\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}\| < \tau/2$ and (2) $\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}\| \ge \tau/2$. For each case, we show that both $\hat{\boldsymbol{\alpha}}_{\mathcal{H}}$ and $\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}$ are the local minimizers of $S(\boldsymbol{\alpha}_{\mathcal{H}})$ and $\hat{\boldsymbol{\alpha}}_{\mathcal{H}} = \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}$ on $\mathcal{J}$.

1. $\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}\| < \tau/2$. On the set $\mathcal{J}$, $\hat{\alpha}_{\mathcal{H}_k} \ge \hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} - |\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_k}^{\text{ols}}| \ge 2\tau - \tau/2 > \tau$ if $\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} > 2\tau$; $|\hat{\alpha}_{\mathcal{H}_k}| < |\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}}| + |\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_k}^{\text{ols}}| < \tau/2$ if $|\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}}| = 0$; $|\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_l}| \ge -|\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_k}^{\text{ols}}| - |\hat{\alpha}_{\mathcal{H}_l} - \hat{\alpha}_{\mathcal{H}_l}^{\text{ols}}| + |\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} - \hat{\alpha}_{\mathcal{H}_l}^{\text{ols}}| \ge \tau$ if $|\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} - \hat{\alpha}_{\mathcal{H}_l}^{\text{ols}}| \ge 2\tau$; $|\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_l}| \le |\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_k}^{\text{ols}}| + |\hat{\alpha}_{\mathcal{H}_l} - \hat{\alpha}_{\mathcal{H}_l}^{\text{ols}}| + |\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} - \hat{\alpha}_{\mathcal{H}_l}^{\text{ols}}| < \tau$ if $|\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} - \hat{\alpha}_{\mathcal{H}_l}^{\text{ols}}| = 0$. It implies that both $\hat{\boldsymbol{\alpha}}_{\mathcal{H}}$ and $\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}$ are the local minimizers of $S(\boldsymbol{\alpha}_{\mathcal{H}})$ and $\hat{\boldsymbol{\alpha}}_{\mathcal{H}} = \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}$ on $\mathcal{J}$.
2. $\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}\| \ge \tau/2$. By Cauchy–Schwarz inequality,

$$\left| \boldsymbol{\varphi}_1^{\top} (\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}) \right| \le \frac{2}{\tau} \big(\lambda_1 s^* + \lambda_2 |\mathcal{N}| \big) \|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}\|.$$

It is easy to verify that $(\hat{\alpha}_{\mathcal{H}_k} I_{\{\hat{\alpha}_{\mathcal{H}_k} < 0\}} - \hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} I_{\{\hat{\alpha}_{\mathcal{H}_k}^{\text{ols}} < 0\}})(\hat{\alpha}_{\mathcal{H}_k} - \hat{\alpha}_{\mathcal{H}_k}^{\text{ols}}) \ge 0$, followed by

$$\boldsymbol{\varphi}_2^{\top} (\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}) \ge 0.$$

By the assumption (A4),

$$\begin{aligned} &\left(\frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} - \frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} \right)^{\top} \frac{\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}}{\|\hat{\boldsymbol{\alpha}}_{\mathcal{H}} - \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}\|} \\ &\ge \min_{K(\mathcal{H}) \le K^*} \frac{\tau}{2} \lambda_{\min} \Big(\frac{1}{n} Z_{\mathcal{H}}^{\top} Z_{\mathcal{H}} \Big) - \frac{2}{\tau} \big(\lambda_1 s^* + \lambda_2 |\mathcal{N}| \big) > 0. \end{aligned} \tag{35}$$

On the other hand, $\frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} = 0$ and $\frac{\partial S(\hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}})}{\partial \boldsymbol{\alpha}_{\mathcal{H}}} = 0$ on $\mathcal{J}$ if $\hat{\boldsymbol{\alpha}}_{\mathcal{H}} \ne \hat{\boldsymbol{\alpha}}_{\mathcal{H}}^{\text{ols}}$, which contracts to (35). Therefore, the problem (34) has a unique solution on $\mathcal{J}$. That is $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{ols}}$ on $\mathcal{J}$, which yields that

$$\text{pr}(\hat{\boldsymbol{\beta}} \ne \hat{\boldsymbol{\beta}}^{\text{ols}}) \le \text{pr}(J^c) \le \text{pr}(\mathcal{J}_{11}^c) + \text{pr}(\mathcal{J}_{12}^c) + \text{pr}(\mathcal{J}_{21}^c) + \text{pr}(\mathcal{J}_{12}^c). \tag{36}$$

Next, we show the bounds of $\text{pr}(\mathcal{J}_{11}^c), \text{pr}(\mathcal{J}_{12}^c), \text{pr}(\mathcal{J}_{21}^c), \text{pr}(\mathcal{J}_{12}^c)$.

Before proceeding, we provide the following inequality, for $x > 0$, $\Phi(-x) \le (2\pi)^{-1/2} x^{-1} \exp(-x^2/2)$. If $x^2 \ge 2\log\{2na/(2\pi)^{1/2}\}$, $a \ge 1$, $x > 0$, then $2a\Phi(-x) \le cn^{-1}(\log n)^{-1/2}$.

For $\mathcal{J}_{11}^c$, by the assumptions (A1)–(A2), $\hat{\beta}_j^{\text{ols}} \sim N(\beta_j^0, var(\hat{\beta}_j^{\text{ols}}))$, where $var(\hat{\beta}_j^{\text{ols}}) \le n^{-1}\sigma^2\lambda_{\min}^{-1}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})$. If $\gamma_{\min} > 2\tau$, and $\{(\gamma_{\min} - 2\tau)n^{1/2}\lambda_{\min}^{1/2}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})\sigma^{-1}\}^2 \ge 2\log\{2n(p - |\mathcal{G}_0^0|)/(2\pi)^{1/2}\}$, then

$$\begin{aligned}
\text{pr}(\mathcal{J}_{11}^c) &\le \sum_{j\in\mathcal{G}_0^{0c}} \text{pr}\Big(\hat{\beta}_j^{\text{ols}} \le 2\tau\Big) \le \sum_{j\in\mathcal{G}_0^{0c}} \text{pr}(\beta_j^0 - |\hat{\beta}_j^{\text{ols}} - \beta_j^0| \le 2\tau) \\
&\le 2\big(p - |\mathcal{G}_0^0|\big)\boldsymbol{\Phi}\Big(-(\gamma_{\min} - 2\tau)n^{1/2}\lambda_{\min}^{1/2}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})\sigma^{-1}\Big) \qquad (37)\\
&= O\bigg(\frac{1}{n(\log n)^{1/2}}\bigg).
\end{aligned}$$

For $\mathcal{J}_{12}^c$, by (A1)–(A2), $\boldsymbol{x}_{(j)}^\top(\boldsymbol{y} - X^\top\hat{\boldsymbol{\beta}}^{\text{ols}}) = \boldsymbol{x}_{(j)}^\top(I - P_{Z_{\mathcal{G}_0^{0c}}})\boldsymbol{\epsilon} \sim N(0, \sigma^2\|(I - P_{Z_{\mathcal{G}_0^{0c}}})\boldsymbol{x}_{(j)}\|^2)$, and $\|(I - P_{Z_{\mathcal{G}_0^{0c}}})\boldsymbol{x}_{(j)}\|^2 \le \|\boldsymbol{x}_{(j)}\|^2$. If $(n\lambda_1\tau^{-1}\sigma^{-1}/\max_{1\le j\le p}\|\boldsymbol{x}_{(j)}\|)^2 \ge 2\log\{2n|\mathcal{G}_0^0|/(2\pi)^{1/2}\}$, then

$$\begin{aligned}
\text{pr}(\mathcal{J}_{12}^c) &\le \sum_{j\in\mathcal{G}_0^0} \text{pr}\left(\Big|\boldsymbol{x}_{(j)}^\top(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\text{ols}})\Big| > n\frac{\lambda_1}{\tau}\right) \\
&\le 2|\mathcal{G}_0^0|\boldsymbol{\Phi}\left(-\frac{n\lambda_1/\tau}{\sigma\max\limits_{1\le j\le p}\|\boldsymbol{x}_{(j)}\|}\right) = O\bigg(\frac{1}{n(\log n)^{1/2}}\bigg). \qquad (38)
\end{aligned}$$

For $\mathcal{J}_{21}^c$, by (A1)–(A2), $\hat{\alpha}_k^{\text{ols}} - \hat{\alpha}_l^{\text{ols}} \sim N(\alpha_k^0 - \alpha_l^0, var(\hat{\alpha}_k^{\text{ols}} - \hat{\alpha}_l^{\text{ols}}))$, where $var(\hat{\alpha}_k^{\text{ols}} - \hat{\alpha}_l^{\text{ols}}) \le 4n^{-1}\sigma^2\lambda_{\min}^{-1}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})$. If $\gamma_{\min} > 2\tau$, and $\{2^{-1}\sigma^{-1}(\gamma_{\min} - 2\tau)n^{1/2}\lambda_{\min}^{1/2}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})\}^2 \ge 2\log\{nK^0(K^0 - 1)/(2\pi)^{1/2}\}$, then

$$\begin{aligned}
\text{pr}(\mathcal{J}_{21}^c) &\le \sum_{1\le k<l\le K^0} \text{pr}(|\hat{\alpha}_k - \hat{\alpha}_l| \le 2\tau) \\
&\le \sum_{1\le k<l\le K^0} \text{pr}(|\alpha_k^0 - \alpha_l^0| - |(\hat{\alpha}_k - \hat{\alpha}_l) - (\alpha_k^0 - \alpha_l^0)| \le 2\tau) \\
&\le K^0(K^0 - 1)\boldsymbol{\Phi}\Big(-2^{-1}\sigma^{-1}(\gamma_{\min} - 2\tau)n^{1/2}\lambda_{\min}^{1/2}(n^{-1}Z_{\mathcal{G}_0^{0c}}^\top Z_{\mathcal{G}_0^{0c}})\Big) \qquad (39)\\
&= O\bigg(\frac{1}{n(\log n)^{1/2}}\bigg).
\end{aligned}$$

For $\mathcal{J}_{22}^c$, by (A1)–(A2), $(X_A\mathbf{1})^\top(\boldsymbol{y} - X^\top\hat{\boldsymbol{\beta}}^{\text{ols}}) = (X_A\mathbf{1})^\top(I - P_{Z_{\mathcal{G}_0^{0c}}})\boldsymbol{\epsilon} \sim N(0, \sigma^2\|(I - P_{Z_{\mathcal{G}_0^{0c}}})X_A\mathbf{1}\|^2)$, and $\|(I - P_{Z_{\mathcal{G}_0^{0c}}})X_A\mathbf{1}\|^2 \le \|X_A\mathbf{1}\|^2$. Denote $\mathcal{D} = \max_{k,A\subset\mathcal{G}_k^0}\|X_A\mathbf{1}\|/|\varepsilon \cap \{A \times (\mathcal{G}_k^0\backslash A)\}|$. If $(2^{-1}n\lambda_2\tau^{-1}\sigma^{-1}/\mathcal{D})^2 \ge 2\log\{2n|\mathcal{N}|/(2\pi)^{1/2}\}$, then

$$\begin{aligned}\mathrm{pr}(\mathcal{J}_{22}^c) &\leq \sum_{k=1,\ldots,K^0; A\subset\mathcal{G}_k^0} \mathrm{pr}\left(\left|(X_A\mathbf{1})^\top(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}^{\mathrm{ols}})\right| > n\frac{\lambda_2}{\tau}\left|\varepsilon \cap \{A \times (\mathcal{G}_k^0\backslash A)\}\right|\right)\\ &\leq 2|\mathcal{N}|\boldsymbol{\Phi}\left(-\frac{n\lambda_2/\tau}{2\sigma\mathcal{D}}\right) = O\left(\frac{1}{n(\log n)^{1/2}}\right).\end{aligned} \tag{40}$$

By (36)–(40), we thus have $\mathrm{pr}(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{\mathrm{ols}}) = O\left(\frac{1}{n(\log n)^{1/2}}\right)$, which, together with Lemma 2.1, yields that

$$\mathrm{pr}(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{\mathrm{ora}}) \leq \mathrm{pr}(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{\mathrm{ols}}) + \mathrm{pr}(\hat{\boldsymbol{\beta}}^{\mathrm{ora}} \neq \hat{\boldsymbol{\beta}}^{\mathrm{ols}}) = O\left(\frac{1}{n(\log n)^{1/2}}\right).$$

(2) Note that, $\hat{\boldsymbol{\alpha}}$ satisfies that $-Z_{\mathcal{G}_0^c}^\top\left(\boldsymbol{y} - Z_{\mathcal{G}_0^c}\hat{\boldsymbol{\alpha}}\right) + 2n\lambda_3 M_0\hat{\boldsymbol{\alpha}} + n\hat{\boldsymbol{\delta}} = 0$, where M_0 is a $K \times K$ diagonal matrix with diagonal elements $|\mathcal{G}_k|I_{\{\hat{\boldsymbol{\alpha}}_k<0\}}$ for $k = 1, \ldots, K$; $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \ldots, \hat{\delta}_K)^\top$, $\hat{\delta}_k = \sum_{j\in\mathcal{G}_k} \Upsilon_j(\hat{\boldsymbol{\beta}})$, and $\Upsilon_j(\boldsymbol{\beta}) = \lambda_1\tau^{-1}\mathrm{sign}(\beta_j)I_{\{|\beta_j|\leq\tau\}} + \lambda_2\tau^{-1} \sum_{j':(j',j)\in\varepsilon} \mathrm{sign}(\beta_j - \beta_{j'})I_{\{|\beta_j-\beta_{j'}|\leq\tau\}}$. Note that $\|\hat{\boldsymbol{\delta}}\|^2 \leq \tau^{-2}(\lambda_1 s^* + \lambda_2|\mathcal{N}|)^2$.We obtain that $\hat{\boldsymbol{\alpha}} = (Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c} + 2n\lambda_3 M_0)^{-1}(Z_{\mathcal{G}_0^c}^\top\boldsymbol{y} - n\hat{\boldsymbol{\delta}})$, followed by

$$\begin{aligned}&\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2\\ &= \|Z_{\mathcal{G}_0^c}(Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c} + 2n\lambda_3 M_0)^{-1}(Z_{\mathcal{G}_0^c}^\top\boldsymbol{y} - n\hat{\boldsymbol{\delta}}) - Z_{\mathcal{G}_0^{0c}}\boldsymbol{\alpha}^0\|^2\\ &= \|\{I - Z_{\mathcal{G}_0^c}(Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c} + 2n\lambda_3 M_0)^{-1}Z_{\mathcal{G}_0^c}^\top\}Z_{\mathcal{G}_0^{0c}}\boldsymbol{\alpha}^0 - Z_{\mathcal{G}_0^c}(Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c} + 2n\lambda_3 M_0)^{-1}Z_{\mathcal{G}_0^c}^\top\boldsymbol{\epsilon}\\ &\quad + nZ_{\mathcal{G}_0^c}(Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c} + 2n\lambda_3 M_0)^{-1}\hat{\boldsymbol{\delta}})\|^2\\ &\leq 3\|X\boldsymbol{\beta}^0\|^2 + 3\|\boldsymbol{\epsilon}\|^2 + \frac{3\tau^2 n}{16}\min_{K(\mathcal{G}_0^c)\leq K^*}\lambda_{\min}\left(\frac{1}{n}Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c}\right).\end{aligned} \tag{41}$$

Denote $T_1 = n^{-1}E(\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2 I_{\{G\}})$ and $T_2 = n^{-1}E(\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2 I_{\{G^c\}})$, where $G = \{n^{-1}\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2 \geq D\}$. By the definition, we have $n^{-1}E(\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2) = T_1 + T_2$. Next, we work on T_1, T_2. Let

$$D = \frac{3}{n}\|X\boldsymbol{\beta}_0\|^2 + 10\sigma^2 + \frac{3\tau^2}{16}\min_{K(\mathcal{G}_0^c)\leq K^*}\lambda_{\min}\left(\frac{1}{n}Z_{\mathcal{G}_0^c}^\top Z_{\mathcal{G}_0^c}\right). \tag{42}$$

For T_1, it follows that

$$
\begin{aligned}
&\int_D^\infty \mathrm{pr}\big(n^{-1}\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2 \geq x\big)\mathrm{d}x \\
&\leq \int_{10\sigma^2}^\infty \mathrm{pr}\big(3n^{-1}\|\boldsymbol{\epsilon}\|^2 \geq x\big)\mathrm{d}x \\
&\leq \int_{10\sigma^2}^\infty E\left\{\exp\left(\frac{t\|\boldsymbol{\epsilon}\|^2}{\sigma^2}\right)\exp\left(-\frac{ntx}{3\sigma^2}\right)\right\}\mathrm{d}x \\
&\leq \int_{10\sigma^2}^\infty \exp\left\{-\frac{n}{9\sigma^2}(x-9\sigma^2)\right\}\mathrm{d}x \\
&\leq \frac{9\sigma^2}{n}\exp\left(-\frac{n}{9}\right) = o\left(\frac{K^0\sigma^2}{n}\right).
\end{aligned} \tag{43}
$$

By (41) and (42), thus the first '≤' follows. In view of the moment generating function for Chi-squared distribution, taking $t = 1/3$, the third '≤' holds. For T_2,

$$
T_2 = E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G^c\}} I_{\{\hat{\boldsymbol{\beta}}\neq\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\}}\right) + E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 (1 - I_{\{G\}}) I_{\{\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\}}\right). \tag{44}
$$

For the first term in (44), if $D = o\{K^0(\log n)^{1/2}\}$, then

$$
E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G^c\}} I_{\{\hat{\boldsymbol{\beta}}\neq\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\}}\right) \leq D\mathrm{pr}\left(\hat{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}^{\mathrm{ols}}\right) = o\left(\frac{K^0\sigma^2}{n}\right). \tag{45}
$$

For the second term in (44),

$$
E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G\}} I_{\{\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\}}\right) \leq E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{G\}}\right) = o\left(\frac{K^0\sigma^2}{n}\right), \tag{46}
$$

$$
E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\right\|^2 I_{\{\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{\mathrm{ols}}\}}\right) \leq E\left(\frac{1}{n}\left\|X\hat{\boldsymbol{\beta}}^{\mathrm{ols}} - X\boldsymbol{\beta}^0\right\|^2\right) = \frac{K^0\sigma^2}{n}. \tag{47}
$$

By (43), (44)–(47), $n^{-1}E(\|X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^0\|^2) = T_1 + T_2 = n^{-1}K^0\sigma^2(1 + o(1))$. □

References

Arnold, T. B., Tibshirani, R. J. (2016). Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, *25*(1), 1–27.

Esser, E., Lou, Y. F., Xin, J. (2013). A method for finding structured sparse solutions to nonnegative least squares problems with applications. *SIAM Journal on Imaging Sciences*, *6*(4), 2010–2046.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Frank, L. E., Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109–135.

Friedman, J., Hastie, T., Simon, N., Tibshirani, R. (2016). Lasso and elastic-net regularized generalized linear models. *R-Package Version*, *2*(0–5), 2016.

Fu, A., Narasimhan, B., Boyd, S. (2017). *CVXR: An R package for disciplined convex optimization*. arXiv:1711.07582.

Goeman, J. J. (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal*, *52*(1), 70–84.

Hu, Z., Follmann, D. A., Miura, K. (2015). Vaccine design via nonnegative lasso-based variable selection. *Statistics in Medicine*, *34*(10), 1791–1798.

Huang, J., Ma, S., Xie, H., Zhang, C. H. (2009). A group bridge approach for variable selection. *Biometrika*, *96*(2), 339–355.

Itoh, Y., Duarte, M. F., Parente, M. (2016). Perfect recovery conditions for non-negative sparse modeling. *IEEE Transactions on Signal Processing*, *65*(1), 69–80.

Jang, W., Lim, J., Lazar, N., Loh, J. M., McDowell, J., Yu, D. (2011). Regression shrinkage and equality selection for highly correlated predictors with HORSES. *Biometrics*, *64*, 1–23.

Koike, Y., Tanoue, Y. (2019). Oracle inequalities for sign constrained generalized linear models. *Econometrics and Statistics*, *11*, 145–157.

Luenberger, D. G., Ye, Y. (2015). *Linear and nonlinear programming*, Vol. 228. New York: Springer.

Mandal, B. N., Ma, J. (2016). l_1 regularized multiplicative iterative path algorithm for non-negative generalized linear models. *Computational Statistics and Data Analysis*, *101*, 289–299.

Meinshausen, N. (2013). Sign-constrained least squares estimation for high-dimensional regression. *Electronic Journal of Statistics*, *7*, 1607–1631.

Mullen, K. M., van Stokkum, I. H. (2012). *The Lawson–Hanson algorithm for nonnegative least squares (NNLS)*. CRAN: R package. https://cran.r-project.org/web/packages/nnls/nnls.pdf.

Rekabdarkolaee, H. M., Boone, E., Wang, Q. (2017). Robust estimation and variable selection in sufficient dimension reduction. *Computational Statistics and Data Analysis*, *108*, 146–157.

Renard, B. Y., Kirchner, M., Steen, H., Steen, J. A., Hamprecht, F. A. (2008). NITPICK: Peak identification for mass spectrometry data. *BMC Bioinformatics*, *9*(1), 355.

Shadmi, Y., Jung, P., Caire, G. (2019). *Sparse non-negative recovery from biased sub-Gaussian measurements using NNLS*. arXiv:1901.05727.

She, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics*, *4*, 1055–1096.

Shen, X., Huang, H. C., Pan, W. (2012a). Simultaneous supervised clustering and feature selection over a graph. *Biometrika*, *99*(4), 899–914.

Shen, X., Pan, W., Zhu, Y. (2012b). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, *107*(497), 223–232.

Shen, X., Pan, W., Zhu, Y., Zhou, H. (2013). On constrained and regularized high-dimensional regression. *Annals of the Institute of Statistical Mathematics*, *65*(5), 807–832.

Slawski, M., Hein, M. (2010). Sparse recovery for protein massspectrometry data. In *NIPS workshop on practical applications of sparse modelling*.

Slawski, M., Hein, M. (2013). Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization. *Electronic Journal of Statistics*, *7*, 3004–3056.

Slawski, M., Hussong, R., Tholey, A., Jakoby, T., Gregorius, B., Hildebrandt, A., Hein, M. (2012). Isotope pattern deconvolution for peptide mass spectrometry by non-negative least squares/least absolute deviation template matching. *BMC Bioinformatics*, *13*(1), 291.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R., Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, *9*(1), 18–29.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(1), 91–108.

Tibshirani, R. J., Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, *39*(3), 1335–1371.

Wen, Y. W., Wang, M., Cao, Z., Cheng, X., Ching, W. K., Vassiliadis, V. S. (2015). Sparse solution of nonnegative least squares problems with applications in the construction of probabilistic Boolean networks. *Numerical Linear Algebra with Applications*, *22*(5), 883–899.

Wu, L., Yang, Y. (2014). Nonnegative elastic net and application in index tracking. *Applied Mathematics and Computation*, *227*, 541–552.

Wu, L., Yang, Y., Liu, H. (2014). Nonnegative-lasso and application in index tracking. *Computational Statistics and Data Analysis*, *70*, 116–126.

Xiang, S., Shen, X., Ye, J. (2015). Efficient nonconvex sparse group feature selection via continuous and discrete optimization. *Artificial Intelligence*, *224*, 28–50.

Yang, S., Yuan, L., Lai, Y. C., Shen, X., Wonka, P., Ye, J. (2012). Feature grouping and selection over an undirected graph. *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 922–930). ACM. New York.

Yang, Y., Wu, L. (2016). Nonnegative adaptive lasso for ultra-high dimensional regression models and a two-stage method applied in financial modeling. *Journal of Statistical Planning and Inference*, *174*, 52–67.

Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*: *Series B (Statistical Methodology)*, *68*(1), 49–67.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942.

Zhu, Y., Shen, X., Pan, W. (2013). Simultaneous grouping pursuit and feature selection over an undirected graph. *Journal of the American Statistical Association*, *108*(502), 713–725.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society)*: *Series B (Statistical Methodology*, *67*(2), 301–320.

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