The Annals of Statistics
2024, Vol. 52, No. 1, 285–310
https://doi.org/10.1214/23-AOS2347
© Institute of Mathematical Statistics, 2024

SUPERVISED HOMOGENEITY FUSION: A COMBINATORIAL APPROACH

BY WEN WANG[1,a], SHIHAO WU[2,c], ZIWEI ZHU[2,d], LING ZHOU[3,e] AND PETER X.-K. SONG[1,b]

[1]*Department of Biostatistics, University of Michigan, Ann Arbor,* [a]*wangwen@umich.edu,* [b]*pxsong@umich.edu*

[2]*Department of Statistics, University of Michigan, Ann Arbor,* [c]*wshihao@umich.edu,* [d]*zzw9348ustc@gmail.com*

[3]*Center of Statistical Research, Southwestern University of Finance and Economics,* [e]*lingzhou@swufe.edu.cn*

Fusing regression coefficients into homogeneous groups can unveil those coefficients that share a common value within each group. Such groupwise homogeneity reduces the intrinsic dimension of the parameter space and unleashes sharper statistical accuracy. We propose and investigate a new combinatorial grouping approach called L_0-Fusion that is amenable to mixed integer optimization (MIO). On the statistical aspect, we identify a fundamental quantity called *MSE grouping sensitivity* that underpins the difficulty of recovering the true groups. We show that L_0-Fusion achieves grouping consistency under the weakest possible requirement of the grouping sensitivity: if this requirement is violated, then the minimax risk of group misspecification will fail to converge to zero. Moreover, we show that in the high-dimensional regime, one can apply L_0-Fusion with a sure screening set of features without any essential loss of statistical efficiency, while reducing the computational cost substantially. On the algorithmic aspect, we provide an MIO formulation for L_0-Fusion along with a warm start strategy. Simulation and real data analysis demonstrate that L_0-Fusion exhibits superiority over its competitors in terms of grouping accuracy.

1. Introduction. Identifying homogeneous groups of regression coefficients has received increasing attention, because the resulting regression model provides better scientific interpretations and enhances predictive performance in many applications. In some occasions, features or covariates naturally act in groups to influence outcomes, so knowing group structures of the features help scientists gain new knowledge about a physical system of interest. From a modeling perspective, aggregating covariates with similar effects along with the response reduces model complexity and improves interpretability, especially in the high-dimensional regime. There have been a flurry of works under this direction; see, for example, Bondell and Reich (2008), Jeon, Kwon and Choi (2017), Ke, Fan and Wu (2015), Shen and Huang (2010), Zhou et al. (2022), Zhu, Shen and Pan (2013), among others. There is a vast literature in discovering homogeneous groups of observations or individuals in an overly heterogeneous population. However, these existing methods cannot be applied to our problem that aims to group regression parameters. Identifying group structures of regression parameters is crucial to learn the underlying heterogeneous covariates' effects, which is then leveraged to reach a more appropriate model for data analyses. A partial list of the literature includes Ke, Li and Zhang (2016), Lian, Qiao and Zhang (2021), Ma and Huang (2017), Shen and He (2015), just name a few. The focus of this paper is on pursuing homogeneous groups of regression coefficients in which we do not have any prior knowledge about their true group structures.

Homogeneity fusion is carried out routinely in environmental health sciences in a manual and subjective manner to evaluate the effect of a given set of toxicants on certain health

Received November 2022; revised October 2023.

MSC2020 subject classifications. Primary 60J05, 90C11; secondary 62P10.

Key words and phrases. Clustering, dimension reduction, mixed integer optimization, linear model, supervised learning.

outcomes. Consider p toxicants, whose concentrations are denoted by $X_1, \dots, X_p$, respectively, q other covariates $\{Z_k\}_{k=1}^q$ and a outcome variable Y. Scientists typically consider a linear regression model $Y \sim \sum_{j=1}^p \beta_j X_j + \sum_{k=1}^q \alpha_k Z_k$ to evaluate effect of a mixture $A := \sum_{j=1}^p \beta_j X_j$ on outcome Y. One common practice to reduce model complexity and facilitate scientific interpretation is aggregating the exposure of similar toxicants to yield a sum-mixture (e.g., $\sum_j X_j$). For example, SumDEHP is a sum of four phthalates, MECPP, MEOHP, MEHHP and MEHP, which quantifies total DEHP exposure from products such as PVC plastics used in food processing/packaging materials, building materials and medical devices (Schettler (2006), Kobrosly et al. (2012), Braun et al. (2012)). See also Marie, Vendittelli and Sauvant-Rochat (2015), Marsee et al. (2006) for another sum-mixture called SumAA that adds three extra phthalates MBP, MiBP and MBzP to SumDEHP. Learning such a sum-mixture requires the toxicants within the same mixture to share the same regression coefficients in the linear model. Unfortunately, in practice the formation of a sum-mixture is done manually by scientists in an ad hoc fashion. There has been long of interest to develop a data-driven homogeneity fusion methodology that provides a needed statistical toolbox for scientists to identify and include important toxicants, while excluding unimportant ones, in the formation of a toxic mixture. This new approach can greatly reduce subjectivity in data processing and yield robust scientific conclusions and insights on the relationship between toxicants and outcome. This motivates us to pursue parsimony by regularizing coefficients β_j in addition to homogeneity pursuit of those nonzero coefficients in our methodology.

Suppose that the true linear model with K_0 groups of nonzero coefficients takes the form

$$Y = \sum_{j=1}^p \beta_j^* X_j + \sum_{k=1}^q \alpha_k^* Z_k + \varepsilon, \quad \beta_j^* \in \{0, \gamma_1^*, \gamma_2^*, \dots, \gamma_{K_0}^*\}, j = 1, \dots, p, \tag{1.1}$$

where random error ε is mean zero and sub-Gaussian with ψ_2-norm bounded by σ, namely $\|\varepsilon\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}\exp(\varepsilon^2/t^2) \le 2\} \le \sigma$ (Vershynin (2018)), and the coefficients $\{\beta_j^*\}_{j=1}^p$ belong to a set including 0 and K_0 unknown different nonzero values $\{\gamma_k^*\}_{k=1}^{K_0}$. Note that the group membership of each nonzero β_j is not observed in data collection. Write $\boldsymbol{\alpha}^* = (\alpha_1^*, \dots, \alpha_q^*)^\top \in \mathbb{R}^q$, $\boldsymbol{\beta}^* = (\beta_1^*, \dots, \beta_p^*)^\top \in \mathbb{R}^p$ and $\boldsymbol{\gamma}^* = (\gamma_1^*, \dots, \gamma_{K_0}^*)^\top \in \mathbb{R}^{K_0}$. Our main goal in this paper is to estimate $\boldsymbol{\gamma}^*$, $\boldsymbol{\beta}^*$ and $\boldsymbol{\alpha}^*$ simultaneously based on an independent and identically distributed ($i.i.d.$) sample $\{(\mathbf{x}_i, \mathbf{z}_i, y_i)\}_{i=1}^n$ of size n. In the case of high dimension, $\boldsymbol{\beta}$ is often assumed sparse so that we perform feature selection and grouping simultaneously to ensure statistical consistency.

We now review and discuss some important works related to model (1.1). Shen and Huang (2010) considered model (1.1) without covariates $\{Z_k\}_{k=1}^q$ and proposed to minimize the following objective with respect to $\boldsymbol{\beta}$: $S_1(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda_1 \sum_{j<j'} J_\tau(|\beta_j - \beta_{j'}|)$, where λ_1 is a tuning parameter that is associated with fusion strength, and $J_\tau(z) = \min(z\tau^{-1}, 1)$ is a surrogate of the indicator function $1_{z\neq 0}(z)$, with $\tau > 0$ representing the approximation error of $J_\tau(z)$ to the L_0 penalty $1_{z\neq 0}(z)$. Such penalty on the pairwise difference can lead to redundant comparisons and extra computational complexity. Note that there is no sparsity regularization in $S_1(\boldsymbol{\beta})$. As an extension, Zhu, Shen and Pan (2013) considered simultaneous grouping pursuit and feature selection by further penalizing individual coefficients, that is, minimizing $S_2(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2 + \lambda_1 \sum_{(j,j')\in E} J_\tau(||\beta_j| - |\beta_{j'}||) + \lambda_2 \sum_{j=1}^p J_\tau(|\beta_j|)$. Here, E is the edge set of an undirected graph with p nodes representing $\{X_j\}_{j=1}^p$. If X_i and X_j can be grouped, then there is an edge between nodes i and j; otherwise, there is no edge. Available prior knowledge of E reduces computational burden and improves estimation efficiency. However, it is always challenging in practice to obtain a plausible estimate of E, which makes the method less appealing. Ke, Fan and Wu (2015) proposed a different method named as clustering algorithm in regression via data-driven segmentation (CARDS). They use a preliminary estimate to determine "adjacent" coefficient pairs for fusion and only penalize distances between the two coefficients in each adjacent pairs by folded concave penalty function. Therefore, the CARDS estimator depends on the initial ordering of the coefficients, which could be unstable especially when the effect sizes are small (e.g., weak signals).

We propose to pursue homogeneity and sparsity simultaneously through a combinatorial approach called L_0-Fusion. Specifically, we estimate $\boldsymbol{\beta}^*$ by the least squares with an exact group constraint (the estimated β_j's can only take at most K distinct nonzero values) and an L_0 sparsity constraint. To obtain this estimator, we formulate the corresponding optimization problem as a mixed integer optimization (MIO) problem. Bertsimas, King and Mazumder (2016) demonstrated that MIO provides a computationally tractable approach to solve the classical best subset selection (BSS) problem of a practical scale: With the sample size in thousands and the dimension in hundreds, a MIO algorithm can achieve provable optimality in minutes. Such success of MIO and the similar combinatorial nature of the BSS problem inspire us to seek for a MIO formulation of the L_0-Fusion problem. Our main contributions are summarized as follows: (a) To the best of our knowledge, it is the first time that we formulate the group pursuing as a MIO problem; (b) we show that the estimator derived from the L_0-Fusion problem achieves grouping consistency once the loss function is reasonably sensitive to a certain degree of grouping error; (c) we discover that the grouping sensitivity requirement in (b) turns out to be necessary (up to a universal constant) for any approach to achieve selection and grouping consistency; and (d) we provide a warm start algorithm with convergence guarantee for the L_0-Fusion problem, in order to accelerate the MIO solver.

The rest of the article is organized as follows. Section 2 introduces the L_0-Fusion method and a "screen then group" strategy to tackle high dimension. Section 3 presents the statistical theory on the selection and grouping consistency of the L_0-Fusion method and a necessary condition to achieve such consistency, and some other related theoretical analyses. Section 4 studies the estimation and prediction error of the estimator obtained by L_0-Fusion. Section 5 introduces our MIO formulation for the L_0-Fusion problem together with a warm up algorithm. Section 6 demonstrates significant superiority of the L_0-Fusion approach over existing ones in terms of grouping accuracy in both low-dimensional and high-dimensional regimes. We also apply L_0-Fusion to a metabolomics data set to aggregate concentration of similar lipids to predict the body mass index (BMI). The Supplementary Material (Wang et al. (2024)) includes all technical details, including the proofs of major theoretical results.

2. Methodology: L_0-fusion.

2.1. *Notation.* We use regular letters, bold regular letters and bold capital letters to denote scalars, vectors and matrices, respectively. For a positive integer n, denote $\{1,\dots,n\}$ by $[n]$. Let $\mathbb{Z}$ be the collection of all integer numbers. For any $x \in \mathbb{R}$, denote $\lfloor x \rfloor := \max\{m \in \mathbb{Z} : m \le x\}$ and $\lceil x \rceil := \min\{n \in \mathbb{Z} : n > x\}$. For any two sets $\mathcal{A}$ and $\mathcal{B}$, let $\mathcal{A}\backslash\mathcal{B} := \mathcal{A} \cap \mathcal{B}^c$. For any two subsets $\mathcal{A}$ and $\mathcal{B}$ of $\mathbb{R}^n$, define $\mathcal{A} - \mathcal{B} = \{a - b \mid a \in \mathcal{A}, b \in \mathcal{B}\}$. For any matrix $\mathbf{A}$, $\mathbf{A}^\top$ denotes the transpose of $\mathbf{A}$ and $\mathbf{A}^+$ denotes its Moore–Penrose inverse. Given $\mathcal{B} = \{i_1,\dots,i_{|\mathcal{B}|}\} \subset [p]$, $\mathbf{X}_\mathcal{B}$ denotes the submatrix of $\mathbf{X}$ with columns indexed in $\mathcal{B}$, and $\boldsymbol{\beta}_\mathcal{B}$ denotes $(\beta_{i_1},\dots,\beta_{i_{|\mathcal{B}|}})^\top$. Given any $a, b \in \mathbb{R}$, we say $a \lesssim b$ if there exists a universal constant $C > 0$ such that $a \le Cb$; we say $a \gtrsim b$ if there exists a universal constant $c > 0$ such that $a \ge cb$; we say $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. For any two positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \gg b_n$ if $a_n/b_n \to \infty$ as $n \to \infty$. For any event $\mathcal{A}$, we use $I(\mathcal{A})$ to denote the indicator function associated with $\mathcal{A}$, that is, $I(\mathcal{A}) = 1$ if $\mathcal{A}$ occurs, and $I(\mathcal{A}) = 0$ otherwise.

2.2. *L_0-fusion with feature screening.* Consider a random sample of n independent observations $(\mathbf{x}_i, \mathbf{z}_i, y_i)_{i\in[n]}$ from model (1.1). Our paper concerns the following combinatorial optimization problem to perform feature selection and homogeneity fusion simultaneously:

$$
\begin{aligned}
&\min_{\boldsymbol{\alpha}\in\mathbb{R}^q,\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\gamma}\in\mathbb{R}^K} && \sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - \mathbf{z}_i^\top\boldsymbol{\alpha})^2, \\
&\text{subject to:} && \beta_j \in \{0, \gamma_1, \gamma_2, \dots, \gamma_K\} \quad \forall j \in [p], \\
& && \|\boldsymbol{\beta}\|_0 := \sum_{j=1}^{p} I(\beta_j \neq 0) \le s.
\end{aligned}
\tag{2.1}
$$

The first constraint in (2.1) requires the nonzero group number to be bounded by K, and the second constraint requires the sparsity of $\boldsymbol{\beta}$ to be bounded by s. Clearly, the problem in (2.1) not only restricts the ℓ_0-norm of $\boldsymbol{\beta}$ by K but also fuses similar components of $\boldsymbol{\beta}$. Thus, we refer to (2.1) as the L_0-Fusion problem in this paper. Without the grouping constraint, problem in (2.1) boils down to the well-known best subset selection (BSS) problem (Garside (1965), Hocking and Leslie (1967), Beale, Kendall and Mann (1967)) with subset size s. Of note, problem (2.1) is NP-hard in the presence of both cardinality and grouping constraints. Despite such computational challenge, in Section 5.1 we provide a MIO formulation of (2.1) that is amenable to modern integer optimization solvers such as GUROBI and MOSEK. As shown in our numerical examples in this paper, when the dimension $p \le 100$, GUROBI can solve the L_0-Fusion problem within seconds.

In many practical occasions, the number of features, p, is often in thousands or even millions. In such cases, directly solving the above L_0-Fusion problem is computationally burdensome or even prohibitive. To tackle this challenge, we propose a "*screen then group*" strategy. In the screening stage, we obtain a screening set, denoted by $\widetilde{\mathcal{S}}$, which is generated by a certain preliminary feature screening procedure. Examples of popular screening techniques include, but are not limited to, penalized least squares methods (Tibshirani (1996), Fan and Li (2001), Zhang (2010)), sure independence screening (Fan and Lv (2008)) and sparsity constraint method (Needell and Tropp (2009), Guo, Zhu and Fan (2020)). We hope to choose a screening method so that the resulting set $\widetilde{\mathcal{S}}$ enjoys the sure screening property in the sense that the true support set $\mathcal{S}^0 \subseteq \widetilde{\mathcal{S}}$ with high probability. Then in the next *grouping* stage, we solve the L_0-Fusion problem in (2.1) on the reduced design $\mathbf{X}_{\widetilde{\mathcal{S}}}$ to estimate nonzero coefficients and obtain their groups.

In this paper, we choose CoSaMP (Compressive Sampling Matching Pursuit), a two-stage iterative hard thresholding (IHT) algorithm proposed by Needell and Tropp (2009), as our choice of screener for dimension reduction. Algorithm 1 presents its pseudocode of CoSaMP that performs two rounds of hard thresholding in each iteration. First, it expands the model by recruiting the largest coordinates of the gradient (lines 3–4); second, it contracts the model by discarding the smallest components of the refitted signal on the expanded model (lines 5–7). Guo, Zhu and Fan (2020) showed that under a high-dimensional sparse regression setup, CoSaMP through the IHT procedure can achieve sure screening properties within few iterations under highly correlated designs. In addition, Zhu and Wu (2021) showed numerically that CoSaMP yields much fewer false discoveries than LASSO, SCAD and MCP on early solution paths, particularly in the presence of high correlations among predictors. These results signify CoSaMP as an efficient and reliable screener to help substantially reduce the dimension while retaining the true signals. We emphasize that the low false discovery rate (FDR) in our two-stage analysis strategy is crucial to reasonably controlling the dimension of the reduced design on which the L_0-Fusion procedure becomes computationally tractable.

Algorithm 1: CoSaMP($\mathbf{X}$, $\mathbf{y}$, $\widehat{\boldsymbol{\beta}}_0$, π, l, τ)

```
Input: Design matrix $\mathbf{X}$, response $\mathbf{y}$, initial value $\widehat{\boldsymbol{\beta}}_0$, projection size $\pi$, expansion size $l$, convergence threshold $\tau > 0$
$t \leftarrow 0$
repeat
    $\mathcal{G}_t \leftarrow \mathcal{T}_{\mathrm{abs}}(\nabla\mathcal{L}(\widehat{\boldsymbol{\beta}}_t), l)$
    $\mathcal{S}_t^\star \leftarrow \mathrm{supp}(\widehat{\boldsymbol{\beta}}_t) \cup \mathcal{G}_t$
    $\widehat{\boldsymbol{\beta}}_t^\star \leftarrow (\mathbf{X}_{\mathcal{S}_t^\star}^\top \mathbf{X}_{\mathcal{S}_t^\star})^+ \mathbf{X}_{\mathcal{S}_t^\star}^\top \mathbf{y}$
    $\mathcal{S}_t \leftarrow \mathcal{T}_{\mathrm{abs}}(\widehat{\boldsymbol{\beta}}_t^\star, \pi)$
    $\widehat{\boldsymbol{\beta}}_{t+1} \leftarrow (\mathbf{X}_{\mathcal{S}_t}^\top \mathbf{X}_{\mathcal{S}_t})^+ \mathbf{X}_{\mathcal{S}_t}^\top \mathbf{y}$
    $t \leftarrow t + 1$
until $\|\widehat{\boldsymbol{\beta}}_t - \widehat{\boldsymbol{\beta}}_{t-1}\|_2 < \tau$
$\widehat{\boldsymbol{\beta}}^{\mathrm{cs}} \leftarrow \widehat{\boldsymbol{\beta}}_t$
Output: $\widehat{\boldsymbol{\beta}}^{\mathrm{cs}}$
```

3. Group recovery theory.

In this section, we show that the global minimizers of problem (2.1) reconstruct the ideal "oracle estimator," provided that the MSE objective in (2.1) is reasonably sensitive to grouping error. By oracle estimator, we mean the estimator obtained under the prior knowledge of the true grouping, We establish both sufficient and necessary conditions to achieve grouping consistency as well as selection consistency. Define the parameter space $\boldsymbol{\Theta}(K, s) := \{\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top \in \mathbb{R}^{p+q} \mid \boldsymbol{\gamma} \in \mathbb{R}^K, \beta_j \in \{0, \gamma_1, \ldots, \gamma_K\}, \forall j \in [p], \|\boldsymbol{\beta}\|_0 \le s\}$. Let $\mathcal{G}(\boldsymbol{\beta}; r) := \{j \in [p] \mid \beta_j = r\}$ be the set containing all indices of the components of $\boldsymbol{\beta}$ equal to r; denote the collection of groups of $\boldsymbol{\beta}$ by $\mathbb{G}(\boldsymbol{\beta}) = \{\mathcal{G}(\boldsymbol{\beta}; r) \mid r \neq 0, \mathcal{G}(\boldsymbol{\beta}; r) \neq \varnothing\}$. Let $|\mathcal{G}(\boldsymbol{\beta}; r)|$ and $|\mathbb{G}(\boldsymbol{\beta})|$ be the cardinality of $\mathcal{G}(\boldsymbol{\beta}; r)$ and $\mathbb{G}(\boldsymbol{\beta})$, respectively. We write the $n \times p$ design matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ and the $n \times q$ matrix $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_q)$, where $\mathbf{X}_j$ and $\mathbf{Z}_k$ are the jth and kth columns of $\mathbf{X}$ and $\mathbf{Z}$, respectively. We consider the fixed design $\mathbf{X}$ and $\mathbf{Z}$ throughout this paper.

3.1. *MSE grouping sensitivity.*

We first define a incongruity measure $d(\boldsymbol{\beta}, \boldsymbol{\beta}')$ between two groupings that correspond to $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$, respectively.

DEFINITION 3.1 (Degree of grouping incongruity). For any $\boldsymbol{\beta}$, $\boldsymbol{\beta}'$ such that $|\mathbb{G}(\boldsymbol{\beta})| \le |\mathbb{G}(\boldsymbol{\beta}')|$, define

$$d(\boldsymbol{\beta}, \boldsymbol{\beta}') := \min_{f \in \mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\beta}')} \left| \left(\bigcup_{\mathcal{G}_2 \in \mathbb{G}(\boldsymbol{\beta}')} \mathcal{G}_2 \right) \Big\backslash \left(\bigcup_{\mathcal{G}_1 \in \mathbb{G}(\boldsymbol{\beta})} \{\mathcal{G}_1 \cap f(\mathcal{G}_1)\} \right) \right|, \tag{3.1}$$

where $\mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\beta}') := \{f \text{ is injective} : \mathbb{G}(\boldsymbol{\beta}) \to \mathbb{G}(\boldsymbol{\beta}')\}$.

Definition 3.1 may be explained as follows. Note that $\bigcup_{\mathcal{G}_1 \in \mathbb{G}(\boldsymbol{\beta})} \{\mathcal{G}_1 \cap f(\mathcal{G}_1)\}$ collects all the variables that are consistently labeled by $\mathbb{G}(\boldsymbol{\beta})$ and $\mathbb{G}(\boldsymbol{\beta}')$ under mapping f, while $\bigcup_{\mathcal{G}_2 \in \mathbb{G}(\boldsymbol{\beta}')} \mathcal{G}_2$ collects the support variables of $\boldsymbol{\beta}'$. Therefore, $\{\bigcup_{\mathcal{G}_2 \in \mathbb{G}(\boldsymbol{\beta}')} \mathcal{G}_2\} \backslash \{\bigcup_{\mathcal{G}_1 \in \mathbb{G}(\boldsymbol{\beta})} \mathcal{G}_1 \cap f(\mathcal{G}_1)\}$ gives rise to a collection of all the variables with inconsistent group labels in $\mathbb{G}(\boldsymbol{\beta})$ and $\mathbb{G}(\boldsymbol{\beta}')$ under mapping f. Moreover, $d(\boldsymbol{\beta}, \boldsymbol{\beta}')$ measures the minimum number of unmatched variables in the support of $\boldsymbol{\beta}'$ among all mappings $f \in \mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\beta}')$. Figure 1 illustrates a specific setup with two mappings of possible group labels, $f^{(1)}$ and $f^{(2)}$ in $\mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\beta}')$. Clearly, $f^{(1)}$ gives three inconsistent group labels (three crosses) between $\mathbb{G}(\boldsymbol{\beta})$ and $\mathbb{G}(\boldsymbol{\beta}')$, while $f^{(2)}$ gives four. Therefore, $f^{(1)}$ minimizes the objective in (3.1), and thus $d(\boldsymbol{\beta}, \boldsymbol{\beta}') = 3$.

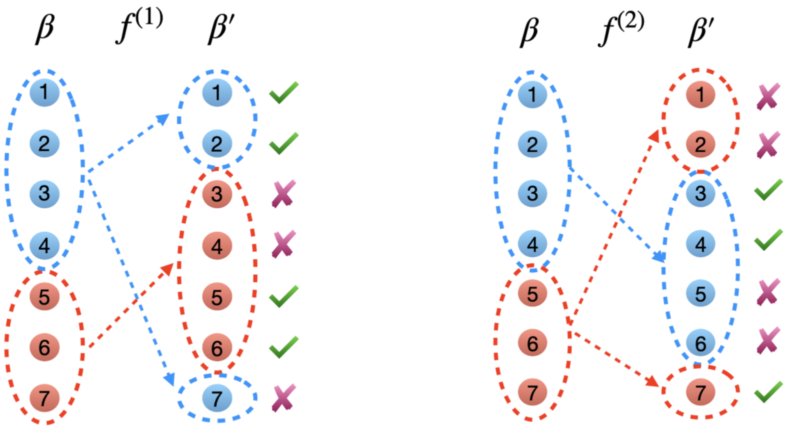


FIG. 1. *Illustration of the grouping maps and grouping incongruity. Here,* $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathbb{R}^7$, $|\mathbb{G}(\boldsymbol{\beta})| = |\mathbb{G}(\boldsymbol{\beta}')| = 2$ *and* $|\mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\beta}')| = 2$. *We write* $\mathcal{F}(\boldsymbol{\beta}, \boldsymbol{\beta}') = \{f^{(1)}, f^{(2)}\}$ *and illustrate these two maps on the left and right panels, respectively. For clarity, we let* $\mathcal{G}$ *and* $f(\mathcal{G})$ *share the same color for any* $\mathcal{G} \in \mathbb{G}(\boldsymbol{\beta})$. *On the right border of each panel, for each feature, we use a check (cross) to indicate the consistency (inconsistency) between* $\mathbb{G}(\boldsymbol{\beta})$ *and* $\mathbb{G}(\boldsymbol{\beta}')$ *according to the given grouping map. Then* $d(\boldsymbol{\beta}, \boldsymbol{\beta}') = 3$.

Next, we define a sensitivity measure via mean squared error (MSE) with respect to degree of grouping incongruity, which plays a key role in quantifying the difficulty of identifying the true grouping in the theoretical analysis.

DEFINITION 3.2 (MSE grouping sensitivity). Under model (1.1) with the true parameter $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*\top}, \boldsymbol{\alpha}^{*\top})^\top \in \mathbb{R}^{p+q}$, define

$$c_{\min} \equiv c_{\min}(\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{Z}) := \min_{\substack{\boldsymbol{\theta} \in \boldsymbol{\Theta}(|\mathbb{G}(\boldsymbol{\beta}^*)|, \|\boldsymbol{\beta}^*\|_0) \\ \mathbb{G}(\boldsymbol{\beta}) \neq \mathbb{G}(\boldsymbol{\beta}^*)}} \frac{\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \mathbf{Z}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)\|_2^2}{n \max(d(\boldsymbol{\beta}, \boldsymbol{\beta}^*), 1)}, \tag{3.2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top \in \mathbb{R}^{p+q}$ is a parameter.

In words, $c_{\min}$ is the resulting minimal increase of MSE from a false grouping result. A small $c_{\min}$ suggests low sensitivity of MSE to false grouping, and consequently, it is difficult to restore the true grouping in the analysis. Given a grouping structure $\mathbb{G}(\boldsymbol{\beta})$, define

$$\mathbf{X}_{\mathbb{G}(\boldsymbol{\beta})} := \left(\sum_{k \in \mathcal{G}(\boldsymbol{\beta}; \gamma_1)} \mathbf{X}_k, \ldots, \sum_{k \in \mathcal{G}(\boldsymbol{\beta}; \gamma_{|\mathbb{G}(\boldsymbol{\beta})|})} \mathbf{X}_k \right),$$

where we collapse matrix $\mathbf{X}$ groupwise by summing up its columns in the same group according to $\mathbb{G}(\boldsymbol{\beta})$. Let $\mathbf{D}(\boldsymbol{\beta}) \in \mathbb{R}^{|\mathbb{G}(\boldsymbol{\beta})| \times |\mathbb{G}(\boldsymbol{\beta})|}$ denote the diagonal matrix with the ith diagonal element $\{D(\boldsymbol{\beta})\}_{ii} = |\mathcal{G}(\boldsymbol{\beta}; \gamma_i)|$, which is the size of the group of parameters equal to γ_i. For a given $\boldsymbol{\beta}$, let

$$\widetilde{\mathbf{X}}(\boldsymbol{\beta}) := (\mathbf{X}_{\mathbb{G}(\boldsymbol{\beta})}\{\mathbf{D}(\boldsymbol{\beta})\}^{-1/2}, \mathbf{Z}) \in \mathbb{R}^{n \times (|\mathbb{G}(\boldsymbol{\beta})| + q)}, \tag{3.3}$$

and the projection operator takes the form $\mathbf{P}(\boldsymbol{\beta}) := \widetilde{\mathbf{X}}(\boldsymbol{\beta})\{\widetilde{\mathbf{X}}(\boldsymbol{\beta})^\top \widetilde{\mathbf{X}}(\boldsymbol{\beta})\}^+ \widetilde{\mathbf{X}}(\boldsymbol{\beta})^\top$. For convenience, we denote the true signal by $\boldsymbol{\mu}^* = (\mathbf{X}, \mathbf{Z})\binom{\boldsymbol{\beta}^*}{\boldsymbol{\alpha}^*}$. To further interpret the term $c_{\min}$, we now introduce an equivalent definition of $c_{\min}$ through the sum of squares of residuals incurred by incorrect grouping as follows:

$$c_{\min}(\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{Z}) := \min_{\substack{\boldsymbol{\theta} \in \boldsymbol{\Theta}(|\mathbb{G}(\boldsymbol{\beta}^*)|, \|\boldsymbol{\beta}^*\|_0) \\ \mathbb{G}(\boldsymbol{\beta}) \neq \mathbb{G}(\boldsymbol{\beta}^*)}} \frac{\|\{\mathbf{I} - \mathbf{P}(\boldsymbol{\beta})\}\boldsymbol{\mu}^*\|_2^2}{n \max(d(\boldsymbol{\beta}, \boldsymbol{\beta}^*), 1)}. \tag{3.4}$$

Definition (3.4) will be extended in Section 3.5 to another MSE type of grouping sensitivity in the case of overestimated group number and sparsity. The equivalency between (3.2) and (3.4) can be easily verified. Note the fact that $\mathbf{P}(\boldsymbol{\beta})$ and $d(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ depend only on the group structure $\mathbb{G}(\boldsymbol{\beta})$. Thus, the equivalency naturally holds by the following operations: first, fix a group structure under which the MSE is minimized by the least squares method; then take the minimum over all possible group structures in Definition (3.2).

3.2. *Sufficient condition.* Here, we present the sufficient conditions for L_0-Fusion to achieve grouping consistency. We first define the oracle least squares estimator.

DEFINITION 3.3 (Oracle least squares estimator). Given the true coefficient $\boldsymbol{\beta}^*$, the oracle least squares estimator $\hat{\boldsymbol{\theta}}^{\mathrm{ol}} = (\hat{\boldsymbol{\beta}}^{\mathrm{ol}\top}, \hat{\boldsymbol{\alpha}}^{\mathrm{ol}\top})^\top$ is defined as

$$\hat{\boldsymbol{\theta}}^{\mathrm{ol}} := \operatorname*{argmin}_{\boldsymbol{\theta}:\mathbb{G}(\boldsymbol{\beta})=\mathbb{G}(\boldsymbol{\beta}^*)} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\alpha}\|_2^2.$$

More specifically, define

$$(\hat{\boldsymbol{\gamma}}^{\mathrm{ol}\top}, \hat{\boldsymbol{\alpha}}^{\mathrm{ol}\top})^\top := (\hat{\gamma}_1^{\mathrm{ol}}, \dots, \hat{\gamma}_{K_0}^{\mathrm{ol}}, \hat{\boldsymbol{\alpha}}^{\mathrm{ol}\top})^\top = \operatorname*{argmin}_{(\boldsymbol{\gamma}^\top, \boldsymbol{\alpha}^\top)^\top \in \mathbb{R}^{K_0+q}} \|\mathbf{Y} - \mathbf{X}_{\mathbb{G}(\boldsymbol{\beta}^*)}\boldsymbol{\gamma} - \mathbf{Z}\boldsymbol{\alpha}\|_2^2.$$

In the oracle solution $\hat{\boldsymbol{\beta}}^{\mathrm{ol}} = (\hat{\beta}_1^{\mathrm{ol}}, \dots, \hat{\beta}_p^{\mathrm{ol}})^\top$, $\hat{\beta}_j^{\mathrm{ol}}$ is equal to either $\hat{\gamma}_k^{\mathrm{ol}}$ if $j \in \mathbb{G}(\boldsymbol{\beta}^*; \gamma_k^*)$, or $\hat{\beta}_j^{\mathrm{ol}}$ is 0 if $j \in \mathbb{G}(\boldsymbol{\beta}^*; 0)$, and $K_0 = |\mathbb{G}(\boldsymbol{\beta}^*)|$.

For any estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\alpha}}^\top)^\top$ of $\boldsymbol{\theta}^*$, define the 0-1 grouping risk $\mathcal{L}_g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}^*) := \mathbb{P}(\mathbb{G}(\hat{\boldsymbol{\beta}}) \neq \mathbb{G}(\boldsymbol{\beta}^*))$. Denote the solution to the L_0-Fusion problem (2.1) by $\hat{\boldsymbol{\theta}}^{\mathrm{g}} = (\hat{\boldsymbol{\beta}}^{\mathrm{g}\top}, \hat{\boldsymbol{\alpha}}^{\mathrm{g}\top})^\top$. Denote $\|\boldsymbol{\beta}^*\|_0$ by s_0 for convenience. Theorem 3.4 below says that the MIO solution $\hat{\boldsymbol{\theta}}^{\mathrm{g}}$ consistently recovers the oracle solution $\hat{\boldsymbol{\theta}}^{\mathrm{ol}}$ if the MSE grouping sensitivity satisfies the condition, $c_{\min} \gtrsim \log(pK_0)/n$, and the following Condition 3.1 holds.

CONDITION 3.1 (Sub-Gaussian random error). For model (1.1), the random errors ε's are independent sub-Gaussian with mean zero and ψ_2-norm bounded by σ.

THEOREM 3.4. *If Condition* 3.1 *holds, under $K = K_0$ and $s = s_0$ in* (2.1), *we have*

$$\mathcal{L}_g(\hat{\boldsymbol{\theta}}^{\mathrm{g}}; \boldsymbol{\theta}^*) \le 4\exp\left[-\frac{3n}{200\sigma^2}\left\{c_{\min} - \frac{\sigma^2}{n}(134\log(pK_0) + 220)\right\}\right].$$

This implies that when $c_{\min} \ge \frac{\sigma^2}{n}\{d_1 \log(pK_0) + 220\}$ for some universal constant $d_1 > 134$, $\hat{\boldsymbol{\theta}}^{\mathrm{g}}$ consistently reconstructs $\hat{\boldsymbol{\theta}}^{\mathrm{ol}}$, namely $\mathcal{L}_g(\hat{\boldsymbol{\theta}}^{\mathrm{g}}; \boldsymbol{\theta}^*) \to 0$ as $n, p \to \infty$.

The proof of Theorem 3.4 is inspired by the proofs of Theorem 2 in Zhu, Shen and Pan (2013) and Theorem 2 in Shen et al. (2013). Note that in Section 3.3 we show that a lower bound of the same order for $c_{\min}$ is necessary to achieve the grouping consistency. In addition, the upper bound in the above theoretical guarantee is obtained for the L_0-Fusion estimator under the true values K_0 and s_0. Although they are unknown in practice, we may employ effective tuning parameter selection methods to learn their values satisfactorily. As illustrated by our simulation experiments in Section 6, optimizing the Bayesian Information Criterion (BIC) with respect to K and s in (2.1) enables us to select the model with the true K_0 and s_0, with high probability. Moreover, in Section 3.5, we relax this ideal setting to a commonly encountered "over-identification" scenario in that tuned values K and s are overestimated and satisfy $K \geq K_0$ and $s \geq s_0$. Consequently, we propose a new concept of *sure partition* to extend Theorem 3.4 in order to establish theoretical guarantees for the L_0-Fusion estimator.

For the sake of computational feasibility and efficiency, we suggest performing a preliminary dimension reduction step to speed up solving (2.1) numerically. This step is not required in the theoretical analysis. In this paper, we adopt the so-called "*screen then group* (StG)" strategy. To elucidate, we define a sure screening event $\mathcal{E}$ as $\{\mathcal{S}^0 \subseteq \widetilde{\mathcal{S}}\}$ with $\widetilde{\mathcal{S}}$ being the resulting parameter set from the screening operation. Let $\hat{\boldsymbol{\theta}}^{\mathrm{sg}} \in \mathbb{R}^{p+q}$ denote an StG solution that consists of the solution to the L_0-Fusion problem in (2.1) on the reduced design matrix $\mathbf{X}_{\widetilde{\mathcal{S}}}$ and a set of zeros corresponding to the screened out features. Corollary 3.1 states that as long as $\mathcal{E}$ holds with high probability and groups are reasonably separable, $\hat{\boldsymbol{\theta}}^{\mathrm{sg}}$ can recover $\hat{\boldsymbol{\theta}}^{\mathrm{ol}}$ with high probability. Note that screening is a separate operation from grouping, which aims to reduce p, the dimension of features X_j's, so that the reduced design matrix contains still the true signals (i.e., K_0 and s_0) prior to the subsequent operation of grouping. Many variable screening techniques provably yield a sure screening set under reasonable assumptions on both signal and design, including Sure Independence Screening (Fan and Lv (2008), Theorem 1), LASSO (Wainwright (2019), Theorem 7.21) and CoSaMP (Guo, Zhu and Fan (2020), Theorem 3.1). In this paper, we choose CoSaMP as the screening technique given its robustness against design collinearity (Guo, Zhu and Fan (2020)).

COROLLARY 3.1. *Under $K = K_0$, $s = s_0$ and $c_{\min} > \frac{\sigma^2}{n}(134\log(pK_0) + 220)$, for any sure screening event $\mathcal{E}$, we have*

$$\mathcal{L}_g(\hat{\boldsymbol{\theta}}^{\mathrm{sg}}; \boldsymbol{\theta}^*) \leq 4\exp\left[-\frac{3n}{200\sigma^2}\left\{c_{\min} - \frac{\sigma^2}{n}(134\log(pK_0) + 220)\right\}\right] + \mathbb{P}(\mathcal{E}^c).$$

PROOF. First, we claim that

$$\{\hat{\boldsymbol{\theta}}^{\mathrm{g}} = \hat{\boldsymbol{\theta}}^{\mathrm{ol}}\} \cap \mathcal{E} \subseteq \{\hat{\boldsymbol{\theta}}^{\mathrm{sg}} = \hat{\boldsymbol{\theta}}^{\mathrm{ol}}\}, \tag{3.5}$$

where event $\{\hat{\boldsymbol{\theta}}^{\mathrm{g}} = \hat{\boldsymbol{\theta}}^{\mathrm{ol}}\}$ means that the L_0-Fusion obtains the oracle solution under a design matrix $\mathbf{X}$, and when this event occurs, $\mathrm{supp}(\hat{\boldsymbol{\theta}}^{\mathrm{g}}) = \mathrm{supp}(\hat{\boldsymbol{\theta}}^{\mathrm{ol}}) = \mathcal{S}^0$. Likewise event $\{\hat{\boldsymbol{\theta}}^{\mathrm{sg}} = \hat{\boldsymbol{\theta}}^{\mathrm{ol}}\}$ means that the StG-based L_0-Fusion obtains the oracle solution under a reduced design matrix $\mathbf{X}_{\widetilde{\mathcal{S}}}$. Note that the sure screening $\mathcal{E}$ only removes some of the null signals, and the resulting reduced design matrix $\mathbf{X}_{\widetilde{\mathcal{S}}}$ contains all the true signals, namely $\{\mathcal{S}^0 \subseteq \widetilde{\mathcal{S}}\}$. Thus, the occurrence of event $\{\hat{\boldsymbol{\theta}}^{\mathrm{g}} = \hat{\boldsymbol{\theta}}^{\mathrm{ol}}\}$ implies that $\hat{\boldsymbol{\theta}}^{\mathrm{g}}$ is one of possible solutions within the scope of the StG operation. Thus, when both events $\{\hat{\boldsymbol{\theta}}^{\mathrm{g}} = \hat{\boldsymbol{\theta}}^{\mathrm{ol}}\}$ and $\{\mathcal{S}^0 \subseteq \widetilde{\mathcal{S}}\}$ occur, event $\{\hat{\boldsymbol{\theta}}^{\mathrm{sg}} = \hat{\boldsymbol{\theta}}^{\mathrm{ol}}\}$ will occur with $\hat{\boldsymbol{\theta}}^{\mathrm{sg}} = \hat{\boldsymbol{\theta}}^{\mathrm{g}} = \hat{\boldsymbol{\theta}}^{\mathrm{ol}}$. It follows from the Bonferroni inequality that

$$\begin{aligned}\mathbb{P}(\hat{\boldsymbol{\theta}}^{\mathrm{sg}} = \hat{\boldsymbol{\theta}}^{\mathrm{ol}}) &\geq \mathbb{P}(\{\hat{\boldsymbol{\theta}}^{\mathrm{g}} = \hat{\boldsymbol{\theta}}^{\mathrm{ol}}\} \cap \mathcal{E}) = 1 - \mathbb{P}(\{\hat{\boldsymbol{\theta}}^{\mathrm{g}} \neq \hat{\boldsymbol{\theta}}^{\mathrm{ol}}\} \cup \mathcal{E}^c) \\ &\geq 1 - \mathbb{P}(\hat{\boldsymbol{\theta}}^{\mathrm{g}} \neq \hat{\boldsymbol{\theta}}^{\mathrm{ol}}) - \mathbb{P}(\mathcal{E}^c).\end{aligned}$$

The conclusion immediately follows by combining this inequality with Theorem 3.4. □

Corollary 3.1 suggests that applying the screening operation prior to grouping can reduce the complexity of L_0-fusion in the case when we correctly identify $K = K_0$, $s = s_0$. There is a small price of $\mathbb{P}(\mathcal{E}^c)$ owing to the screening operation in StG leads to a slight loss in statistical accuracy guarantee as a trade-off for the gain of computational benefits.

3.3. *Necessary condition.* For a constant $c_* > 0$, consider the following parameter subspace of $\boldsymbol{\Theta}(K_0, s_0)$:

$$\boldsymbol{\Theta}_c(K_0, s_0, c_*) := \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \boldsymbol{\Theta}(K_0, s_0), c_{\min}(\boldsymbol{\theta}, \mathbf{X}, \mathbf{Z}) \geq c_*\}.$$

In Theorem 3.5 below, we aim to establish a lower bound for the minimax 0-1 grouping risk $\mathcal{L}_g(\cdot)$ over $\boldsymbol{\Theta}_c(K_0, s_0, c_*)$. As a result, the lower bound of $c_{\min}$ in Theorem 3.4 becomes necessary to simultaneously achieve selection and grouping consistency. To proceed, we postulate the following two conditions on the design matrix, both of which are indeed routinely assumed in the high-dimensional statistics literature.

CONDITION 3.2. There exists a constant $\kappa_u > 0$ such that $\max_{j\in[p]} \|\mathbf{X}_j\|_2/\sqrt{n} \leq \kappa_u$.

CONDITION 3.3 (Weak sparse Riesz condition). There exists a constant $\kappa_l > 0$ such that

$$\frac{1}{\sqrt{n}} \left\| (\mathbf{X}, \mathbf{Z}) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} \right\|_2 \geq \kappa_l \left\| \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} \right\|_2,$$

for any $(\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top \in \mathbb{R}^{p+q} \backslash \{\mathbf{0}\}$ satisfying $\|\boldsymbol{\beta}\|_0 \leq 2K_0$.

THEOREM 3.5. *For model* (1.1), *suppose that errors* $\{\epsilon_i\}_{i\in[n]}$ *are i.i.d.* $\mathcal{N}(0, \sigma^2)$ *and that Conditions* 3.2 *and* 3.3 *hold. Under* $K_0 \geq 1$, $p \geq s_0 \geq K_0$, *for any* $c_* > 0$, *we have the minmax lower bound given by*

$$\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_c(K_0, s_0, c_*)} \mathcal{L}_g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}^*) \geq 1 - \frac{20n\kappa_u^2 \kappa_l^{-2} c_* + \sigma^2 \log(2)}{\sigma^2 \log(\lfloor \frac{K_0+3}{4} \rfloor p)}.$$

Consequently, if there exists a universal constant $\rho \in (0, 1)$ *such that*

$$c_* \leq \frac{\sigma^2 \kappa_l^2 \{\rho \log(\lfloor \frac{K_0+3}{4} \rfloor p) - \log(2)\}}{20n\kappa_u^2},$$

then we have $\inf_{\hat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_c(K_0, s_0, c_*)} \mathcal{L}_g(\hat{\boldsymbol{\theta}}; \boldsymbol{\theta}^*) \geq 1 - \rho > 0$.

3.4. *Characterizing* $c_{\min}$ *by the homogeneous coefficient group separation.* So far, the MSE grouping sensitivity $c_{\min}$ reflects an intrinsic difficulty of the L_0-Fusion problem in terms of achieving the grouping consistency. To further interpret and illustrate $c_{\min}$ intuitively, we now introduce a quantity, denoted by $\gamma_{\min}$, to depict the separation between homogeneous coefficient groups. This metric of separability is given by the minimum distance between different groups of model parameters.

DEFINITION 3.6 (Minimum group gap). $\gamma_{\min} = \min\{\min_{\substack{j, j' \in [p] \\ \beta_j^* \neq \beta_{j'}^*}} |\beta_j^* - \beta_{j'}^*|, \min_{\substack{j \in [p] \\ \beta_j^* \neq 0}} |\beta_j^*|\}$.

We show that $c_{\min}$ and $\gamma_{\min}$ are closely related in various cases; thus, theoretical guarantees for grouping consistency may be intuitively presented and interpreted through the corresponding conditions of $\gamma_{\min}$ in the rest of this paper. Clearly, the larger $\gamma_{\min}$, the easier to distinguish different groups of parameters and identify the true group structure. Arguably, the relationship between $\gamma_{\min}$ and $c_{\min}$ help to better understand conditions required to achieve grouping consistency. For this investigation, we actually need a stronger condition than Condition 3.3 on the design matrix, given as follows.

CONDITION 3.4 (Strong sparse Reisz condition). There exists a constant $\kappa_l' \in (0, \infty)$ such that

$$\frac{1}{\sqrt{n}} \left\| (\mathbf{X}, \mathbf{Z}) \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} \right\|_2 \geq \kappa_l' \left\| \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{pmatrix} \right\|_2,$$

for any $(\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top \in \mathbb{R}^{p+q} \backslash \{\mathbf{0}\}$ such that $\|\boldsymbol{\beta}\|_0 \leq 2s_0$.

Of note, Condition 3.4 has been routinely assumed in the high-dimensional statistics literature. Also, it is easy to see that Condition 3.4 implies Condition 3.3. Below we study the two examples with two groups of coefficients to gain some primary insights.

EXAMPLE 3.7 (Balanced groups). Consider $K_0 = 2$, $p = s_0$ and an arbitrary q. Assume that $c^* := |\mathcal{G}(\boldsymbol{\beta}^*; \gamma_1^*)|/s_0 \in (0, 1)$. Under Condition 3.4, we have

$$c_{\min} \geq \min\left\{ c^*, 1 - c^*, \frac{2}{c^* + 2}, \frac{2}{3 - c^*}, \frac{2(1 - c^*)}{3 - 2c^*}, \frac{2c^*}{1 + 2c^*} \right\} \kappa_l'^2 \gamma_{\min}^2.$$

Example 3.7 implies that when there are two disjoint groups of relatively balanced sizes, if $\gamma_{\min} \gtrsim \sigma\{\log(pK_0)/n\}^{1/2}$, where the lower bound accords with the rate of the signal strength in the β-min condition (Wainwright (2009)), then $c_{\min} \gtrsim \sigma^2 \log(pK_0)/n$. Moreover, applying Theorem 3.4 leads to grouping consistency of $\hat{\boldsymbol{\theta}}^{\mathrm{g}}$ in this balanced case. In the next example with the group sizes being very unbalanced, a more stringent lower bound for $\gamma_{\min}$ is required to achieve grouping consistency.

EXAMPLE 3.8 (Unbalanced groups). Suppose $p = s_0$, $q = 0$, $K_0 = 2$, $|\mathcal{G}(\boldsymbol{\beta}^*, \gamma_1^*)| = 1$, $|\mathcal{G}(\boldsymbol{\beta}^*, \gamma_2^*)| = s_0 - 1$, $\gamma_1^* = \gamma$ and $\gamma_2^* = 2\gamma$. It follows that $\gamma_{\min} = \gamma$. Assume that there exists a universal constant $\kappa_u' \in (0, \infty)$ such that the square-root of the maximum eigenvalue of the sample covariance is bounded above, $\lambda_{\max}^{1/2}(\mathbf{X}^\top\mathbf{X}/n) \leq \kappa_u'$. Then we have $c_{\min} \leq \kappa_u'^2 \gamma_{\min}^2 / \lceil s_0/2 \rceil$.

In Example 3.8, to yield $c_{\min} \gtrsim \sigma^2 \log(pK_0)/n$ we need at least $\gamma_{\min} \gtrsim \sigma\{s_0 \log(pK_0)/n\}^{1/2}$, which includes an extra $s_0^{1/2}$ factor in comparison to that in Example 3.7. Proofs of Examples 3.7 and 3.8 are in Sections 2.3 and 2.4 of the Supplementary Material (Wang et al. (2024)). Theorem 3.9 below presents a general result for the connection between $c_{\min}$ and $\gamma_{\min}$.

THEOREM 3.9. *Suppose Condition* 3.4 *holds for some* $\kappa_l' \in (0, \infty)$. *Then* $c_{\min} \geq \kappa_l'^2 \gamma_{\min}^2 / (4s_0)$. *Moreover, if* $\gamma_{\min} \gtrsim \sigma\{s_0 \log(pK_0)/n\}^{1/2}$, *then* $c_{\min} \gtrsim \sigma^2 \log(pK_0)/n$.

It is interesting to note that Ke, Fan and Wu (2015) introduced a similar group gap quantity (referred to as b_n therein), where to achieve group consistency in Ke, Fan and Wu (2015) the $\gamma_{\min}$ term needs to satisfy $\gamma_{\min}^2 \gg \max_k\{|\mathcal{G}(\boldsymbol{\beta}^*; \gamma_k^*)|\}\sigma^2 \log(n \vee p)/n$. Under a highly challenging setup as in Example 3.8, Theorem 3.9 agrees with their result in Ke, Fan and Wu (2015) where $\max_k\{|\mathcal{G}(\boldsymbol{\beta}^*; \gamma_k^*)|\} = s_0 - 1$ holds.

3.5. *Sufficient conditions with overestimated group number and sparsity.* In practice, both K_0 and s_0 are unknown and need to be determined. Here, we investigate the L_0-Fusion estimation with overestimated group number and sparisity, denoted by $K^\dagger$ and $s^\dagger$ that are obtained by a tuning step that typically results in $K = K^\dagger \geq K_0$ and $s = s^\dagger \geq s_0$. In this case of overidentification, we change our analytic goal from a full restoring of the true grouping structure to a realistically achievable optimality in terms of *sure partition*.

DEFINITION 3.10 (Sure partition estimator). An statistic $\hat{\boldsymbol\beta}$ is called a sure partition estimator of $\boldsymbol\beta^*$ if (i) the null signals given by $\hat{\boldsymbol\beta}$ form a subset of the true null signals, namely $\mathcal{G}(\hat{\boldsymbol\beta}; 0) \subseteq \mathcal{G}(\boldsymbol\beta^*; 0)$, and (ii) sure partition $\mathbb{G}(\hat{\boldsymbol\beta}) \cup \{\mathcal{G}(\hat{\boldsymbol\beta}; 0)\}$ is a finer partition of $[p]$ than the true grouping $\mathbb{G}(\boldsymbol\beta^*) \cup \{\mathcal{G}(\boldsymbol\beta^*; 0)\}$.

In words, a sure partition corresponds to a situation where the groups given by estimator $\hat{\boldsymbol\beta}$ are resulted from further partitioning of the true groups. For any sure partition estimator $\hat{\boldsymbol\beta}$, the true signal $\boldsymbol\mu^*$ is in the column space spanned by $\widetilde{\mathbf{X}}(\hat{\boldsymbol\beta})$ (see (3.3)), or equivalently, $\mathbf{P}(\hat{\boldsymbol\beta})\boldsymbol\mu^* = \boldsymbol\mu^*$. This implies no shrinkage of the feature space with respect to the design matrix of an overientified group structure. Thus, the least squares estimator based on a resulting design matrix from sure partition $\mathbb{G}(\hat{\boldsymbol\beta})$ enjoys both unbiasedness and parsimony. These properties make the sure partition estimator a desirable choice to deal with the L_0-Fusion problem along with tuning on group size and sparsity. We denote the collection of all sure partition estimators of $(\boldsymbol\beta^{*\top}, \boldsymbol\alpha^*)^\top$ with a group size $K^\dagger \ge K_0$ and sparsity $s^\dagger \ge s_0$ by

$$\Pi(\boldsymbol\beta^*, K^\dagger, s^\dagger) := \{(\boldsymbol\beta^\top, \boldsymbol\alpha^\top)^\top \mid \boldsymbol\beta \text{ is a sure partition estimate of } \boldsymbol\beta^*,\\ (\boldsymbol\beta^\top, \boldsymbol\alpha^\top)^\top \in \boldsymbol\Theta(K^\dagger, s^\dagger), \boldsymbol\alpha \in \mathbb{R}^q\}.$$

Below we provide a toy example to facilitate understanding of sure partition estimator.

EXAMPLE 3.11. Consider $\boldsymbol\beta^* \in \mathbb{R}^{20}$ and $\mathbb{G}(\boldsymbol\beta^*) = \{\{1,2\},\{3,4\}\}$. Obviously, $K_0 = 2$ and $s_0 = 4$. At $K^\dagger = 4$, $\{\mathbb{G}(\boldsymbol\beta) : \boldsymbol\theta = (\boldsymbol\beta^\top, \boldsymbol\alpha^\top)^\top \in \Pi(\boldsymbol\beta^*, 4, s^\dagger) \text{ for some } s^\dagger \ge 6\}$ includes

$$\{\{1,2\},\{3,4\},\mathcal{G}_1,\mathcal{G}_2\}, \qquad \{\{1,2\},\{3\},\{4\},\mathcal{G}\} \quad \text{and} \quad \{\{1\},\{2\},\{3,4\},\mathcal{G}\}$$

for any nonempty subset $\mathcal{G} \subset \{5,6,\dots,20\}$ such that $|\mathcal{G}| \le s^\dagger - 4$ and for any nonempty sets $\mathcal{G}_1, \mathcal{G}_2 \subset \{5,6,\dots,20\}$ such that $\mathcal{G}_1 \cap \mathcal{G}_2 = \varnothing$ and $|\mathcal{G}_1| + |\mathcal{G}_2| \le s^\dagger - 4$. Of note, $\mathbb{G}(\boldsymbol\beta)$ only collects the groups of nonzero coefficients. For example, when $\mathbb{G}(\boldsymbol\beta) = \{\{1,2\},\{3,4\},\mathcal{G}_1,\mathcal{G}_2\}$, we obtain its support as $\operatorname{supp}(\boldsymbol\beta) = \{1,2,3,4\} \cup \mathcal{G}_1 \cup \mathcal{G}_2$.

Next, to discuss optimality under sure partition, we introduce a new measure $d^\dagger(\boldsymbol\beta; \boldsymbol\beta^*, K^\dagger, s^\dagger)$ that characterizes the incongruity of $\boldsymbol\beta$ over the set of sure partition estimators of $\boldsymbol\beta^*$. This leads to a modified version of the MSE sensitivity to grouping incongruity, denoted by $c^\dagger_{\min}$.

DEFINITION 3.12 (Degree of grouping incongruity with sure partition). Given any $K^\dagger \ge K_0$ and $s^\dagger \ge s_0$, for any $\boldsymbol\theta = (\boldsymbol\beta^\top, \boldsymbol\alpha^\top)^\top \in \boldsymbol\Theta(K^\dagger, s^\dagger)$, define

$$d^\dagger(\boldsymbol\beta; \boldsymbol\beta^*, K^\dagger, s^\dagger) := \min_{\boldsymbol\theta^\dagger \in \{\boldsymbol\theta^\dagger \in \Pi(\boldsymbol\beta^*, K^\dagger, s^\dagger) : |\mathbb{G}(\boldsymbol\beta^\dagger)| \ge |\mathbb{G}(\boldsymbol\beta)|\}} d(\boldsymbol\beta, \boldsymbol\beta^\dagger).$$

DEFINITION 3.13 (MSE sensitivity to incongruity with sure partition). Given $\boldsymbol\theta^*$, for any $K^\dagger \ge K_0$ and $s^\dagger \ge s_0$, define

$$\begin{aligned} c^\dagger_{\min}(\boldsymbol\theta^*, \mathbf{X}, \mathbf{Z}, K^\dagger, s^\dagger) &:= \min_{\substack{\boldsymbol\theta = (\boldsymbol\beta^\top, \boldsymbol\alpha^\top)^\top \in \boldsymbol\Theta(K^\dagger, s^\dagger)\\ \boldsymbol\theta \notin \Pi(\boldsymbol\beta^*, K^\dagger, s^\dagger)}} \frac{\|\mathbf{X}(\boldsymbol\beta - \boldsymbol\beta^*) + \mathbf{Z}(\boldsymbol\alpha - \boldsymbol\alpha^*)\|_2^2}{n \max(d^\dagger(\boldsymbol\beta; \boldsymbol\beta^*, K^\dagger, s^\dagger), 1)} \\ &= \min_{\substack{\boldsymbol\theta = (\boldsymbol\beta^\top, \boldsymbol\alpha^\top)^\top \in \boldsymbol\Theta(K^\dagger, s^\dagger)\\ \boldsymbol\theta \notin \Pi(\boldsymbol\beta^*, K^\dagger, s^\dagger)}} \frac{\|\{\mathbf{I} - \mathbf{P}(\boldsymbol\beta)\}\boldsymbol\mu^*\|_2^2}{n \max(d^\dagger(\boldsymbol\beta; \boldsymbol\beta^*, K^\dagger, s^\dagger), 1)}. \end{aligned}$$

Note that when $K^\dagger = K_0$ and $s^\dagger = s_0$, we have $c^\dagger_{\min}(\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{Z}, K_0, s_0) = c_{\min}(\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{Z})$. We use $\mathcal{L}^\dagger_g(\hat{\boldsymbol{\theta}}^{\mathrm{g}}; \boldsymbol{\theta}^*, K^\dagger, s^\dagger)$ to denote the 0-1 sure partition risk $\mathbb{P}(\hat{\boldsymbol{\theta}}^{\mathrm{g}} \notin \Pi(\boldsymbol{\beta}^*, K^\dagger, s^\dagger))$. Theorem 3.14 affirms that the L_0-Fusion solution $\hat{\boldsymbol{\theta}}^{\mathrm{g}}$ under the overestimated group size and sparsity can achieve sure partition with high probability when the group incongruity satisfies $c^\dagger_{\min}(\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{Z}, K^\dagger, s^\dagger) \gtrsim \log(pK^\dagger)/n$.

THEOREM 3.14. *Let $K = K^\dagger \geq K_0$ and $s = s^\dagger \geq s_0$ in the L_0-Fusion problem where a solution $\hat{\boldsymbol{\theta}}^{\mathrm{g}}$ is obtained. Under Condition 3.1, we have*

$$\mathcal{L}^\dagger_g(\hat{\boldsymbol{\theta}}^{\mathrm{g}}; \boldsymbol{\theta}^*, K^\dagger, s^\dagger) \leq 4\exp\biggl[-\frac{3n}{200\sigma^2}\biggl\{c^\dagger_{\min}(\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{Z}, K^\dagger, s^\dagger) - \frac{\sigma^2}{n}\bigl(134\log(pK^\dagger) + 220\bigr)\biggr\}\biggr].$$

Moreover, when $c^\dagger_{\min}(\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{Z}, K^\dagger, s^\dagger) \geq \frac{\sigma^2}{n}\{d_1 \log(pK^\dagger) + 220\}$ for some universal constant $d_1 > 134$, $\hat{\boldsymbol{\theta}}^{\mathrm{g}}$ achieves sure partition with probability going to one as $n, p \to \infty$.

In the setting of sure partition with estimates $K^\dagger$ and $s^\dagger$, when screening is applied prior to grouping, similar to Corollary 3.1, a certain small loss appears to slightly weaken the upper bound given in Theorem 3.14. Below Corollary 3.2 presents an extension of Corollary 3.1.

COROLLARY 3.2. *Let $K = K^\dagger \geq K_0$ and $s = s^\dagger \geq s_0$ for an StG solution $\hat{\boldsymbol{\theta}}^{\mathrm{sg}}$. Under Condition 3.1, if $c^\dagger_{\min}(\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{Z}, K^\dagger, s^\dagger) \geq \frac{\sigma^2}{n}\{d_1 \log(pK^\dagger) + 220\}$ holds for some universal constant $d_1 > 134$, we have*

$$\begin{aligned}&\mathcal{L}^\dagger_g(\hat{\boldsymbol{\theta}}^{\mathrm{sg}}; \boldsymbol{\theta}^*, K^\dagger, s^\dagger)\\ &\quad\leq 4\exp\biggl[-\frac{3n}{200\sigma^2}\biggl\{c^\dagger_{\min}(\boldsymbol{\theta}^*, \mathbf{X}, \mathbf{Z}, K^\dagger, s^\dagger) - \frac{\sigma^2}{n}\bigl(134\log(pK^\dagger) + 220\bigr)\biggr\}\biggr] + \mathbb{P}(\mathcal{E}_1^c),\end{aligned}$$

where $\mathcal{E}_1 = \{\operatorname{supp}(\hat{\boldsymbol{\theta}}^{\mathrm{g}}) \subseteq \widetilde{\mathcal{S}}\}$ with $\widetilde{\mathcal{S}}$ being the resulting parameter set from screening.

PROOF. Using the similar arguments in the proof of Corollary 3.1, we have

$$\{\hat{\boldsymbol{\theta}}^{\mathrm{g}} \in \Pi(\boldsymbol{\beta}^*, K^\dagger, s^\dagger)\} \cap \mathcal{E}_1 \subseteq \{\hat{\boldsymbol{\theta}}^{\mathrm{sg}} \in \Pi(\boldsymbol{\beta}^*, K^\dagger, s^\dagger)\}.$$

The remaining proof follows the exact steps in the proof of Corollary 3.1 with little effort. □

4. Estimation and prediction error. This section focuses on the L_2 estimation and prediction error bounds for the oracle estimator $\hat{\boldsymbol{\theta}}^{\mathrm{ol}}$. The same bounds apply to the L_0-Fusion estimator if it recovers the true groups. Our analysis requires a regularity condition on the collapsed design matrix $\widetilde{\mathbf{X}}(\boldsymbol{\beta}^*)$ based on the true groups. Similar oracle error bounds are obtained by Ke, Fan and Wu (2015).

CONDITION 4.1. There exist universal constants $\kappa_l^*, \kappa_u^* > 0$ such that the square-roots of the maximum and minimum eigenvalues of the sample covariance are bounded above and below, respectively,

$$\kappa_l^* \leq \lambda^{1/2}_{\min}\biggl(\frac{\widetilde{\mathbf{X}}(\boldsymbol{\beta}^*)^\top \widetilde{\mathbf{X}}(\boldsymbol{\beta}^*)}{n}\biggr) \leq \lambda^{1/2}_{\max}\biggl(\frac{\widetilde{\mathbf{X}}(\boldsymbol{\beta}^*)^\top \widetilde{\mathbf{X}}(\boldsymbol{\beta}^*)}{n}\biggr) \leq \kappa_u^*$$

where $\widetilde{\mathbf{X}}(\boldsymbol{\beta}^*)$ is the collapsed design matrix given in (3.3).

PROPOSITION 4.1 (Oracle estimation and prediction error). *For model* (1.1), *an oracle solution to the* L_0*-Fusion problem* (2.1) *is found as* $\hat{\boldsymbol{\beta}}^{\mathrm{ol}}$.

(i) (*Estimation error*) *Under Condition* 4.1, *we have for any* $\xi > 1$,

$$\left\| \begin{pmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{ol}} \\ \hat{\boldsymbol{\alpha}}^{\mathrm{ol}} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta}^{*} \\ \boldsymbol{\alpha}^{*} \end{pmatrix} \right\|_2 \le \frac{\sigma \kappa_u^*}{(\kappa_l^*)^2} \sqrt{\frac{2\xi(K_0 + q)\log p}{n}},$$

with probability at least $1 - 2p^{-c\xi}$, *where* $c > 0$ *is a universal constant.*

(ii) (*Prediction error*) *For any* $\xi > 0$, *we have*

$$\frac{\left\| (\mathbf{X}, \mathbf{Z}) \begin{pmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{ol}} \\ \hat{\boldsymbol{\alpha}}^{\mathrm{ol}} \end{pmatrix} - (\mathbf{X}, \mathbf{Z}) \begin{pmatrix} \boldsymbol{\beta}^{*} \\ \boldsymbol{\alpha}^{*} \end{pmatrix} \right\|_2}{n^{1/2}} \le \sigma \sqrt{\frac{2\xi(K_0 + q)\log p}{n}},$$

with probability at least $1 - 2p^{-c\xi}$, *where c is a universal constant.*

In Appendix 3, we study the L_2 estimation error and prediction error of L_0-Fusion in the case of overidentification with $K^\dagger \ge K_0$ and $s^\dagger \ge s_0$. When the true grouping structure is misspecified, which is a common case in practice, we wish to have L_0-Fusion maintain reasonable estimation and prediction accuracy. It turns out that the resulting statistical rates lose the blessing of the original group structure due to the model misspecification: the rates reduce to those obtained in the standard high-dimensional sparse regression with no homogeneous groups of coefficients (see, e.g., Wainwright (2009)).

5. Mixed integer optimization formulation.

Given the strong statistical guarantee established for the L_0-Fusion in the previous sections, we now switch our focus to the computational aspect of the problem. In this paper, we leverage mixed integer optimization techniques to solve the combinatorial problem (2.1). Recently, Bertsimas, King and Mazumder (2016) proposed a MIO approach to solve the best subset selection problem of a remarkably enhanced scale. This inspires us to formulate L_0-Fusion as a MIO problem, for which we can resort to modern MIO solvers. In Section 5.1, we introduce the MIO formulation of L_0-Fusion. Then we present a warm start algorithm in Section 5.2 to further accelerate the MIO solver.

5.1. *MIO formulations for homogeneity fusion.*

Generally speaking, a MIO problem is formulated as follows:

$$\begin{aligned} \min_{\boldsymbol{\alpha} \in \mathbb{R}^m} \quad & \boldsymbol{\alpha}^\top \mathbf{Q} \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{a} \\ \text{s.t.} \quad & \mathbf{A}\boldsymbol{\alpha} \le \mathbf{b}, \\ & \alpha_j \in \{0, 1\}, \quad j \in \mathcal{I}, \\ & \alpha_j \ge 0, \quad j \notin \mathcal{I}, \end{aligned} \tag{5.1}$$

where $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{h \times m}$, $\mathbf{b} \in \mathbb{R}^h$ and $\mathbf{Q} \in \mathbb{R}^{m \times m}$ is positive semidefinite. The symbol "$\le$" represents elementwise inequalities. $\mathcal{I}$, an index subset of $[m]$, identifies the binary components of $\boldsymbol{\alpha}$. The mixture of discrete and continuous components of $\boldsymbol{\alpha}$ justifies the name of mixed integer programming. For more comprehensive background of MIO, we refer the readers to Bertsimas and Weismantel (2005) and Jünger and Reinelt (2013). Some popular MIO solvers include CPLEX, GLPK, MOSEK and GUROBI. Thanks to the branch-and-bound techniques (Cook, Lovász and Seymour (1995)), these solvers can provide both feasible solutions and lower bounds of the optimal objective value, from which we can learn how far a current solution is from the global optimum.

Now we introduce the MIO formulation for problem (2.1):

$$
\begin{aligned}
\min_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^q,\boldsymbol{\beta}\in\mathbb{R}^p,\\ \boldsymbol{\gamma}\in\mathbb{R}^K,\boldsymbol{\Omega}\in\{0,1\}^{p\times(K+1)}}} \quad & \sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\beta} - \mathbf{z}_i^\top\boldsymbol{\alpha})^2, \\
\text{subject to:} \quad & \omega_{jk}\in\{0,1\} \quad \forall k\in\{0\}\cup[K], j\in[p] \\
& \omega_{jk}(\beta_j - \gamma_k) = 0 \quad \forall k\in[K], j\in[p] \\
& \omega_{j0}\beta_j = 0 \quad \forall j\in[p] \\
& \gamma_k < \gamma_{k+1} \quad \forall k\in[K-1] \\
& \sum_{k=0}^{K}\omega_{jk} = 1 \quad \forall j\in[p] \\
& \sum_{j=1}^{p}\omega_{j0} \geq p - s.
\end{aligned}
\tag{5.2}
$$

Here, the number of groups K and the sparsity s are prespecified, which will be tuned by, for example, cross-validation. For any $j\in[p]$ and $k\in\{0\}\cup[K]$, we use ω_{jk} to denote the $(j,k+1)$th entry of $\boldsymbol{\Omega}$. For any $j\in[p]$ and $k\in[K]$, $\omega_{jk}=1$ ($\omega_{jk}=0$) means that the jth covariate is present (absent) in the kth group. To see why this is true, note that $\omega_{jk}(\beta_j-\gamma_k)=0$ enforces $\beta_j=\gamma_k$ when $\omega_{jk}=1$. Similarly, $\omega_{j0}=1$ implies that $\beta_j=0$, given the constraint that $\omega_{j0}\beta_j=0$. These types of constraints correspond to Specially Ordered Sets of type 1 (SOS-1) in Beale and Tomlin (1970) and can be replaced by linear constraints (Vielma and Nemhauser (2011), Markowitz and Manne (1957), Dantzig (1960)). The constraint $\gamma_k<\gamma_{k+1}$ resolves the identifiability issue so that $\{\gamma_k\}_{k\in[K]}$ can be uniquely determined. Constraint $\sum_{k=1}^{K}\omega_{jk}=1$ implies that each covariate belongs to exactly one group. Finally, $\sum_{j=1}^{p}\omega_{j0}\geq p-s$ ensures the size of the zero-valued group to be bigger than $p-s$, thereby constraining the sparsity of $\boldsymbol{\beta}$ below s. It is noteworthy that the solution of problem (5.2) can have fewer than K groups.

Formulation (5.2) can be easily extended to accommodate prior knowledge regarding group structures. For instance, some covariates are known in advance to be in the same group, *say*, $\beta_{j\in\mathcal{J}}$ are equal for a set $\mathcal{J}\subset[p]$. Then we can incorporate this information into (5.2) by adding the constraint that $\omega_{j_1k}=\omega_{j_2k}, \forall j_1, j_2\in\mathcal{J}, j_1\neq j_2, k\in\{0\}\cup[K]$. In another example with the absence of pair covariates among $\{X_j\}_{j\in\mathcal{J}}$ belonging to the same group, we can add the constraint $\sum_{j\in\mathcal{J}}\omega_{jk}\leq 1, k=\{0\}\cup[K]$.

5.2. *Warm start algorithm.* This section introduces a discrete first-order algorithm to provide a warm start for the MIO problem (5.2). Our algorithm is inspired by Bertsimas, King and Mazumder (2016), who proposed a similar algorithm to initialize a MIO solver to solve the BSS problem. Since this algorithm is not limited to the square loss objective in the L_0-Fusion problem, we extend the original L_0-Fusion problem to embrace a wider range of objective functions.

Suppose we are interested in a convex objective function $g(\boldsymbol{\theta})$ satisfying:

(i) $g(\boldsymbol{\theta})\geq C_2>-\infty$ for some universal constant C_2;

(ii) $g(\boldsymbol{\theta})$ has Lipschitz continuous gradient, that is, $\|\nabla g(\boldsymbol{\theta})-\nabla g(\widetilde{\boldsymbol{\theta}})\|_2\leq l\|\boldsymbol{\theta}-\widetilde{\boldsymbol{\theta}}\|_2$ for some positive l and any $\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}}\in\boldsymbol{\Theta}(K,s)$, which is defined in the beginning of Section 3.

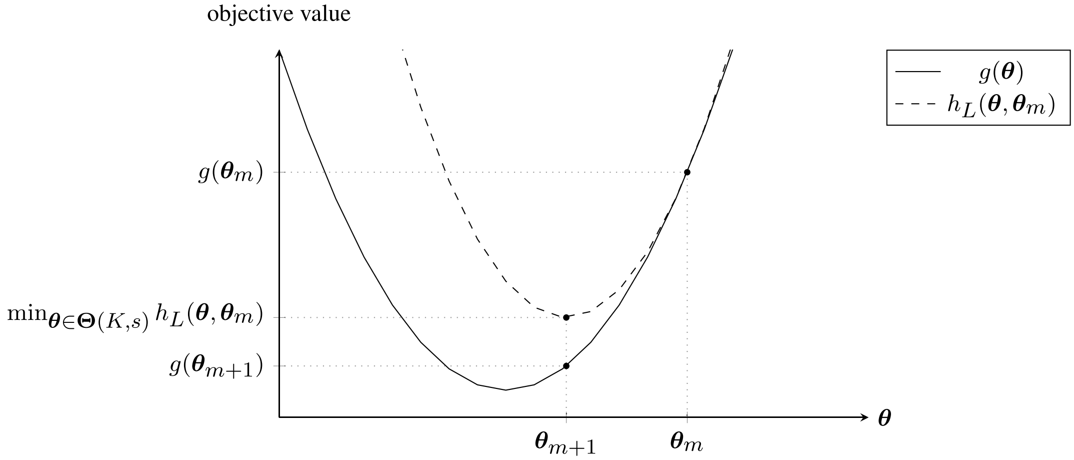


FIG. 2. *An illustration of $g(\boldsymbol{\theta})$ and $h_L(\boldsymbol{\theta},\boldsymbol{\theta}_m)$ in Proposition 5.1. The solid curve is $g(\boldsymbol{\theta})$ and the dashed curve is $h_L(\boldsymbol{\theta},\boldsymbol{\theta}_m)$.*

Consider the following generalized L_0-Fusion problem:

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}(K,s)} g(\boldsymbol{\theta}). \tag{5.3}$$

We propose an algorithm to attain a feasible point close to the solution of problem (5.3), based on ideas from projected gradient descent methods (Nesterov (2004), Nesterov (2013)). Note that this point can serve as a starting point for MIO solvers and the objective function value at this point is an upper bound of the global minimum. To do so, we construct a curve $h_L(\boldsymbol{\theta},\boldsymbol{\theta}')$ defined in the following proposition, which lies above $g(\boldsymbol{\theta})$ and is tangent to $g(\boldsymbol{\theta})$ at $\boldsymbol{\theta}'$.

PROPOSITION 5.1 (Nesterov (2004), Nesterov (2013)). *For a convex function $g(\boldsymbol{\theta})$ satisfying* (ii), *and for any $L \geq l$, we have*

$$g(\boldsymbol{\theta}) \leq h_L(\boldsymbol{\theta},\boldsymbol{\theta}') := g(\boldsymbol{\theta}') + (\boldsymbol{\theta}-\boldsymbol{\theta}')^\top \nabla g(\boldsymbol{\theta}') + \frac{L}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2, \tag{5.4}$$

for all $\boldsymbol{\theta}$, $\boldsymbol{\theta}'$ with equality holding at $\boldsymbol{\theta}=\boldsymbol{\theta}'$.

As illustrated in Figure 2, given a point $\boldsymbol{\theta}_m$, we can always improve the current objective value $g(\boldsymbol{\theta}_m)$ through the descending route:

$$g(\boldsymbol{\theta}_{m+1}) \leq h_L(\boldsymbol{\theta}_{m+1},\boldsymbol{\theta}_m) \leq h_L(\boldsymbol{\theta}_m,\boldsymbol{\theta}_m) = g(\boldsymbol{\theta}_m), \tag{5.5}$$

where

$$\boldsymbol{\theta}_{m+1} \in \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}(K,s)} h_L(\boldsymbol{\theta},\boldsymbol{\theta}_m) = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}(K,s)} \left\|\boldsymbol{\theta} - \left(\boldsymbol{\theta}_m - \frac{1}{L}\nabla g(\boldsymbol{\theta}_m)\right)\right\|_2^2.$$

For convenience, for any constant vector $\mathbf{c} = (c_1,\dots,c_{p+q})^\top$, define

$$\mathcal{H}_{K,s}(\mathbf{c}) := \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}(K,s)} \|\boldsymbol{\theta}-\mathbf{c}\|_2^2.$$

Then $\boldsymbol{\theta}_{m+1} \in \mathcal{H}_{K,s}(\boldsymbol{\theta}_m - \frac{1}{L}\nabla g(\boldsymbol{\theta}_m))$. By carrying out this improvement iteratively, we implement Algorithm 2 below that supplies decent warm starts to solve problem (5.3).

Algorithm 2: Warm Start

```
Input: Loss function $g(\boldsymbol{\theta})$, number of groups $K$, sparsity constraint $s$, step size
       parameter $L$ and convergence tolerance $\varepsilon$.
1: Initialize with $\boldsymbol{\theta}_1 \in \mathbb{R}^{p+q}$.
2: For $m \geq 1$, $\boldsymbol{\theta}_{m+1} \in \mathcal{H}_{K,s}(\boldsymbol{\theta}_m - \frac{1}{L}\nabla g(\boldsymbol{\theta}_m))$.
3: Repeat Step 2 until $g(\boldsymbol{\theta}_m) - g(\boldsymbol{\theta}_{m+1}) \leq \varepsilon$.
Output: $\boldsymbol{\theta}_{m+1}$.
```

Algorithm 2 is essentially a projected gradient descent algorithm: In each iteration, we perform a gradient descent step followed by projection onto $\mathcal{H}_{K,s}$. To obtain an element in $\mathcal{H}_{K,s}(\mathbf{c})$ for any $\mathbf{c} \in \mathbb{R}^{p+q}$, we can exploit the subroutine Algorithm 1 in Section 1 of the Supplementary Material (Wang et al. (2024)), which is a generalization of the segment neighbourhood method (Auger and Lawrence (1989)) with sparsity constraint. To investigate the algorithmic convergence of Algorithm 2, we first define the first-order stationary points of problem (5.3) as follows.

DEFINITION 5.1 (First-order stationary point). We say a vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}(K, s)$ is a first-order stationary point for problem (5.3) if $\boldsymbol{\theta} \in \mathcal{H}_{K,s}(\boldsymbol{\theta} - \frac{1}{L}\nabla g(\boldsymbol{\theta}))$ for some positive constant $L \geq l$.

The following proposition establishes two important properties of the first-order stationary points that underpin the effectiveness and stability of our warm start Algorithm 2.

PROPOSITION 5.2. *Suppose a positive constant $L > l$.*

(i) *If $\boldsymbol{\theta}$ is a solution to problem* (5.3), *then it is a first-order stationary point*; *and*
(ii) *If $\boldsymbol{\theta}$ is a first-order stationary point*, *then the set $\mathcal{H}_{K,s}(\boldsymbol{\theta} - \frac{1}{L}\nabla g(\boldsymbol{\theta}))$ has exactly one element $\boldsymbol{\theta}$.*

Now we present the convergence property and convergence rate of Algorithm 2 through Proposition 5.3 and Theorem 5.2, respectively.

PROPOSITION 5.3. *For problem* (5.3) *and some positive constant $L > l$, let $\boldsymbol{\theta}_m, m \geq 1$ be the sequence generated by Algorithm* 2. *We have*:

(i) $g(\boldsymbol{\theta}_m) - g(\boldsymbol{\theta}_{m+1}) \geq \frac{L-l}{2}\|\boldsymbol{\theta}_m - \boldsymbol{\theta}_{m+1}\|_2^2$; *and*
(ii) $\|\boldsymbol{\theta}_{m+1} - \boldsymbol{\theta}_m\|_2 \to 0$ *as* $m \to \infty$.

THEOREM 5.2. *For the sequence $\{\boldsymbol{\theta}_m\}_{m=1}^{\infty}$ generated by Algorithm* 2, *if $L > l$, then there exists $c \in \mathbb{R}$, such that for any $M \in \mathbb{Z}^+$ we have*

$$\min_{m=1,\ldots,M} \|\boldsymbol{\theta}_{m+1} - \boldsymbol{\theta}_m\|_2^2 \leq \frac{2(g(\boldsymbol{\theta}_1) - c)}{M(L-l)},$$

where $g(\boldsymbol{\theta}_m) \downarrow c$ as $m \to \infty$.

Finally, we show that Algorithm 2 gives a feasible solution whose objective value is the same as some first-order stationary point under mild conditions.

PROPOSITION 5.4. *Consider problem* (5.3) *and some constant $L > l$, let $\boldsymbol{\theta}_m, m \geq 1$ be the sequence generated by Algorithm* 2. *Suppose g satisfies the following conditions*:

(i) *g has second-order derivative*;

(ii) *there exists $l' > 0$ such that $l'\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_2 \le \|\nabla g(\boldsymbol{\theta}) - \nabla g(\widetilde{\boldsymbol{\theta}})\|_2$ for any $\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \in \Theta(K, s)$ satisfying $\mathbb{G}(\boldsymbol{\beta}) = \mathbb{G}(\widetilde{\boldsymbol{\beta}})$*;

(iii) *$\{\boldsymbol{\theta} \in \boldsymbol{\Theta}(K, s) \mid g(\boldsymbol{\theta}) \le C\}$ is bounded for any $C \in \mathbb{R}$.*

Then $g(\boldsymbol{\theta}_m)$ converges to $g(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a first-order stationary point.

The detailed proof of Proposition 5.4 is given in Proposition 2.1 and Remark 2.1 in Section 2.11 of the Supplementary Material (Wang et al. (2024)).

6. Numerical studies. We conduct a variety of numerical experiments to assess the performance of the proposed L_0-Fusion methodology. We use the *normalized mutual information* (NMI, Ana and Jain (2003)) to evaluate grouping accuracy. Specifically, given $\mathbb{G}_1 = \{G_1^{(1)}, G_1^{(2)}, \dots\}$ and $\mathbb{G}_2 = \{G_2^{(1)}, G_2^{(2)}, \dots\}$ as two sets of disjoint clusters of $[p]$, the mutual information $I(\mathbb{G}_1; \mathbb{G}_2)$ between $\mathbb{G}_1$ and $\mathbb{G}_2$ takes the form

$$I(\mathbb{G}_1; \mathbb{G}_2) := \sum_{i \in [|\mathbb{G}_1|], j \in [|\mathbb{G}_2|]} \frac{|G_1^{(i)} \cap G_2^{(j)}|}{p} \log\left(\frac{p|G_1^{(i)} \cap G_2^{(j)}|}{|G_1^{(i)}||G_2^{(j)}|}\right),$$

and the entropy of $\mathbb{G}_1$ is

$$H(\mathbb{G}_1) := I(\mathbb{G}_1; \mathbb{G}_1) = -\sum_{i \in [|\mathbb{G}_1|]} \frac{|G_1^{(i)}|}{p} \log\left(\frac{|G_1^{(i)}|}{p}\right).$$

Now we are ready to define the NMI between $\mathbb{G}_1$ and $\mathbb{G}_2$ as

$$\mathrm{NMI}(\mathbb{G}_1, \mathbb{G}_2) := \frac{I(\mathbb{G}_1; \mathbb{G}_2)}{\{H(\mathbb{G}_1) + H(\mathbb{G}_2)\}/2}.$$

Note that if $\mathbb{G}_1$ and $\mathbb{G}_2$ share the identical group structure, then $\mathrm{NMI}(\mathbb{G}_1, \mathbb{G}_2) = 1$.

The rest of the section is organized as follows. Section 6.1 compares L_0-Fusion with its competitors in terms of grouping accuracy and investigates the effectiveness of the warm start Algorithm 2 under low-dimensional regimes. Section 6.2 implements the "*screening then grouping*" strategy discussed in Section 2.2 to perform homogeneity fusion under ultrahigh-dimensional sparse setups. Finally, Section 6.3 applies L_0-Fusion to group lipids in a study of metabolomic effects on body mass index (BMI).

6.1. *Low-dimensional regime.* We consider a collection of low-dimensional setups where the design vectors $\{\mathbf{x}_i\}_{i=1}^n$ are independent realizations from a p-dimensional multivariate normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with mean zero and covariance matrix $\boldsymbol{\Sigma} := (\Sigma_{ij})$. We adopt the autoregressive design in the sense that $\Sigma_{ij} = \rho^{|i-j|}$ with $\rho \in \{0, 0.5\}$. In particular, $\rho = 0$ gives the independent measures design. For each fixed $\mathbf{X}$, we generate the responses $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. Throughout this section, we always set the group number $K_0 = 4$, while n, p and $\boldsymbol{\beta}^*$ are specified in given settings under investigation.

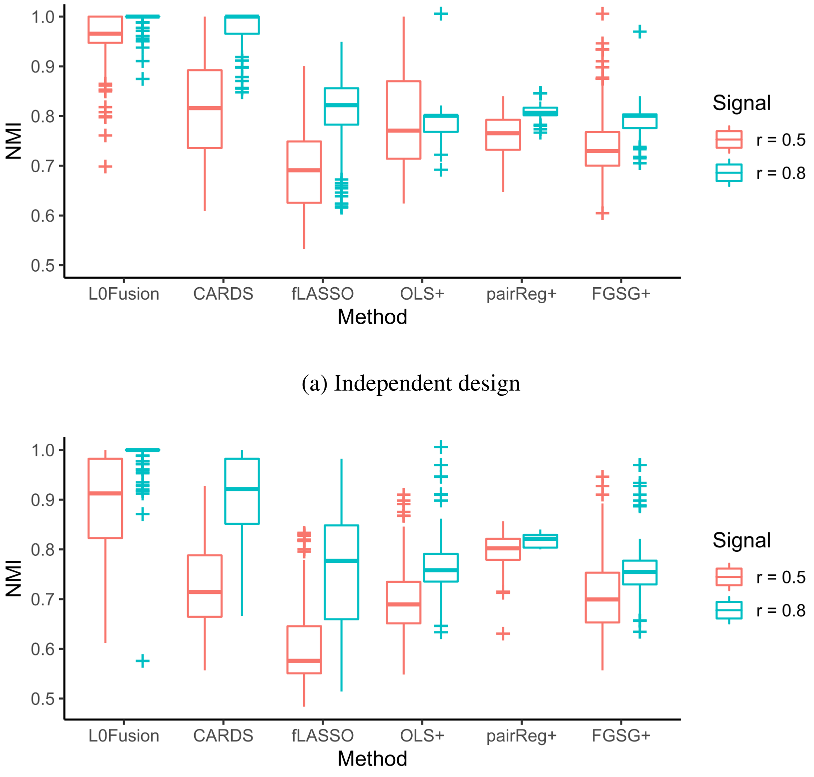
In Sections 6.1.1 and 6.1.2, we compare six methods when group sizes are equal and unequal respectively: L_0-Fusion, ordinary least squares (OLS), fused LASSO (fLASSO, Tibshirani et al. (2005)), pairwise fusion (pairReg, Ma and Huang (2017)), feature grouping and selection over an undirected graph (FGSG, Zhu, Shen and Pan (2013)) and clustering algorithm in regression via data-driven segmentation (CARDS, Ke, Li and Zhang (2016)). To compare L_0-Fusion with CARDS and fLASSO, tuning parameters are chosen via Bayesian information criterion (BIC). For OLS, FGSG and pairReg, we first tune the parameters (if any) in these methods via 10-fold cross-validation in terms of mean squared error (MSE).

Note that these methods encourage coefficients within the same group to be close but not exactly the same. To derive grouping structures and gauge their accuracy, we perform *k-means* clustering on the solutions given respectively by OLS, FGSG and pairReg under the oracle cluster number $K = 4$. We use OLS+, FGSG+ and pairReg+ to represent the corresponding post-clustering results. In Section 6.1.3, we present the NMI of the warm-start solution in Section 5.2 with varying n and p, and illustrate how warm starts help the convergence of L_0-Fusion especially in the early search stage. All the results are based on 200 rounds of independent Monte Carlo experiments.

6.1.1. *Equal group sizes.* We start with the case where all the coefficient groups have equal sizes. Specifically, set $n = 120$, $p = 80$ and $K_0 = 4$ coefficient groups of size 20, which take values $-2r, -r, r, 2r$, respectively, with $r \in \{0.5, 0.8\}$. Figure 3 displays the boxplots of NMI for correlation coefficients $\rho \in \{0, 0.5\}$ and signal strengths $r \in \{0.5, 0.8\}$. Our findings are summarized as follows:

(i) L_0-Fusion exhibits significantly higher NMI than the other methods in all the cases, even though OLS+, pairReg+ and FGSG+ used the oracle knowledge of the true number of groups in the analyses.

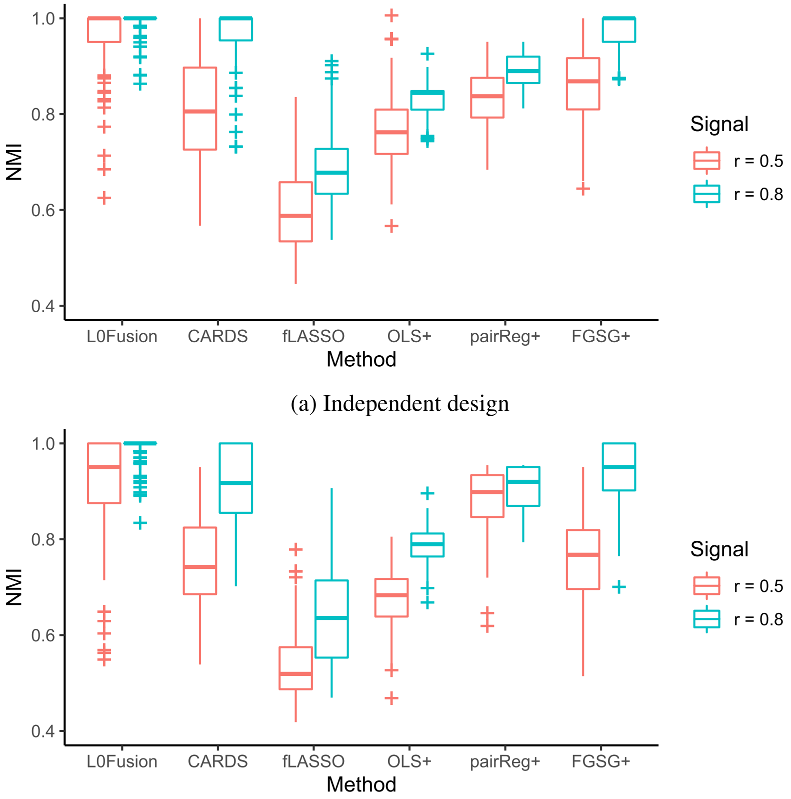
(ii) In the cases where $r = 0.8$, the lower quartile of NMI of L_0-Fusion is 1. This means that in more than 75 % repetitions, L_0-Fusion achieved grouping consistency, and BIC correctly chose the true group size $K = K_0$ and the true number of signals $s = s_0$.



(a) Independent design

(b) Autoregressive design with $\rho = 0.5$

FIG. 3. *Grouping accuracy with equal group sizes under different covariance designs and signal strengths. We set the correlation coefficient* $\rho \in \{0, 0.5\}$ *and signal strength* $r \in \{0.5, 0.8\}$.



(a) Independent design

(b) Autoregressive design with $\rho = 0.5$

FIG. 4. *Grouping accuracies with unequal group sizes under different covariance designs and signal strengths. We set the correlation coefficient* $\rho \in \{0, 0.5\}$ *and signal strength* $r \in \{0.5, 0.8\}$.

(iii) Almost all the methods yield higher grouping accuracy when r is larger by comparing the left red boxplox to the right blue boxplot, or ρ is smaller by comparing panel (a) to panel (b).

6.1.2. *Unequal group sizes.* When groups have different sizes, presumably challenges arise from identifying small groups. To assess the capability of detecting small groups, set $n = 120$, $p = 80$ and divide the true predictors into $K_0 = 4$ groups of sizes 1, 20, 20, 39, whose coefficient values are set at $-4r$, $-r$, r, $2r$, respectively. Figure 4 shows the boxplots of NMI obtained by all the aforementioned approaches under $\rho \in \{0, 0.5\}$ and $r \in \{0.5, 0.8\}$. The results are summarized as follows:

(i) Similar to Section 6.1.1, L_0-Fusion outperforms the competing methods in terms of NMI uniformly in all the cases.

(ii) Once again, all the methods enjoy better grouping accuracy when r is larger through a comparison of the red boxplot with the blue boxplot or ρ is smaller through a comparison of panel (a) and panel (b).

(iii) The performance gap on NMI between L_0-Fusion and fLASSO is further enlarged here compared with the case of equal group sizes. This suggests the robustness of L_0-Fusion with respect to group size heterogeneity.

6.1.3. *Warm-start algorithm.* We begin with a simulation experiment to assess the grouping detection accuracy for the solution obtained by the discrete first-order algorithm, which is

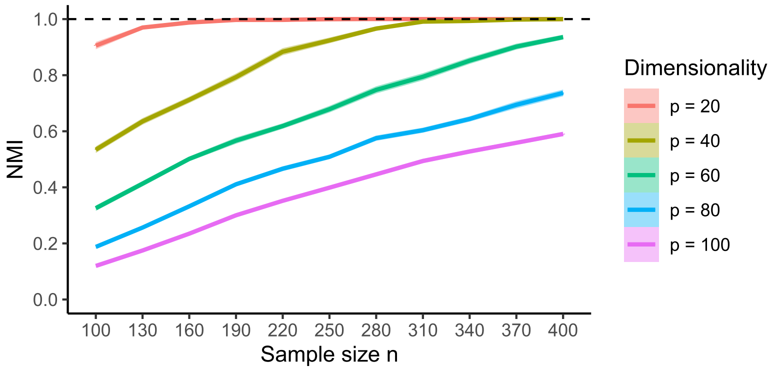


FIG. 5. *Error bands for NMIs of the discrete first-order algorithm with respect to different sample size and dimensionality. Here,* $s_0 = p$, $q = 0$ *and* $K_0 = 4$. *Rows of design matrix* $\mathbf{X}$ *are i.i.d. from* $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. *The true predictors are divided into* 4 *groups of size* $(p/4, p/4, p/4, p/4)$. *True coefficients within each group are* $(-1, -0.5, 0.5, 1)$, *respectively.*

introduced in Section 5.2 with initial value $\boldsymbol{\theta}_0 = \mathbf{0}_{p+q}$. Figure 5 presents the results of NMI produced by this algorithm with varying oracle K_0 as n and p. With no surprise, this figure shows a deteriorating trend of the performance as p grows or n decreases. However, clearly this warm-start algorithm is capable of recovering the group structure with a sufficiently large sample.

Next, we exploit the discrete first-order algorithm to provide a warm start for L_0-Fusion. The MIO solver in *Gurobi* (Gurobi Optimization, LLC (2021)) terminates when the gap between the lower and upper objective bounds is less than the Mixed-Integer Programming (MIP) Gap (a user-determined parameter between 0 and 1) times the absolute value of the incumbent objective value. More precisely, let z_P be the incumbent primal objective value, which is an upper bound for the global minimum, and z_D be the dual objective value, which is a lower bound for the global minimum. Then the MIP Gap is defined as $|z_P - z_D|/|z_P|$. Figure 6 tracks the MIP Gap and the NMI of L_0-Fusion against its running time on the University of Michigan High Performance Linux cluster. Each job uses 4 CPUs and 16 GB memory, which is satisfied on most personal computers. Below is a summary of our findings:

(i) The warm-start solution yields NMI $= 0.6$, which is plausible but far from optimal.

(ii) L_0-Fusion with a warm start yields significantly higher NMI than that with a cold start within the first 50 seconds.

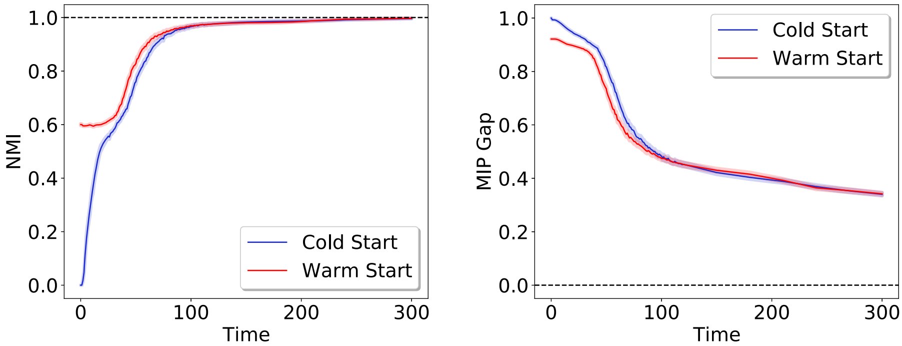


FIG. 6. *Error bands for comparing warm start and cold start after* 200 *Monte Carlo repetitions. Here, we set* $n = 250$, $p = 120$, $s_0 = 120$ *and* $K_0 = 4$. *Rows of design matrix* $\mathbf{X}$ *are i.i.d. from* $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. *The true predictors are divided into* 4 *groups of size* $(30, 30, 30, 30)$. *True coefficients within each group are* $(-1, -0.5, 0.5, 1)$.

(iii) Even when the MIP Gap is not exactly 0, L_0-Fusion can achieve near perfect group recovery. Therefore, one can still expect decent grouping results even if the algorithm has to halt before the MIP Gap vanishes.

6.2. *Ultrahigh-dimensional regime.* For ultrahigh-dimensional cases, we let $p = 20{,}000$, $s_0 = 60$, $n = \lfloor 2s_0 \log p \rfloor$ and $K_0 = 4$. All the entries of the design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ are independent standard Gaussian random variables. The true predictors are set as the first 60 predictors and divided into 4 groups of size 15 with coefficient values $-2r$, $-r$, r, $2r$, respectively, where $r \in \{0.15, 0.2, 0.25, 0.3\}$. All the results in this section are based on 100 independent Monte Carlo repetitions.

In the screening step, we first estimate the true sparsity s_0 using the model size obtained from MCP (Zhang (2010)) under the lowest 10-fold cross-validation (CV) MSE. To speed up the computation in this simulation study, we set *a prior* upper limit of sparsity at 100 or lower. Denote the resulting sparsity estimator by $\widehat{s}$. Then we use CoSaMP with projection size $\pi = \widehat{s}$ and expansion size $l = \lceil \widehat{s}/2 \rceil$ to generate a screening set $\widehat{\mathcal{S}}$ of size $\widehat{s}$. To evaluate the quality of the screened set $\widehat{\mathcal{S}}$, in Figure 7 we investigate the cardinality and true positive proportion (TPP) of $\widehat{\mathcal{S}}$, the latter of which is defined as

$$\text{TPP}(\widehat{\mathcal{S}}) := \frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^0|}{|\mathcal{S}^0|}, \tag{6.1}$$

where $\mathcal{S}^0$ denotes the true support set.

We then perform grouping on the reduced design. The implementations for all the grouping methods are similar as those in Section 6.1. All the zero coefficients are considered as forming one group when we calculate the NMI. Figure 8 reports the NMI of L_0-Fusion, CARDS, fLASSO, OLS+, pairReg+ and FGSG+. We summarize key results as follows:

(i) Figure 7 shows that as signal strength grows, CoSaMP yields higher TPP and smaller screening sizes, meaning that both accuracy and efficiency of CoSaMP improve.

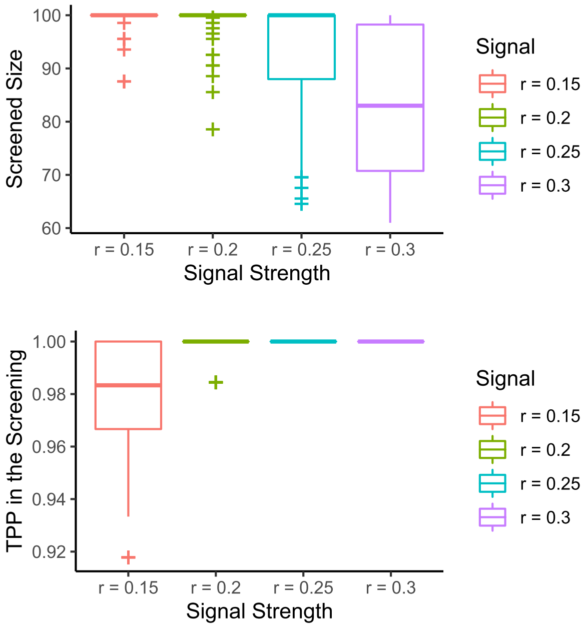


FIG. 7. *Screening results for ultrahigh-dimensional problems. Here,* $p = 20{,}000$, $s_0 = 60$, $n = \lfloor 2s_0 \log p \rfloor$.

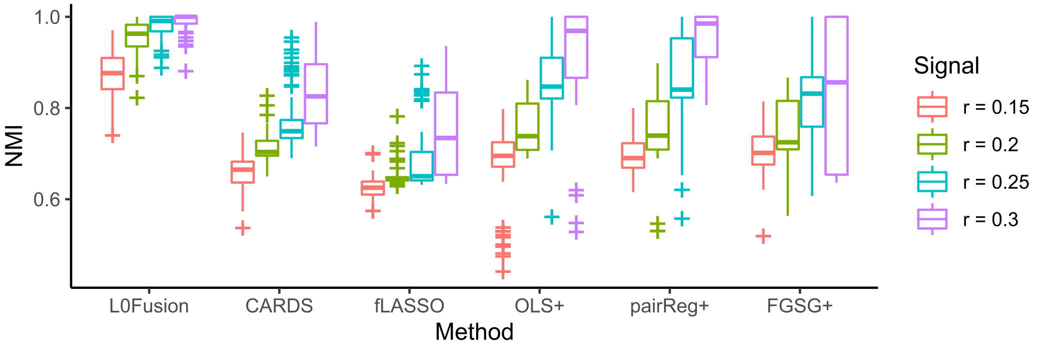


FIG. 8. *Grouping results for ultrahigh-dimensional problems after* 100 *Monte Carlo repetitions.*

(ii) Combining Figures 7 and 8, we observe that when $r \in \{0.2, 0.25, 0.3\}$, CoSaMP achieves sure screening while the grouping can be far from the truth. This suggests that the grouping error is due to the grouping stage rather than the screening stage.

(iii) L_0-Fusion still outperforms all the competing methods on the reduced designs.

6.3. *Real data analysis.* We further illustrate the proposed method by an empirical study of metabolomics data collected from $n = 397$ adolescents consisting of 197 boys and 200 girls aged 8 to 18 years during a critical period of growth and sexuality maturation. Early onset of obesity in the adolescent years has been found to be associated with an increased risk of many diseases (e.g., hypertension, diabetics and cancer) during adulthood. Thus, it is of great scientific interest to detect key groups of lipids (largest metabolites among metabolomics) that predict body mass index (BMI), adjusted by age and sex (1 for boy and 0 for girl). We investigate a total of $p = 234$ lipids to determine the number of lipid groups, group memberships and associated average contribution of a group predictor to BMI. We fit the following linear model with a group homogeneity pursuit on the outcome of BMI:

$$\text{BMI} = \alpha_0 + \alpha_1 \,\text{age} + \alpha_2 \,\text{sex} + \sum_{j=1}^{234} \beta_j \,\text{lip}_j + \varepsilon, \quad \beta_j \in \{0, \gamma_1, \gamma_2, \dots, \gamma_K\}, \forall j \in [p], \tag{6.2}$$

where the number of signal groups (K) as well as the group of null lipids with $\gamma_0 = 0$ is determined by 10-fold CV. Each group represents a subset of lipids with a shared nonzero effect size $\gamma_k, k = 1, \dots, K$. We normalize the design matrix to ensure mean 0 and variance 1 except intercept and sex. To calibrate the effect sizes with respect to different group sizes, for each group of lipids, we use their average measurement as the group's overall measurement; therefore, the corresponding group-level effect size is γ_k multiplied by the group size.

We adopt the aforementioned "*screen then group*" strategy. Similar to Section 6.2, we first estimate the true sparsity s by MCP with 10-fold cross-validation (CV) based on MSE and apply CoSaMP to identify promising individual lipids from the pool of 234 lipids. This screening step selects 18 potential lipids together with intercept, age and gender. In the second phase, we perform L_0-Fusion on these selected lipids. Through 10-fold CV over $K = \{1, \dots, 10\}$, we detect six groups with nonzero effect sizes. It is noteworthy that GUROBI solves the L_0-Fusion problem within few seconds. The results are summarized in Table 1. For group 1 consisting of 7 similar lipids, the group-level average lipid measurement has a 4.0 effect size on BMI.

We also conduct a confirmatory semisimulation using the metabolomics design matrix of this real data set. Assume that a variable of interest y relates to the metabolomics as follows:

$$y = \alpha_1 \,\text{age} + \alpha_2 \,\text{sex} + \sum_{j=1}^{234} \beta_j \,\text{lip}_j + \varepsilon, \quad \beta_j \in \{0, \gamma_1, \gamma_2, \dots, \gamma_K\}, \forall j \in [p]. \tag{6.3}$$

TABLE 1

The data analysis results of lipid groups and effect sizes among 18 *promising metabolites from preliminary screening by CoSaMP*

Group (size)	Features	Group-level effect size
	Intercept	21.88
	Boys versus girls	−1.05
	Age	0.62
Group 1 (7)	"cholesterol biosynthesis"	4.00
Group 2 (6)	"nutritional energy support and regulation"	−3.13
Group 3 (2)	"energy transport"	4.44
Group 4 (1)	"diet signaling"	−2.00
Group 5 (1)	"energy production"	−1.28
Group 6 (1)	"peptide hormones on food consumption"	0.99

We set $\alpha_1 = 0.5$, $\alpha_2 = 1$ and randomly assign the coefficients $\boldsymbol{\beta}$ with a sparse (20 nonzero values) and grouped (4 groups of size 5) structure, where the true coefficients within individual groups are equal to $-2r, -r, r, 2r$, respectively, with $r \in \{0.5, 1\}$. Then we generate $n = 397$ responses from the metabolomics design according to (6.3), where ϵ's are i.i.d. $\mathcal{N}(0, 1)$. We assess the prediction performance with/without group structure along with the grouping accuracy by randomly splitting the observations into a training set and a testing set at each repetition. We then implement both the screening procedure alone and the StG procedure on the training set and compare their prediction accuracy in terms of MSE on the testing set. Table 2 reports the testing MSE with standard error as well as the quantiles of NMI based on 100 independent Monte Carlo repetitions. This table clearly suggests that leveraging the existing group structure by L_0-Fusion improves the prediction accuracy.

7. Discussion. This paper studies a combinatorial approach called L_0-Fusion that enables simultaneous operation of clustering and estimation for regression coefficients in a linear model. This analytic task addresses a practical need for learning homogeneous groups of nonzero regression coefficients in the regression analysis to assess the relationship between outcomes and clustered signal features. We propose to formulate the L_0-Fusion problem as a mixed integer optimization (MIO) problem and then leverage modern MIO solvers to compute the corresponding parameter estimates. When the dimension is too high for the MIO solver to handle, we invoke CoSaMP as a preliminary variable screening procedure to reduce dimension prior to the L_0-Fusion. As shown theoretically and numerically in in this paper, such a "screen then group" strategy dramatically broadens the applicability of the homogeneity fusion technique, which can scale L_0-Fusion up to the ambient dimension $p = 20{,}000$ with high accuracy of recovering the true group structure of regression coefficients. This level of methodological capacity allows to handle a large number of modern biomedical data

TABLE 2

Results of grouping and prediction accuracy for the semisimulation

Signal strength	Prediction error (with grouping)	Prediction error (without grouping)	NMI (quartiles)
$r = 1$	**1.092 (0.031)**	1.186 (0.029)	Max: 1.00; Q3: 1.00; Median: 1.00; Q1:1.00; Min: 0.75
$r = 0.5$	**1.120 (0.023)**	1.223 (0.001)	Max: 1.00; Q3: 1.00; Median: 0.91; Q1: 0.85; Min: 0.58

sets. Thus, this two-stage approach as well as its variants provide efficient toolboxes to solve many homogeneity fusion problems on large-scale data sets.

Theoretically, we establish grouping consistency of the L_0-Fusion estimator, for which the sample size n only needs to grow at the same rate as the sum of logarithms of the true sparsity and true group number, that is, $\log(pK)$. This sample size requirement is also shown to be necessary for any procedure to achieve grouping consistency. These technical results are not only of theoretical interest, but also useful to guide practical work such as sample size determination in a study design.

An important future work concerns statistical inference after the operation of L_0-Fusion. A thorough investigation on the influence of selection errors on statistical inference, in both aspects of finite-sample and large-sample properties, is of great interest. This L_0-Fusion may be extended to other regression problems with the framework of generalized linear models where iterative procedures used in the parameter estimation rely on weighted least squares objective functions. Thus, this extension is technically manageable but may require substantial computational effort. Also, we would consider an extension of this method to the setting of estimating equations, which could cover a broad range of important statistical models, such as GEE regression, Cox regression and quantile regression.

Acknowledgments. We thank the Co-editors, the Associate Editor and two anonymous reviewers for constructive comments that have led to a significantly improved manuscript. We also thank Drs. Z.T. Ke and X. Shen for providing their Matlab code and R code for CARDS and FGSG methods, respectively. Wang and Wu are co-first authors of this paper.

Funding. Dr. Song's research is supported by National Institutes of Health grants R01ES024732 and R01ES033656 and National Science Foundation grants DMS2113564 and DMS2015366.

Dr. Zhou's research was partially supported by the National Natural Science Foundation of China (Nos. 12271441 and 11931014).

SUPPLEMENTARY MATERIAL

Supplement to "Supervised homogeneity fusion: A combinatorial approach" (DOI: 10.1214/23-AOS2347SUPP; .pdf). The Supplementary Material contains analytic proofs of the main results, technical lemmas and additional theoretical results.

REFERENCES

ANA, L. F. and JAIN, A. K. (2003). Robust data clustering. In 2003 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003. *Proceedings*. **2** II–II. IEEE, New York.

AUGER, I. E. and LAWRENCE, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* **51** 39–54. MR0978902 https://doi.org/10.1016/S0092-8240(89)80047-3

BEALE, E. M. L., KENDALL, M. G. and MANN, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika* **54** 357–366. MR0221656 https://doi.org/10.1093/biomet/54.3-4.357

BEALE, E. M. L. and TOMLIN, J. A. (1970). Special facilities in a general mathematical programming system for non-convex problems using ordered sets of variables. *Oper. Res.* **69** 99.

BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. MR3476618 https://doi.org/10.1214/15-AOS1388

BERTSIMAS, D. and WEISMANTEL, R. (2005). *Optimization over Integers* **13**. Dynamic Ideas Belmont.

BONDELL, H. D. and REICH, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123. MR2422825 https://doi.org/10.1111/j.1541-0420.2007.00843.x

BRAUN, J. M., HOFFMAN, E., SCHWARTZ, J., SANCHEZ, B., SCHNAAS, L., MERCADO-GARCIA, A., SOLANO-GONZALEZ, M., BELLINGER, D. C., LANPHEAR, B. P. et al. (2012). Assessing windows of susceptibility to lead-induced cognitive deficits in Mexican children. *Neurotoxicolog* **33** 1040–1047.

COOK, W., LOVÁSZ, L. and SEYMOUR, P., eds. (1995) *Combinatorial Optimization*: *Papers from the DIMACS Special Year*. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* **20**. Amer. Math. Soc., Providence, RI. MR1338611 https://doi.org/10.1090/dimacs/020

DANTZIG, G. B. (1960). On the significance of solving linear programming problems with some integer variables. *Econometrica* **28** 30–44. MR0112748 https://doi.org/10.2307/1905292

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

GARSIDE, M. (1965). The best sub-set in multiple regression analysis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **14** 196–200.

GUO, Y., ZHU, Z. and FAN, J. (2020). Best subset selection is robust against design dependence. Preprint. Available at arXiv:2007.01478.

GUROBI OPTIMIZATION, LLC (2021). Gurobi optimizer reference manual.

HOCKING, R. R. and LESLIE, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics* **9** 531–540. MR0219192 https://doi.org/10.2307/1266192

JEON, J.-J., KWON, S. and CHOI, H. (2017). Homogeneity detection for the high-dimensional generalized linear model. *Comput. Statist. Data Anal.* **114** 61–74. MR3660839 https://doi.org/10.1016/j.csda.2017.04.001

JÜNGER, M. and REINELT, G. (2013). *Facets of Combinatorial Optimization*. Springer, Berlin.

KE, Y., LI, J. and ZHANG, W. (2016). Structure identification in panel data analysis. *Ann. Statist.* **44** 1193–1233. MR3485958 https://doi.org/10.1214/15-AOS1403

KE, Z. T., FAN, J. and WU, Y. (2015). Homogeneity pursuit. *J. Amer. Statist. Assoc.* **110** 175–194. MR3338495 https://doi.org/10.1080/01621459.2014.892882

KOBROSLY, R. W., PARLETT, L. E., STAHLHUT, R. W., BARRETT, E. S. and SWAN, S. H. (2012). Socioeconomic factors and phthalate metabolite concentrations among United States women of reproductive age. *Environ. Res.* **115** 11–17. https://doi.org/10.1016/j.envres.2012.03.008

LIAN, H., QIAO, X. and ZHANG, W. (2021). Homogeneity pursuit in single index models based panel data analysis. *J. Bus. Econom. Statist.* **39** 386–401. MR4235184 https://doi.org/10.1080/07350015.2019.1665531

MA, S. and HUANG, J. (2017). A concave pairwise fusion approach to subgroup analysis. *J. Amer. Statist. Assoc.* **112** 410–423. MR3646581 https://doi.org/10.1080/01621459.2016.1148039

MARIE, C., VENDITTELLI, F. and SAUVANT-ROCHAT, M. P. (2015). Obstetrical outcomes and biomarkers toassess exposure to phthalates: A review. *Environ. Int.* **83** 116–136.

MARKOWITZ, H. M. and MANNE, A. S. (1957). On the solution of discrete programming problems. *Econometrica* **25** 84–110. MR0085968 https://doi.org/10.2307/1907744

MARSEE, K., WOODRUFF, T. J., AXELRAD, D. A., CALAFAT, A. M. and SWAN, S. H. (2006). Estimated dailyphthalate exposures in a population of mothers of male infants exhibiting reduced anogenital distance. *Environ. Health Perspect.* **114** 805–809.

NEEDELL, D. and TROPP, J. A. (2009). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.* **26** 301–321. MR2502366 https://doi.org/10.1016/j.acha.2008.07.002

NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization*: *A Basic Course*. *Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598 https://doi.org/10.1007/978-1-4419-8853-9

NESTEROV, Y. (2013). Gradient methods for minimizing composite functions. *Math. Program.* **140** 125–161. MR3071865 https://doi.org/10.1007/s10107-012-0629-5

SCHETTLER, T. (2006). Human exposure to phthalates via consumer products. *J. Androl.* **29** 134–139.

SHEN, J. and HE, X. (2015). Inference for subgroup analysis with a structured logistic-normal mixture model. *J. Amer. Statist. Assoc.* **110** 303–312. MR3338504 https://doi.org/10.1080/01621459.2014.894763

SHEN, X. and HUANG, H.-C. (2010). Grouping pursuit through a regularization solution surface. *J. Amer. Statist. Assoc.* **105** 727–739. MR2724856 https://doi.org/10.1198/jasa.2010.tm09380

SHEN, X., PAN, W., ZHU, Y. and ZHOU, H. (2013). On constrained and regularized high-dimensional regression. *Ann. Inst. Statist. Math.* **65** 807–832. MR3105798 https://doi.org/10.1007/s10463-012-0396-3

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641 https://doi.org/10.1111/j.1467-9868.2005.00490.x

VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

VIELMA, J. P. and NEMHAUSER, G. L. (2011). Modeling disjunctive constraints with a logarithmic number of binary variables and constraints. *Math. Program.* **128** 49–72. MR2810952 https://doi.org/10.1007/s10107-009-0295-4

WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory* **55** 5728–5741. MR2597190 https://doi.org/10.1109/TIT.2009.2032816

WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

WANG, W., WU, S., ZHU, Z., ZHOU, L. and SONG, P. X. (2024). Supplement to "Supervised homogeneity fusion: A combinatorial approach." https://doi.org/10.1214/23-AOS2347SUPP

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHOU, L., SUN, S., FU, H. and SONG, P. X.-K. (2022). Subgroup-effects models for the analysis of personal treatment effects. *Ann. Appl. Stat.* **16** 80–103. MR4400504 https://doi.org/10.1214/21-aoas1503

ZHU, Y., SHEN, X. and PAN, W. (2013). Simultaneous grouping pursuit and feature selection over an undirected graph. *J. Amer. Statist. Assoc.* **108** 713–725. MR3174654 https://doi.org/10.1080/01621459.2013.770704

ZHU, Z. and WU, S. (2021). On the early solution path of best subset selection. Preprint. Available at arXiv:2107.06939.