Strong Screening Rules for Group-based SLOPE Models

Fabio Feser
Imperial College London

Marina Evangelou
Imperial College London

Abstract

Tuning the regularization parameter in penalized regression models is an expensive task, requiring multiple models to be fit along a path of parameters. Strong screening rules drastically reduce computational costs by lowering the dimensionality of the input prior to fitting. We develop strong screening rules for group-based Sorted L-One Penalized Estimation (SLOPE) models: Group SLOPE and Sparse-group SLOPE. The developed rules are applicable to the wider family of group-based OWL models, including OSCAR. Our experiments on both synthetic and real data show that the screening rules significantly accelerate the fitting process. The screening rules make it accessible for group SLOPE and sparse-group SLOPE to be applied to high-dimensional datasets, particularly those encountered in genetics.

1 INTRODUCTION

As the amount of data collected increases, the emergence of high-dimensional data, where the number of features (p) is much larger than the number of observations (n), is becoming increasingly common in fields ranging from genetics to finance. Performing regression and discovering relevant features on these datasets is a challenging task, as classical statistical methods tend to break down. The most popular approach to meeting this challenge is the lasso (Tibshirani, 1996), which has given rise to the general penalized regression framework

$$\hat{\beta}(\lambda) \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ f(\beta) + \lambda J(\beta; v) \right\}, \tag{1}$$

where f is a differentiable and convex loss function, J is a convex penalty norm, v are penalty weights, and $\lambda > 0$ is the regularization parameter.

A key aspect of fitting a penalized model is to tune the value of λ along an l-length path $\lambda_1 \geq \ldots \geq \lambda_l \geq 0$. Several approaches exist for tuning this parameter, including cross-validation (Homrighausen and McDonald, 2018; Chetverikov et al., 2021) and exact solution path algorithms (Efron et al., 2004), but these can be computationally expensive. Screening rules help alleviate these costs by discarding variables that are inactive at the optimal solution, thus reducing input dimensionality prior to optimization.

Denote the *active set* of coefficients at path point λ_{k+1}, for $k \in [l-1] := \{1, \ldots, l-1\}$, by $\mathcal{A}_v(\lambda_{k+1}) = \{i \in [p] : \hat{\beta}_i(\lambda_{k+1}) \neq 0\}$. The goal of a (sequential) screening rule is to use the solution at λ_k to recover a *screened set* of features, $\mathcal{S}_v(\lambda_{k+1})$, which is a superset of $\mathcal{A}_v(\lambda_{k+1})$. The screened set is then used as input for calculating the fitted values, leading to significant computational savings.

There are two types of screening rules: *safe* and *heuristic*. Safe rules are guaranteed to only discard inactive variables and mostly follow the Safe Feature Elimination (SAFE) framework (El Ghaoui et al., 2010), in which safe regions for variables are constructed. Other notable examples include Slores (Wang et al., 2014), the dome test (Xiang and Ramadge, 2012), and Dual Polytope Projections (DPP) (Wang et al., 2013), as well as sample screening (Shibagaki et al., 2016) and other examples given in Ogawa et al. (2013); Fercoq et al. (2015); Atamturk and Gomez (2020).

Heuristic rules tend to follow the strong screening rule framework, proposed by Tibshirani et al. (2010), and discard considerably more variables than safe rules. However, they can incorrectly discard active variables, so are complemented by a check of the Karush–Kuhn–Tucker (KKT) (Kuhn and Tucker, 1950) stationarity conditions. Other strong rules include Blitz (Johnson and Guestrin, 2015), SIS (Fan and Lv, 2008), and ExSIS (Ahmed and Bajwa, 2019). Hybrid schemes exist, using both safe and heuristic

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

rules (Zeng et al., 2021; Wang and Breheny, 2022).

A strong screening rule is formulated through the KKT stationarity conditions for Equation 1, given by

$$\mathbf{0} \in \nabla f(\beta) + \lambda \partial J(\beta; v). \tag{2}$$

If the gradient were available, the active set could be identified exactly by checking the subdifferential of the norm at zero:

$$\nabla f(\beta) \in \lambda \partial J(\mathbf{0}; v) = \{x \in \mathbb{R}^p : J^*(x; \lambda v) \leq 1\}, \tag{3}$$

where J^* is the dual norm of J and $\partial J(\mathbf{0}; v)$ is the unit ball of the dual norm. So, $J^*(\nabla f(\beta); \lambda v) \leq 1$ indicates that $\beta = 0$. As the gradient at λ_{k+1} is not available, a model-specific approximation of the gradient is derived to find a screened subset of the features, $\mathcal{S}_v(\lambda_{k+1})$, such that $\mathcal{A}_v(\lambda_{k+1}) \subset \mathcal{S}_v(\lambda_{k+1})$.

1.1 Screening Approaches for SLOPE

As the lasso is inconsistent under certain scenarios (Zou, 2006), several adaptive extensions have been proposed, including the Sorted L-One Penalized Estimation (SLOPE) model (Bogdan et al., 2015). SLOPE applies the sorted ℓ_1 norm $J_{\text{slope}}(\beta; v) = \sum_{i=1}^{p} v_i |\beta|_{(i)}$, where $v_1 \geq \ldots \geq v_p \geq 0, |\beta|_{(1)} \geq \ldots \geq |\beta|_{(p)}$. One key advantage of SLOPE is its ability to control the variable false discovery rate (FDR) under orthogonal data. Additional powerful properties include: it clusters strongly correlated features, it finds the minimum total squared error loss across different sparsity levels, removing the need for prior knowledge of sparsity, and it is asymptotically minimax (Figueiredo and Nowak, 2014; Su and Candès, 2016). All of these useful properties have meant that SLOPE has found widespread use in machine learning and genetics (Gossmann et al., 2015; Virouleau et al., 2017; Kremer et al., 2020; Frommlet et al., 2022; Riccobello et al., 2023).

Both safe (Bao et al., 2020; Elvira and Herzet, 2021) and strong (Larsson et al., 2020) rules have been proposed for SLOPE, as well as exact solution path algorithms (Nomura, 2020; Dupuis and Tardivel, 2023). As SLOPE is a non-separable penalty, safe screening requires repeated screening during optimization, which is expensive due to the repeated dual norm evaluations required for the safe regions (Larsson et al., 2020).

Group-based SLOPE Models In genetics, the analysis of grouped features is frequently encountered, as genes are grouped into pathways for the completion of a specific biological task. To use this grouping information, SLOPE has been extended to group (gSLOPE) and sparse-group (SGS) regression.

For a set of m non-overlapping groups, $\mathcal{G}_1, \ldots, \mathcal{G}_m$ of sizes $p_1, \ldots, p_m$, Group SLOPE (gSLOPE) (Gossmann et al., 2015; Brzyski et al., 2019) is given by

$$J_{\text{gslope}}(\beta; w) = \sum_{g=1}^{m} \sqrt{p_g} w_g \|\beta^{(g)}\|_2, \tag{4}$$

such that $\beta^{(g)} \in \mathbb{R}^{p_g}$ is a vector of the group coefficients. The norm has ordered penalty weights $w_1 \geq \ldots \geq w_g \geq 0$ (described in Appendix A.1) which are matched to $\sqrt{p_1}\|\beta^{(1)}\|_2 \geq \ldots \geq \sqrt{p_m}\|\beta^{(m)}\|_2$.

Sparse-group SLOPE (SGS) (Feser and Evangelou, 2023) was further proposed as a convex combination of SLOPE and gSLOPE for concurrent variable and group selection. For $\alpha \in [0, 1]$, with weights (v, w) (described in Appendix B.1), the norm is given by

$$J_{\text{sgs}}(\beta; \alpha, v, w) = \alpha J_{\text{slope}}(\beta; v) + (1 - \alpha) J_{\text{gslope}}(\beta; w).$$

Both approaches control the FDR under orthogonal data: gSLOPE at the group-level (Brzyski et al., 2019) and SGS at both levels (Feser and Evangelou, 2023). SGS has been found to outperform other methods at selection and prediction (Feser and Evangelou, 2023).

1.2 Contributions

No screening rules exist for group-based SLOPE models. The strong screening rule framework introduced by Tibshirani et al. (2010) facilitated the extension from the lasso to the group lasso by requiring only knowledge of the dual norm. However, this framework is restricted to separable penalties, making it unsuitable for SLOPE. Instead, in Larsson et al. (2020), the subdifferential of SLOPE is used to derive strong rules.

Motivated by this, we propose a new strong screening framework for sparse-group norms (Section 2), used to develop screening for gSLOPE (Section 3) and SGS (Section 4) (with proofs of the results provided in Appendices A.3 and B.3). The framework applies two screening layers, drastically reducing dimensionality (Figure 1). The screening requires knowledge of the subdifferentials, necessitating a general derivation for gSLOPE (Theorem 3.1).

The choice of strong screening over safe is motivated by two main reasons. First, strong rules discard significantly more variables than safe rules (Tibshirani et al., 2010). Second, safe rules require calculating the dual norm set. For SGS, this is a sum of convex sets. Determining if a point lies in this set requires knowledge of the summation's decomposability, which is a difficult task (Wang and Ye, 2014). This challenge is addressed in Section 4.1.

The analysis of synthetic and real datasets shows that our proposed screening rules considerably improve runtime (Section 6). The reduced input dimensionality from the screening also eases convergence issues

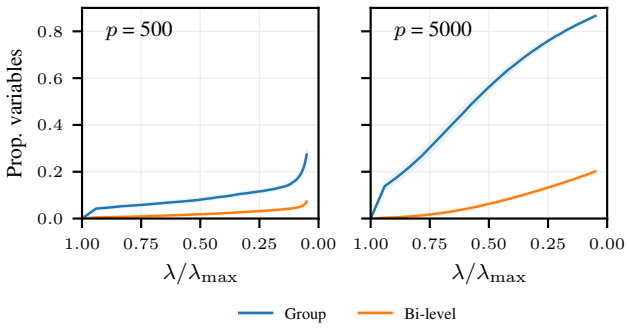


Figure 1: The proportion of variables in $\mathcal{S}_v$ relative to p for group-only and bi-level screening applied to SGS, plotted along the regularization path with 95% confidence intervals. Synthetic data was generated under a linear model for $p = 500, 5000$ (Section 6.1), with results averaged over 100 repetitions.

with large datasets. These improvements are achieved without affecting solution optimality.

As the proposed screening rules only require that the penalty sequences (v, w) are ordered, they apply to the wider class of group-based Ordered Weighted ℓ_1 (OWL) models. Popular special cases of OWL models include SLOPE and the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) model (Bondell and Reich, 2008) (the description of group-based OSCAR models is provided in Section 5).

Notation The sets of active and inactive groups are given by $\mathcal{A}_g = \{g \in [m] : \|\hat{\beta}^{(g)}\|_2 \neq 0\}$ and $\mathcal{Z} = \{g \in [m] : \|\hat{\beta}^{(g)}\|_2 = 0\}$. Their corresponding set of variable indices are denoted by $\mathcal{G}_\mathcal{A}$ and $\mathcal{G}_\mathcal{Z}$. The cardinality of a vector x is denoted by $\text{card}(x)$. The operators $(\cdot)_\downarrow$ and $(\cdot)_{|\downarrow|}$ sort a vector into decreasing and decreasing absolute form. We use $\preceq$ to denote element-wise inequality signs. The operator $\mathcal{O}(\cdot)$ returns the index of a vector sorted into decreasing absolute form. The cumulative summation operator applied on a vector is denoted by $\text{cumsum}(x) = [x_1, x_1 + x_2, \dots, \sum_{i=1}^{\text{card}(x)} x_i]$.

2 SPARSE-GROUP STRONG SCREENING

Sparse-group models, such as SGS and the sparse-group lasso (SGL) (Simon et al., 2013), apply both variable and group penalization so that bi-level screening is possible. Safe rules that perform bi-level screening exist for SGL (Wang and Ye, 2014; Ndiaye et al., 2016a), but there are no such strong rules. The strong screening framework (Tibshirani et al., 2010) does not extend to sparse-group or non-separable norms, and the strong rule derived for SGL in Liang et al. (2022) applies only group-level screening.

Framework We introduce a new sparse-group framework (Algorithm 1), based on the strong framework by Tibshirani et al. (2010), for applying strong screening to sparse-group norms, allowing for bi-level screening. By applying bi-level screening for SGS, a substantially larger proportion of variables are discarded than with group screening alone (Figure 1).

First, a screened set of groups is computed, $\mathcal{S}_g$. An additional layer of screening is then performed to compute $\mathcal{S}_v$ using $\mathcal{S}_g$. This, combined with the previously active variables, forms the reduced input set for fitting, $\mathcal{E}_v$. KKT checks are performed on $\mathcal{E}_v$ to ensure no violations have occurred (Appendices A.4 and B.4). Based on this framework, the SGS rules are derived in Section 4. The screening for gSLOPE (Section 3) also uses this framework but does not perform the variable screening and instead takes $\mathcal{S}_v$ as all variables in the groups of $\mathcal{S}_g$. The KKT checks are then performed only on the groups. Appendix D describes the implementation of the framework for gSLOPE and SGS.

The main cost of the framework is calculating the fitted values. For t iterations of ATOS (a proximal algorithm used to fit SGS, see Section 6.1), the convergence rate is $O(1/t)$ (Pedregosa and Gidel, 2018) and we expect a time complexity of $O(tp^2)$ for a proximal algorithm (Zhao and Huo, 2023).

Algorithm 1 Sparse-group screening framework

Input: $(\lambda_1, \dots, \lambda_l) \in \mathbb{R}^l$, $\mathbf{X} \in \mathbb{R}^{n \times p}, y \in \mathbb{R}^n$
for $k = 1$ **to** $l - 1$ **do**
  $\mathcal{S}_g(\lambda_{k+1}) \leftarrow$ group screening on full input
  $\mathcal{S}_v(\lambda_{k+1}) \leftarrow$ variable screening on $g \in \mathcal{S}_g(\lambda_{k+1})$
  $\mathcal{E}_v \leftarrow \mathcal{S}_v(\lambda_{k+1}) \cup \mathcal{A}_v(\lambda_k)$
  compute $\hat{\beta}_{\mathcal{E}_v}(\lambda_{k+1})$
  $\mathcal{K}_v \leftarrow$ variable KKT violations on $\hat{\beta}(\lambda_{k+1})$
  while $\text{card}(\mathcal{K}_v) > 0$ **do**
    $\mathcal{E}_v \leftarrow \mathcal{E}_v \cup \mathcal{K}_v$
    compute $\hat{\beta}_{\mathcal{E}_v}(\lambda_{k+1})$
    $\mathcal{K}_v \leftarrow$ variable KKT violations on $\hat{\beta}(\lambda_{k+1})$
  end while
end for
Output: $\hat{\beta}(\lambda_1), \dots, \hat{\beta}(\lambda_l) \in \mathbb{R}^p$

3 GROUP SLOPE

The strong rule for gSLOPE is formulated by checking the zero condition of the subdifferential (as per Equation 3) derived in Theorem 3.1. To derive the subdifferential, we define the operator

$$[b]_{\mathcal{G},q} := (p_1^q \|b^{(1)}\|_2, \dots, p_m^q \|b^{(m)}\|_2)^\top.$$

In particular, $[b]_{\mathcal{G}_\mathcal{Z},-0.5}$ is the operator applied only to the inactive groups using the quotient $q = -0.5$.

Theorem 3.1 (gSLOPE subdifferential)**.** *The subdifferential for gSLOPE is given by*

$$\partial J_{gslope}(\beta; w) = \begin{cases} \left\{x \in \mathbb{R}^{\operatorname{card} \mathcal{G}_\mathcal{Z}} : \right. \\ \left. [x]_{\mathcal{G}_\mathcal{Z},-0.5} \in \partial J_{slope}(0; w_\mathcal{Z})\right\}, & \text{at 0.} \\ \left\{ w_g \sqrt{p_g} \dfrac{\beta^{(g)}}{\|\beta^{(g)}\|_2} \right\}, & \text{otherwise.} \end{cases}$$

The choice of $q = -0.5$ leads to $J^*_{\text{gslope}}(x; w) = J^*_{\text{slope}}([x]_{\mathcal{G},-0.5})$, which allows the gSLOPE subdifferential to be written in terms of the SLOPE one (Brzyski et al., 2019). Combining the KKT conditions at zero (Equation 3) with the gSLOPE subdifferential (Theorem 3.1) reveals that a group is inactive if $h(\lambda) := ([\nabla f(\hat{\beta}(\lambda))]_{\mathcal{G},-0.5})_\downarrow \in \partial J_{\text{slope}}(\mathbf{0}; \lambda w)$. Using the subdifferential of SLOPE (Appendix A.2) (Larsson et al., 2020), this is given by

$$\operatorname{cumsum}(h(\lambda) - \lambda w) \preceq \mathbf{0}. \tag{5}$$

This condition is checked efficiently using the algorithm proposed for the SLOPE strong rule (Algorithm A1) leading to the strong rule for gSLOPE (Proposition 3.2). The algorithm assumes that the indices for the inactive predictors will be ordered last in the input c and the features $|\hat{\beta}|_\downarrow$ (Larsson et al., 2020).

Proposition 3.2 (Strong screening rule for gSLOPE)**.** *Taking $c = h(\lambda_{k+1})$ and $\phi = \lambda_{k+1} w$ as inputs for Algorithm A1 returns a superset $\mathcal{S}_g(\lambda_{k+1})$ of the active set $\mathcal{A}_g(\lambda_{k+1})$.*

The gradient at path index $k+1$ is not available for the computation of $h(\lambda_{k+1})$, so an approximation is required that does not lead to any violations in Algorithm A1. By the cumsum condition in this algorithm, an approximation for a group $g \in [m]$ is sought such that $h_g(\lambda_{k+1}) \leq h_g(\lambda_k) + R_g$, where $R_g \geq 0$ needs to be determined. An approximation is found by assuming that $h_g(\lambda_{k+1})$ is a Lipschitz function of λ_{k+1} with respect to the ℓ_1 norm, that is,

$$|h_g(\lambda_{k+1}) - h_g(\lambda_k)| \leq w_g |\lambda_{k+1} - \lambda_k|.$$

By again noting that $J^*_{\text{gslope}}(x) = J^*_{\text{slope}}([x]_{\mathcal{G},-0.5})$, it can be seen that the assumption is equivalent to the Lipschitz assumptions used for the lasso and SLOPE strong rules (Tibshirani et al., 2010; Larsson et al., 2020). By the reverse triangle inequality,

$$|h_g(\lambda_{k+1})| \leq |h_g(\lambda_k)| + \lambda_k w_g - \lambda_{k+1} w_g,$$

leading to the choice $R_g = \lambda_k w_g - \lambda_{k+1} w_g$ and the gradient approximation strong screening rule (Proposition 3.3). To apply gSLOPE screening in practice (Section 6.1), Proposition 3.4 describes the calculation of the first path value.

Proposition 3.3 (Gradient approximation strong screening rule for gSLOPE)**.** *Taking $c = h(\lambda_k) + \lambda_k w - \lambda_{k+1} w$ and $\phi = \lambda_{k+1} w$ as inputs for Algorithm A1, and assuming that for any $k \in [l-1]$,*

$$|h_g(\lambda_{k+1}) - h_g(\lambda_k)| \leq w_g |\lambda_{k+1} - \lambda_k|, \ \forall g = [m],$$

and $\mathcal{O}(h(\lambda_{k+1})) = \mathcal{O}(h(\lambda_k))$, then the algorithm returns a superset $\mathcal{S}_g(\lambda_{k+1})$ of the active set $\mathcal{A}_g(\lambda_{k+1})$.

Proposition 3.4 (gSLOPE path start)**.** *For gSLOPE, the path value at which the first group enters the model is given by*

$$\lambda = \max \left\{ \operatorname{cumsum}\left(([\nabla f(\mathbf{0})]_{\mathcal{G},-0.5})_\downarrow \right) \oslash \operatorname{cumsum}(w) \right\},$$

where $\oslash$ denotes Hadamard division.

4 SPARSE-GROUP SLOPE

This section presents the group and variable screening rules for SGS. They are derived using the SGS KKT conditions, formulated in terms of SLOPE and gSLOPE (by the sum rule of subdifferentials):

$$\nabla f(\beta) \in \lambda\alpha \partial J_{\text{slope}}(\beta; v) + \lambda(1-\alpha) \partial J_{\text{gslope}}(\beta; w). \tag{6}$$

4.1 Group Screening

For inactive groups, the KKT conditions (Equation 6) for SGS are

$$\begin{aligned} &(\nabla f(\beta) + \lambda\alpha \partial J_{\text{slope}}(\mathbf{0}; v))_{\mathcal{G}_\mathcal{Z}} \in \lambda(1-\alpha) \partial J_{\text{gslope}}(\mathbf{0}; w_\mathcal{Z}), \\ &\underset{\text{Equation 5}}{\Longrightarrow} \operatorname{cumsum}\Big(\left([\nabla f(\beta) + \lambda\alpha \partial J_{\text{slope}}(\mathbf{0}; v)]_{\mathcal{G}_\mathcal{Z},-0.5} \right)_\downarrow \\ &\qquad - \lambda(1-\alpha) w_\mathcal{Z} \Big) \preceq \mathbf{0}. \end{aligned} \tag{7}$$

The problem reduces to a form similar to the gSLOPE screening rule (Section 3), with inputs for Algorithm A1 given by $c = ([\nabla f(\beta) + \lambda\alpha \partial J_{\text{slope}}(\mathbf{0}; v)]_{\mathcal{G},-0.5})_\downarrow$ and $\phi = \lambda(1-\alpha) w$.

To determine the form of the quantity $\partial J_{\text{slope}}(\mathbf{0}; v)$, the term inside the $[\cdot]_{\mathcal{G}_\mathcal{Z},-0.5}$ operator needs to be as small as possible for Equation 7 to be satisfied. This term is found to be the soft thresholding operator, $S(\nabla f(\beta), \lambda\alpha v) := \operatorname{sign}(\nabla f(\beta))(|\nabla f(\beta)| - \lambda\alpha v)_+$ by Lemma 4.1 (see Appendix B.2 for the proof).

Lemma 4.1. *In Equation 7, choosing $\partial J_{slope}(\mathbf{0}; v) = S(\nabla f(\beta), \lambda\alpha v)$ minimises $[\nabla f(\beta) + \lambda\alpha \partial J_{slope}(\mathbf{0}; v)]_\mathcal{G}$.*

By using the soft-thresholding operator, a valuable connection between SGS and SGL is found, as the operator is used in the gradient update step for SGL (Simon et al., 2013). This connection has the potential to lead to new and more efficient optimization approaches for SGS that are more closely related to those used to

solve SGL, similar to the recent coordinate descent algorithm for SLOPE (Larsson et al., 2022).

Using this operator, the (non-approximated) strong group screening rule for SGS is shown in Proposition B.1. Applying a similar Lipschitz assumption as for the gSLOPE rule gives the gradient approximation strong group screening rule for SGS (Proposition 4.2).

Proposition 4.2 (Gradient approximation strong group screening rule for SGS)**.** *Let* $\tilde{h}(\lambda) := ([S(\nabla f(\hat{\beta}(\lambda)), \lambda\alpha v)]_{\mathcal{G},-0.5})_{\downarrow}$. *Taking* $c = \tilde{h}(\lambda_k) + \lambda_k(1-\alpha)w - \lambda_{k+1}(1-\alpha)w$ *and* $\phi = \lambda_{k+1}(1-\alpha)w$ *as inputs for Algorithm A1, and assuming that for any* $k \in [l-1]$,

$$\left|\tilde{h}_g(\lambda_{k+1}) - \tilde{h}_g(\lambda_k)\right| \leq (1-\alpha)w_g|\lambda_{k+1} - \lambda_k|, \ \forall g = [m],$$

and $\mathcal{O}(\tilde{h}(\lambda_{k+1})) = \mathcal{O}(\tilde{h}(\lambda_k))$*, then the algorithm returns a superset* $\mathcal{S}_g(\lambda_{k+1})$ *of the active set* $\mathcal{A}_g(\lambda_{k+1})$.

4.2 Variable Screening

By exploiting the sparse-group norm of SGS, the input dimensionality can be reduced further with a second layer of variable screening. The KKT conditions (Equation 6) for zero variables in active groups are

$$\nabla_{\mathcal{G}_{\mathcal{A}_g}} f(\beta) \in \lambda\alpha\partial J_{\text{slope}}(\mathbf{0}; v_{\mathcal{G}_{\mathcal{A}_g}}). \tag{8}$$

The gSLOPE subdifferential term vanishes as the numerator is zero in Theorem 3.1. The problem reduces to that of SLOPE screening, applied to the variables in groups in $\mathcal{A}_g$ and scaled by α. The gradient approximated rule is shown in Proposition 4.3 (see Proposition B.2 for the non-approximated version).

Proposition 4.3 (Gradient approximation strong variable screening rule for SGS)**.** *Let* $\bar{h}(\lambda) := (\nabla f(\hat{\beta}(\lambda)))_{|\downarrow|}$. *Taking* $c = |\bar{h}(\lambda_{k+1})| + \lambda_k\alpha v - \lambda_{k+1}\alpha v$ *and* $\phi = \lambda_{k+1}\alpha v$ *for the variables in the groups in* $\mathcal{A}_g(\lambda_{k+1})$ *as inputs for Algorithm A1, and assuming that for any* $k \in [l-1]$,

$$\left|\bar{h}_j(\lambda_{k+1}) - \bar{h}_j(\lambda_k)\right| \leq \alpha v_j|\lambda_{k+1} - \lambda_k|, \forall j \in \mathcal{G}_{\mathcal{A}_g(\lambda_{k+1})},$$

and $\mathcal{O}(\bar{h}(\lambda_{k+1})) = \mathcal{O}(\bar{h}(\lambda_k))$*, then the algorithm returns a superset* $\mathcal{S}_v(\lambda_{k+1})$ *of* $\mathcal{A}_v(\lambda_{k+1})$.

In practice $\mathcal{A}_g(\lambda_{k+1})$ is not available, as this is exactly what we are trying to superset with any screening rule. However, Proposition 4.2 guarantees that it is contained in $\mathcal{S}_g(\lambda_{k+1})$ so that this can be used instead. To apply SGS in practice (Section 6.1), Proposition 4.4 describes the calculation of the first path value.

Proposition 4.4 (SGS path start)**.** *For SGS, the path value at which the first variable enters the model is*

$$\lambda = \max\{\text{cumsum}(|\nabla f(\mathbf{0})|_{\downarrow}) \oslash \text{cumsum}((1-\alpha)\tau\omega - \alpha v)\},$$

where τ *and* ω *are expanded vectors of the group sizes* $(\sqrt{p_g})$ *and penalty weights* (w_g) *to* p *dimensions.*

5 GROUP-BASED OWL

The screening rule framework presented are also applicable to the wider class of OWL models. The Ordered Weighted ℓ_1 (OWL) framework is defined as (Zeng and Figueiredo, 2014a)

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{\nabla f(\beta) + \lambda J_{\text{owl}}(\beta; v)\right\},$$

where $J_{\text{owl}}(\beta; v) = \sum_{i=1}^{p} v_i|\beta|_{(i)}$, $|\beta|_{(1)} \geq \ldots \geq |\beta|_{(p)}$, and v are non-negative non-increasing weights. SLOPE is a special case of OWL where the weights are taken to be the Benjamini-Hochberg critical values (Bogdan et al., 2015). Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) (Bondell and Reich, 2008) is a further special case of OWL (often referred to as OWL with linear decay) where for a variable $i \in [p]$, the weights are taken to be $v_i = \sigma_1 + \sigma_2(p-i)$, and σ_1, σ_2 are to be set. In Bao et al. (2020) they are set to $\sigma_1 = d_i\|\mathbf{X}^\top y\|_\infty, \sigma_2 = \sigma_1/p$, where $d_i = i \times e^{-2}$, $\mathbf{X}^\top \in \mathbb{R}^{n\times p}$ is the design matrix, and $y \in \mathbb{R}^n$ is the response vector.

Group OSCAR (gOSCAR) and Sparse-group OSCAR (SGO) are defined using the frameworks provided by gSLOPE (Brzyski et al., 2019) and SGS (Feser and Evangelou, 2023), respectively, but instead use the OSCAR weights (see Appendix E.1).

6 RESULTS

This section illustrates the effectiveness of the screening rules for gSLOPE and SGS using synthetic (Section 6.1) and real (Section 6.2) data. References to $\mathcal{E}, \mathcal{A}$ for group and variable metrics denote $\mathcal{E}_g, \mathcal{A}_g$ and $\mathcal{E}_v, \mathcal{A}_v$. For SGS, $\mathcal{E}_g$ represents groups with members in $\mathcal{E}_v$.

6.1 Synthetic Data Analysis

Set up A multivariate Gaussian design matrix, $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \in \mathbb{R}^{400\times p}$, was used and the within-group correlation set to $\Sigma_{i,j} = \rho$, where i and j belong to the same group. The correlation and number of features were varied between $\rho \in \{0, 0.3, 0.6, 0.9\}$ and $p \in \{500, 1625, 2750, 3875, 5000\}$, producing 20 simulation cases. Each simulation case was repeated 100 times. Two models were considered: linear and logistic. For the linear model, the output was generated as $y = \mathbf{X}\beta + \mathcal{N}(0,1)$ and for the logistic model the class probabilities were calculated using $\sigma(\mathbf{X}\beta + \mathcal{N}(0,1))$, where σ is the sigmoid function. Groups of sizes between 3 and 25 were considered, of which 15% were

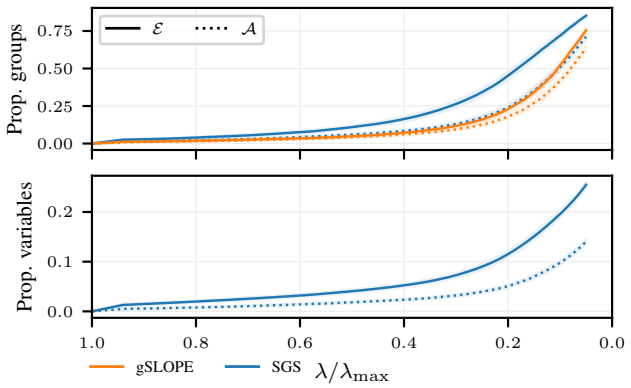


Figure 2: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$, relative to the input, for both gSLOPE and SGS as a function of the path for the linear model with $p = 2750, \rho = 0.6, m = 197$. The results are averaged over 100 repetitions, with 95% confidence intervals shown.

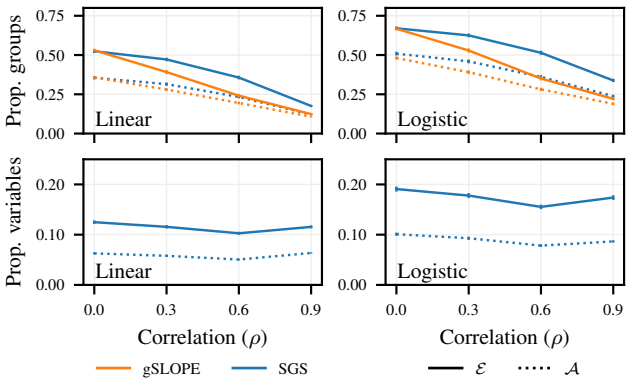


Figure 3: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$, relative to the full input, shown for gSLOPE and SGS. This is shown as a function of the correlation (ρ), averaged over all cases of the input dimension (p), with 100 repetitions for each p, for both linear and logistic models, with standard errors shown.

set to active. Within each active group, 30% of the variables were set to active with signal $\beta \sim \mathcal{N}(0, 5)$. gSLOPE and SGS were fit along a log-linear path of 50 regularization parameters using warm starts, beginning at λ_1 (from Propositions 3.4 and 4.4), and ending at $\lambda_{50} = 0.05\lambda_1$. The data was ℓ_2 standardized and for the linear model an intercept was used. Both models had FDR-control parameters set to 0.05, and $\alpha = 0.95$ for SGS. The models were fit using the adaptive three operator splitting (ATOS) algorithm (Pedregosa and Gidel, 2018), although the screening rules can be applied with any fitting algorithm. Additional computational details are in Appendix F.1.

Primarily, the results for the linear model are presented, and the results for the logistic model are in Appendix F.3.2. The simulations were repeated for group-based OSCAR models (Appendix E).

Screening Efficiency By comparing the sizes of the fitting set ($\mathcal{E}$) to the active set ($\mathcal{A}$), the screening rules are found to be efficient in providing dimensionality reduction close to the minimum possible (the active set size) (Figures 2 and A9). As expected, the sets increase in size as λ decreases, and the difference in size between the sets remains stable along the path, decreasing towards the termination point. The size of the fitting set remains far below the input size across the whole path, showing the benefit of the screening. This is found to be true for any correlation, input dimensionality, and model considered (Appendix F.3).

The screening rules perform well for linear and logistic models (Figure 3), showing robust dimensionality reduction for all correlation cases considered. As the correlation increases, the signal concentrates in fewer groups, causing the active group set to decrease in size. SLOPE models deal well with highly correlated features, as the sorted norm clusters them together (Zeng and Figueiredo, 2014b).

The screening rules are found to efficiently reduce the input dimensionality on average across all cases considered (Table 2).

Table 1: Runtime (in seconds) for fitting 50 models along a path, shown for screening against no screening, for the linear and logistic models. The results are averaged across all cases of the correlation (ρ) and dimensionality (p), with standard errors shown.

Method	Type	Screen (s)	No screen (s)
gSLOPE	Linear	1016 ± 21	1623 ± 27
gSLOPE	Logistic	814 ± 8	1409 ± 11
SGS	Linear	735 ± 15	1830 ± 34
SGS	Logistic	407 ± 2	859 ± 6

Runtime Performance A key metric of performance for a screening rule is the time taken to fit a path of models. Figure 4 shows the significant runtime improvements our screening rules provide as a function of increasing input dimensionality. The gain is observed to be substantial under all correlation cases. Applying screening improves the scaleability of applying gSLOPE and SGS models to larger datasets.

The clear benefit and robustness of our screening method, with regards to runtime, can be seen by aggregating the results of all the simulation cases (Table 1). For both models, in the linear and logistic cases, screening substantially improves the computa-

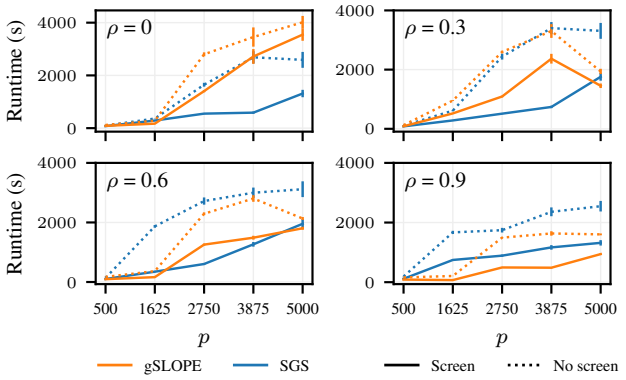


Figure 4: Runtime (in seconds) for fitting 50 models along a path, shown for screening against no screening as a function of p, broken down into different correlation cases, for the linear model. The results are averaged over 100 repetitions, with standard errors shown.

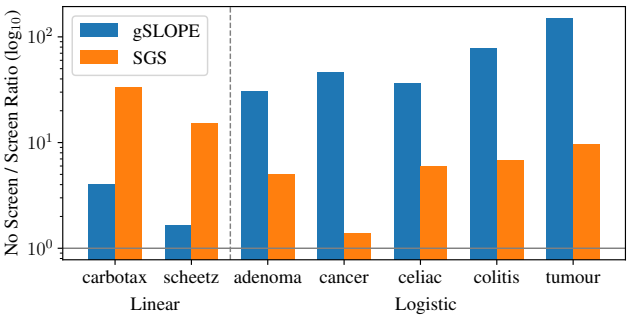


Figure 5: The ratio of no screen time to screen time (↑) of gSLOPE and SGS applied to the real datasets, for fitting 100 path models, split into response type. The horizontal grey line represents no screening improvement.

tional cost, halving the runtime for SGS models.

6.2 Real Data Experiments

Datasets The screening rules were applied to seven real gene datasets, of different response types and dimensionality. Two of the datasets, carbotax and sheetz, had a continuous response so were fit using a linear model. For these, the groups were generated using K-means clustering (Lloyd, 1982). The remaining five (adenoma, cancer, celiac, colitis, and tumour), had binary labels, so a logistic model was used. For these, the design matrices contained gene expression data downloaded from NCBI's GEO database (Edgar et al., 2002), so the genes could be assigned to pathways (groups) using the C3 regulatory target gene sets from MSigDB (Subramanian et al., 2005; Liberzon et al., 2011). All datasets were high-dimensional. See Appendix F.4 for full details.

Both gSLOPE and SGS were fit with their FDR-control parameters set to 0.01 and for SGS $\alpha = 0.99$. Each model was applied along a path of 100 regularization parameters, with $\lambda_{100} = 0.01\lambda_1$. Table A1 describes the algorithmic parameters used for ATOS.

Results For both gSLOPE and SGS, the screening rules led to considerably faster runtimes for all datasets (Figure 5). The screening was more effective for SGS under a linear model and for gSLOPE under a logistic model. The active set sizes for gSLOPE tended to be smaller for the logistic models, allowing for larger feature reduction (Table 2). For SGS, this trend went the other way; the active sets were smaller for the continuous datasets. Another explanation for the reduced screening efficiency for gSLOPE under linear models was due to the grouping structure. For the continuous responses, the groups were generated using K-means clustering, leading to fewer groups and less opportunity for dimensionality reduction (Table A8). With fewer groups, gSLOPE is less likely to discard full groups, as they may contain some signal.

The feature reduction provided by our screening rules led to the alleviation of convergence issues (Table A10). SGS failed to converge for three datasets without screening and none with screening. gSLOPE experienced failed convergences across all datasets without screening, while with screening, it only failed to converge for three datasets, each showing fewer instances of failure. As gSLOPE applies no variable penalization, it is forced to fit all variables within a group. For datasets with large groups, such as those considered here, this leads to a problematic fitting process which can include many noisy variables. Our screening rules help gSLOPE partially overcome this issue, leading to large computational savings and better solution optimality.

The analysis of the real data further illustrates the benefits of the bi-level screening to the runtime and performance of SGS (for synthetic data, see Figure 1). Figure 6 illustrates that for the cancer and celiac datasets, the bi-level screening allows the input dimensionality for SGS to be reduced to a much greater extent than by just group screening (see Figure A17 for the other datasets).

Both screening rules drastically reduce the input dimensionality on all the real datasets (Table 2). On average, for the logistic real data models, the SGS screening rules reduced the input to just 7% of the total space. This comes without affecting solution consistency (Appendices F.2 and F.5). For additional metrics and comparison to the SLOPE strong rule for the real datasets, see Figures A14, A15, and A16. Our rules for gSLOPE and SGS are found to offer similar

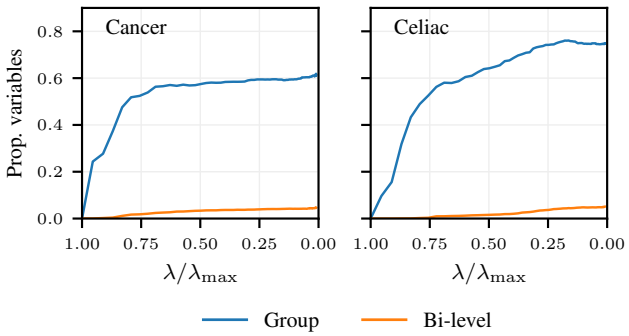


Figure 6: The proportion of variables in $\mathcal{S}_v$ relative to the full input for group-only and bi-level screening applied to SGS, plotted along the regularization path for the cancer and celiac datasets.

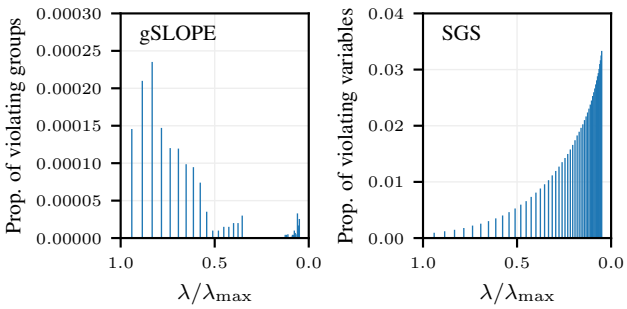


Figure 7: The proportion of KKT violations relative to the full input for gSLOPE and SGS, under linear models, averaged over all synthetic data cases. This is shown as a function of the regularization path.

levels of runtime improvements and feature reduction to that of SLOPE (Larsson et al., 2020).

6.3 KKT Violations

As with any strong screening rule, our approach depends on assumptions that may fail. When this happens, KKT checks are used to ensure no active variables are excluded, with violations added to $\mathcal{E}_v$. For SGS, KKT violations occur at a variable-level (Appendix B.4) and for gSLOPE at a group-level (Appendix A.4).

KKT violations are very rare for gSLOPE (Figure 7), occurring on the simulated data infrequently toward the start of the path. On the real data, only a single dataset had violations, and the number of violations was very small (Table A10).

For SGS, KKT violations are more common, but still infrequent (Figure 7 and Tables A9) due to additional assumptions in the second layer of screening. In Equation 7, minimizing the subdifferential term leads to tighter screened sets, contributing to these violations. In Figure 7, the number of violations increases as a function of the model density for SGS, mirroring the log-linear shape of the regularization path. This pattern is also seen in the strong rule for SLOPE (Larsson et al., 2020).

Table 2: The cardinality of the active ($\mathcal{A}$) and fitting ($\mathcal{E}$) sets for gSLOPE and SGS, averaged across all synthetic and real data cases, split into model type. For gSLOPE, the cardinality is for the group sets, and for SGS the variable ones. Dim. is the input dimensionality. For synthetic, it was $p = 2750$ and $m = 210$.

		Synthetic		Real		
Method	Type	$\mathcal{A}$	$\mathcal{E}$	$\mathcal{A}$	$\mathcal{E}$	Dim.
gSLOPE	Lin.	55±1	76±1	112±3	168±3	240
gSLOPE	Log.	71±1	97±1	49±2	83±3	1634
SGS	Lin.	178±3	364±6	378±37	1139±87	14488
SGS	Log.	230±3	472±6	526±5	979±13	13734

Despite these violations and the computational cost of the checks and refitting (Algorithm 1), the overall screening process yields significant runtime improvements in all cases (Table 1 and Figure 5).

7 DISCUSSION

In this manuscript, we have developed strong screening rules for group-based SLOPE models using our new sparse-group strong screening framework: Group SLOPE and Sparse-group SLOPE, neither of which have any previous screening rules. Our proposed screening rules differ from the existing SLOPE strong rule both in construction and in outcome. The screening rule for gSLOPE screens out irrelevant groups before fitting. The screening rules for SGS perform bi-level screening. Our rules apply to the wider class of OWL models, including group-based OSCAR models.

SLOPE models are finding increasing use in genetics and machine learning, with SGS found to have superior disease prediction performance over other penalized methods (Feser and Evangelou, 2023). Our screening rules will make the group-based versions more accessible by reducing their computational burden. This will allow practitioners to utilize their FDR properties, facilitating their widespread use across various fields.

Through comprehensive analysis of synthetic and real data, we illustrate that the screening rules lead to dramatic improvements in the runtime of gSLOPE and SGS models, as well as for group-based OSCAR models (Appendix E). This is achieved without affecting model accuracy. This is particularly important in datasets where $p \gg n$, such as genetics ones, which is the main motivation behind the proposal of SLOPE

(Bogdan et al., 2015). The screening rules presented in this manuscript allow group-based SLOPE, and more generally group-based OWL models, to achieve computational fitting times more in line with their lasso-based counterparts. The screening rules also helped gSLOPE and SGS overcome convergence issues in large datasets, improving solution optimality.

In our data studies, we have not discovered any scenario where our screening rules did not perform better than no screening. In each case, the rules greatly reduce the input dimensionality and speed up the computational runtime.

Limitations Our screening rules, as any strong rule, rely on assumptions. For both gSLOPE and SGS, Lipschitz assumptions were used that are consistent with those used in the strong screening framework (Tibshirani et al., 2010). For gSLOPE, a Lipschitz assumption was used to derive Proposition 3.3, while for SGS, a separate Lipschitz assumption was made for each layer of screening (Propositions 4.2 and 4.3).

Violations of these assumptions are checked for using the KKT conditions. The SGS KKT checks (Appendix B.4) led to an increased number of violations, as the checks were overly conservative. However, the overall amount of KKT violations for both models was still relatively small, suggesting that these assumptions are not overly restrictive.

An attempt was made to derive less conservative checks (Appendix B.4.3), performed directly on the variables without an initial group check. These were found to be too lenient, incorrectly missing violations. Future work can consider alternative KKT checks.

An additional future direction includes the development of safe rules, which guarantee that only inactive variables are discarded. These could be incorporated into a hybrid scheme together with our strong rules (Zeng et al., 2021; Wang and Breheny, 2022). Deriving safe rules would facilitate a comparison between them and our proposed strong rules, offering further insight into which type of screening is most effective for both non-separable and sparse-group norms.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) through the Modern Statistics and Statistical Machine Learning (StatML) CDT programme, grant no. EP/S023151/1.

References

Talal Ahmed and Waheed U. Bajwa. ExSIS: Extended sure independence screening for ultrahigh-dimensional linear models. *Signal Processing*, 159: 33–48, 2019. ISSN 0165-1684. doi: https://doi.org/10.1016/j.sigpro.2019.01.018.

Alper Atamturk and Andres Gomez. Safe screening rules for ℓ_0-regression from perspective relaxations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 421–430. PMLR, 2020.

Runxue Bao, Bin Gu, and Heng Huang. Fast OSCAR and OWL Regression via Safe Screening Rules. In *Proceedings of the 37th International Conference on Machine Learning*, pages 653–663. PMLR, 2020.

Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. SLOPE—Adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3), 2015. ISSN 1932-6157. doi: 10.1214/15-AOAS842.

Howard D. Bondell and Brian J. Reich. Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics*, 64(1):115–123, 2008. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2007.00843.x.

Damian Brzyski, Alexej Gossmann, Weijie Su, and Małgorzata Bogdan. Group SLOPE – Adaptive Selection of Groups of Predictors. *Journal of the American Statistical Association*, 114(525):419–433, 2019. doi: 10.1080/01621459.2017.1411269.

Michael E Burczynski, Ron L Peterson, Natalie C Twine, Krystyna A Zuberek, Brendan J Brodeur, Lori Casciotti, Vasu Maganti, Padma S Reddy, Andrew Strahs, Fred Immermann, Walter Spinelli, Ulrich Schwertschlag, Anna M Slager, Monette M Cotreau, and Andrew J Dorner. Molecular Classification of Crohn's Disease and Ulcerative Colitis Patients Using Transcriptional Profiles in Peripheral Blood Mononuclear Cells. *The Journal of Molecular Diagnostics*, 8(1):51–61, 2006. ISSN 15251578. doi: 10.2353/jmoldx.2006.050079.

Denis Chetverikov, Zhipeng Liao, and Victor Chernozhukov. On cross-validated Lasso in high dimensions. *The Annals of Statistics*, 49(3):1300 – 1317, 2021. doi: 10.1214/20-AOS2000.

Xavier Dupuis and Patrick J C Tardivel. The Solution Path of SLOPE. Working paper, 2023. URL `https://hal.science/hal-04100441`.

Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30:207–210, 1 2002. ISSN 13624962. doi: 10.1093/nar/30.1.207.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004. doi: 10.1214/009053604000000067.

Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination in sparse supervised learning. Technical Report UCB/EECS-2010-126, EECS Department, University of California, Berkeley, 2010.

Katarzyna A. Ellsworth, Bruce W. Eckloff, Liang Li, Irene Moon, Brooke L. Fridley, Gregory D. Jenkins, Erin Carlson, Abra Brisbin, Ryan Abo, William Bamlet, Gloria Petersen, Eric D. Wieben, and Liewei Wang. Contribution of FKBP5 Genetic Variation to Gemcitabine Treatment and Survival in Pancreatic Adenocarcinoma. *PLoS ONE*, 8(8): e70216, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0070216.

Clément Elvira and Cédric Herzet. Safe rules for the identification of zeros in the solutions of the SLOPE problem. *SIAM Journal on Mathematics of Data Science*, 5(1):147–173, 2021. doi: 10.1137/21m1457631.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008. doi: https://doi.org/10.1111/j.1467-9868.2008.00674.x.

Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Mind the duality gap: safer rules for the Lasso. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 333–342. JMLR, 2015.

Fabio Feser and Marina Evangelou. Sparse-group SLOPE: adaptive bi-level selection with FDR-control. *arXiv preprint arXiv:2305.09467*, 2023.

Mario A. T. Figueiredo and Robert D. Nowak. Sparse estimation with strongly correlated variables using ordered weighted l1 regularization. *arXiv preprint arXiv:1409.4005*, 2014.

Florian Frommlet, Piotr Szulc, Franz König, and Malgorzata Bogdan. Selecting predictive biomarkers from genomic data. *PLOS ONE*, 17(6):1–21, 06 2022. doi: 10.1371/journal.pone.0269369.

Alexej Gossmann, Shaolong Cao, and Yu-Ping Wang. Identification of significant genetic variants via slope, and its extension to group slope. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '15, page 232–240, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450338530.

Graham A Heap, Gosia Trynka, Ritsert C Jansen, Marcel Bruinenberg, Morris A Swertz, Lotte C Dinesen, Karen A Hunt, Cisca Wijmenga, David A vanHeel, and Lude Franke. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Medical Genomics*, 2(1):1, 2009. ISSN 1755-8794. doi: 10.1186/1755-8794-2-1.

Darren Homrighausen and Daniel J. McDonald. A study on tuning parameter selection for the high-dimensional lasso. *Journal of Statistical Computation and Simulation*, 88(15):2865–2892, 2018. doi: 10.1080/00949655.2018.1491575.

Tyler Johnson and Carlos Guestrin. Blitz: A principled meta-algorithm for scaling sparse optimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1171–1179, Lille, France, 07–09 Jul 2015. PMLR.

A Koussounadis, S P Langdon, D J Harrison, and V A Smith. Chemotherapy-induced dynamic gene expression changes in vivo are prognostic in ovarian cancer. *British Journal of Cancer*, 110:2975–2984, 6 2014. ISSN 0007-0920. doi: 10.1038/bjc.2014.258.

Philipp J. Kremer, Sangkyun Lee, Małgorzata Bogdan, and Sandra Paterlini. Sparse portfolio selection via the sorted ℓ_1-norm. *Journal of Banking & Finance*, 110:105687, 2020. ISSN 0378-4266. doi: https://doi.org/10.1016/j.jbankfin.2019.105687.

H W Kuhn and A W Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, Los Angeles, USA, 1950. University of California Press.

Johan Larsson, Małgorzata Bogdan, and Jonas Wallin. The strong screening rule for SLOPE. In *Advances in Neural Information Processing Systems*, volume 33, pages 14592–14603. Curran Associates, Inc., 2020.

Johan Larsson, Quentin Klopfenstein, Mathurin Massias, and Jonas Wallin. Coordinate Descent for SLOPE. *Proceedings of Machine Learning Research*, 206:4802–4821, 2022. ISSN 26403498.

Liang Li, Jian-Wei Zhang, Gregory Jenkins, Fang Xie, Erin E. Carlson, Brooke L. Fridley, William R. Bamlet, Gloria M. Petersen, Robert R. McWilliams, and Liewei Wang. Genetic variations associated with gemcitabine treatment outcome in pancreatic cancer. *Pharmacogenetics and Genomics*, 26(12):527–537, 2016. ISSN 1744-6872. doi: 10.1097/FPC.0000000000000241.

Xiaoxuan Liang, Aaron Cohen, Anibal Solón Heinsfeld, Franco Pestilli, and Daniel J. McDonald.

sparsegl: An R Package for Estimating Sparse Group Lasso. *arXiv preprint arXiv:2208.02942*, 2022.

Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 05 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr260.

S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Xiao-Jun Ma, Zuncai Wang, Paula D Ryan, Steven J Isakoff, Anne Barmettler, Andrew Fuller, Beth Muir, Gayatry Mohapatra, Ranelle Salunga, J.Todd Tuggle, Yen Tran, Diem Tran, Ana Tassin, Paul Amon, Wilson Wang, Wei Wang, Edward Enright, Kimberly Stecker, Eden Estepa-Sabal, Barbara Smith, Jerry Younger, Ulysses Balis, James Michaelson, Atul Bhan, Karleen Habin, Thomas M Baer, Joan Brugge, Daniel A Haber, Mark G Erlander, and Dennis C Sgroi. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell*, 5(6):607–616, 2004. ISSN 15356108. doi: 10.1016/j.ccr.2004.05.015.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. GAP Safe Screening Rules for Sparse-Group Lasso. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016a.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap Safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18, 2016b. ISSN 15337928.

Renato Negrinho and André F T Martins. Orbit regularization. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2, pages 3221–3229. MIT Press, 2014.

Shunichi Nomura. An Exact Solution Path Algorithm for SLOPE and Quasi-Spherical OSCAR. *arXiv preprint arXiv:2010.15511*, 2020.

Kohei Ogawa, Yoshiki Suzuki, and Ichiro Takeuchi. Safe screening of non-support vectors in pathwise svm computation. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1382–1390. PMLR, 2013.

Fabian Pedregosa and Gauthier Gidel. Adaptive three operator splitting. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4085–4094. PMLR, 2018.

Huadong Pei, Liang Li, Brooke L. Fridley, Gregory D. Jenkins, Krishna R. Kalari, Wilma Lingle, Gloria Petersen, Zhenkun Lou, and Liewei Wang. FKBP51 Affects Cancer Cell Response to Chemotherapy by Negatively Regulating Akt. *Cancer Cell*, 16(3):259–266, 2009. ISSN 15356108. doi: 10.1016/j.ccr.2009.07.016.

Riccardo Riccobello, Malgorzata Bogdan, Giovanni Bonaccolto, Philipp J. Kremer, Sandra Paterlini, and Piotr Sobczyk. Sparse graphical modelling via the sorted ℓ_1-norm, 2023.

Jacob Sabates-Bellver, Laurens G. Van der Flier, Mariagrazia de Palo, Elisa Cattaneo, Caroline Maake, Hubert Rehrauer, Endre Laczko, Michal A. Kurowski, Janusz M. Bujnicki, Mirco Menigatti, Judith Luz, Teresa V. Ranalli, Vito Gomes, Alfredo Pastorelli, Roberto Faggiani, Marcello Anti, Josef Jiricny, Hans Clevers, and Giancarlo Marra. Transcriptome Profile of Human Colorectal Adenomas. *Molecular Cancer Research*, 5(12):1263–1275, 2007. ISSN 1541-7786. doi: 10.1158/1541-7786.MCR-07-0267.

Todd E. Scheetz, Kwang-Youn A. Kim, Ruth E. Swiderski, Alisdair R. Philp, Terry A. Braun, Kevin L. Knudtson, Anne M. Dorrance, Gerald F. DiBona, Jian Huang, Thomas L. Casavant, Val C. Sheffield, and Edwin M. Stone. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006. ISSN 0027-8424. doi: 10.1073/pnas.0602562103.

Ulrike Schneider and Patrick Tardivel. The Geometry of Uniqueness, Sparsity and Clustering in Penalized Estimation. *Journal of Machine Learning Research*, 23:1–36, 2022.

Atsushi Shibagaki, Masayuki Karasuyama, Kohei Hatano, and Ichiro Takeuchi. Simultaneous safe screening of features and samples in doubly sparse modeling. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1577–1586. PMLR, 2016.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22 (2):231–245, 2013. ISSN 1061-8600. doi: 10.1080/10618600.2012.681250.

Weijie Su and Emmanuel Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038 – 1068, 2016. doi: 10.1214/15-AOS1397.

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert,

Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102:15545–15550, 10 2005. ISSN 0027-8424. doi: 10.1073/pnas.0506580102.

Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 74(2):245–266, 2010. ISSN 13697412. doi: 10.1111/j.1467-9868.2011.01004.x.

Alain Virouleau, Agathe Guilloux, Stéphane Gaïffas, and Malgorzata Bogdan. High-dimensional robust regression and outliers detection with SLOPE, 2017.

Chuyi Wang and Patrick Breheny. Adaptive hybrid screening for efficient lasso optimization. *Journal of Statistical Computation and Simulation*, 92 (11):2233–2256, 2022. doi: 10.1080/00949655.2021.2025376.

Jie Wang and Jieping Ye. Two-Layer Feature Reduction for Sparse-Group Lasso via Decomposition of Convex Sets. *Advances in Neural Information Processing Systems*, 3:2132–2140, 2014. ISSN 10495258.

Jie Wang, Jiayu Zhou, Peter Wonka, and Jieping Ye. Lasso screening rules via dual Polytope Projection. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 1, pages 1070–1078. Curran Associates Inc., 2013.

Jie Wang, Jiayu Zhou, Jun Liu, Peter Wonka, and Jieping Ye. A Safe Screening Rule for Sparse Logistic Regression. In Z Ghahramani, M Welling, C Cortes, N Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Zhen James Xiang and Peter J Ramadge. Fast lasso screening tests based on correlations. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2137–2140, 2012. doi: 10.1109/ICASSP.2012.6288334.

Xiangrong Zeng and Mário A. T. Figueiredo. Decreasing weighted sorted ℓ_1 regularization. *IEEE Signal Processing Letters*, 21:1240–1244, 2014a.

Xiangrong Zeng and Mário A T Figueiredo. The atomic norm formulation of OSCAR regularization with application to the Frank-Wolfe algorithm. In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pages 780–784, 2014b.

Yaohui Zeng, Tianbao Yang, and Patrick Breheny. Hybrid safe–strong rules for efficient optimization in lasso-type problems. *Computational Statistics & Data Analysis*, 153:107063, 2021. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2020.107063.

Yujie Zhao and Xiaoming Huo. A survey of numerical algorithms that can solve the lasso problems. *WIREs Computational Statistics*, 15(4):e1602, 2023. doi: https://doi.org/10.1002/wics.1602.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 12 2006.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, we have defined our approach and the problem setting.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, we have described the complexity of the fitting algorithm.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes. We have presented our assumptions in the results and have discussed the limitations of these in the discussion.]
   (b) Complete proofs of all theoretical results. [Yes. Proofs have been provided in the Appendix for all of our results.]
   (c) Clear explanations of any assumptions. [Yes, we have extensively discussed our assumptions and any corresponding limitations.]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a

URL). [Yes, code has been provided to reproduce any figure.]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, we have given the full simulation set up in Table A1.]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes, we have described our statistical measures and our error bars.]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes, we have cited our data and model sources.]

   (b) The license information of the assets, if applicable. [Yes]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Yes, this is included in the code supplementary materials.]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Strong Screening Rules for Group-based SLOPE Models: Supplementary Materials

A GROUP SLOPE

A.1 Penalty Weights

The penalty weights for gSLOPE were derived to provide group FDR-control under orthogonal designs (Brzyski et al., 2019). For the FDR-control parameter $q_g \in (0, 1)$, they are given by (where the indexing corresponds to the sorted groups)

$$w_i^{\max} = \max_{j=1,\dots,m} \left\{ \frac{1}{\sqrt{p_j}} F^{-1}_{\chi_{p_j}}(1 - q_g i/m) \right\}, \text{ for } i = 1, \dots, m,$$

where $F_{\chi_{p_j}}$ is the cumulative distribution function of a χ distribution with p_j degrees of freedom. A relaxtion to this sequence is applied in Brzyski et al. (2019) to give

$$w_i^{\text{mean}} = \overline{F}^{-1}_{\chi_{p_j}}(1 - q_g i/m), \text{ where } \overline{F}_{\chi_{p_j}}(x) := \frac{1}{m} \sum_{j=1}^{m} F_{\chi_{p_j}}(\sqrt{p_j} x). \tag{9}$$

The mean sequence weights defined in Equation 9 are used for all gSLOPE numerical simulations in this manuscript (shown in Figure A1).

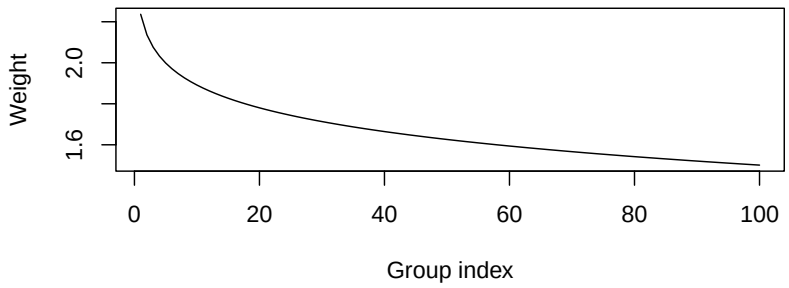


Figure A1: The gSLOPE weights, w, shown for Figure 4 for $p = 500, m = 100, q_g = 0.05$.

A.2 SLOPE Subdifferential

The subdifferential for SLOPE was derived in Larsson et al. (2020) and as it is a vital part of our arguments, it is reproduced here for ease of reference.

Define the function $R : \mathbb{R}^p \to \mathbb{N}^p$ that returns the ranks of the absolute values of its input and the set $\mathcal{C}_i(\beta) = \{j \in \{1, \dots, p\} \ : \ |\beta_i| = |\beta_j|\}$. Then, the subdifferential is given by (Larsson et al., 2020)

$$\partial J_{\text{slope}}(\beta; v) = \begin{cases} \left\{x \in \mathbb{R}^{|\mathcal{C}_i|} : \text{cumsum}(|x|_\downarrow - v_{R(x)_{\mathcal{C}_i}}) \preceq 0\right\} & \text{if } \beta_{\mathcal{C}_i} = 0, \\ \left\{x \in \mathbb{R}^{|\mathcal{C}_i|} : \text{cumsum}(|x|_\downarrow - v_{R(x)_{\mathcal{C}_i}}) \preceq 0 \right. \\ \quad \text{and} \ \ \sum_{j \in \mathcal{C}_i}(|x_j| - v_{R(s)_j}) = 0 \\ \left. \quad \text{and} \ \ \text{sign}(\beta_{\mathcal{C}_i}) = \text{sign}(x)\right\} & \text{otherwise.} \end{cases}$$

The primary use of the subdifferential in this manuscript is the zero condition (Equation 5).

A.3 Theory

Proof of Theorem 3.1. The proof is similar to that of Theorem 2.7 in Brzyski et al. (2019), where the subdifferential of gSLOPE is derived under equal groups. It is derived here under more general terms. The subdifferential needs to be derived under two cases:

1. Inactive groups, $\mathcal{G}_{\mathcal{Z}}$.
2. Active groups, $\mathcal{G}_{\mathcal{A}}$.

Case 1: For inactive groups, we consider the subdifferential at zero. The subdifferential of a norm at zero is given by the dual norm of the unit ball (Schneider and Tardivel, 2022),

$$\partial J_{\text{gslope}}(\mathbf{0}; w) = \mathbf{B}_{J^*_{\text{gslope}}(\mathbf{0}; w)}[0, 1] = \{x : J^*_{\text{gslope}}(x; w) \leq 1\}.$$

The dual norm for gSLOPE is given by (Brzyski et al., 2019)

$$J^*_{\text{gslope}}(x; w) = J^*_{\text{slope}}([x]_{\mathcal{G}, -0.5}).$$

Hence, the dual norm unit ball is

$$\mathbf{B}_{J^*_{\text{gslope}}(\mathbf{0}; w)}[0, 1] = \{x : [x]_{\mathcal{G}, -0.5} \in \mathbf{B}_{J^*_{\text{slope}}(\mathbf{0}; w)}[0, 1]\},$$

where $\mathbf{B}_{J^*_{\text{slope}}(\mathbf{0}; w)}[0, 1] = \{x \in \mathbb{R}^m : \text{cumsum}(|x|_\downarrow - w) \preceq \mathbf{0}\}$ is the unit ball of the dual norm to J_{slope} (Bogdan et al., 2015). Using this, the subdifferential at zero for the inactive groups, $\mathcal{Z}$, is given by

$$\partial J_{\text{gslope}}(\mathbf{0}; w_{\mathcal{Z}}) = \{x \in \mathbb{R}^{\text{card}(\mathcal{G}_{\mathcal{Z}})} : [x]_{\mathcal{G}_{\mathcal{Z}}, -0.5} \in \partial J_{\text{slope}}(\mathbf{0}; w_{\mathcal{Z}})\}.$$

Case 2: Without loss of generality, denote the group index s such that $\|\beta^{(g)}\|_2 = 0$ for $g > s$ (inactive groups) and $\|\beta^{(g)}\|_2 \neq 0$ for $g \leq s$ (active groups). In other words, $g \in \mathcal{G}_{\mathcal{A}}$ if $g \leq s$. Define a set $D = \{d \in \mathbb{R}^p : \|\beta^{(1)} + d^{(1)}\|_2 > \dots > \|\beta^{(s)} + d^{(s)}\|_2, \|\beta^{(s)} + d^{(s)}\|_2 > \|d^{(g)}\|_2, g > s\}$. By definition of a subdifferential, if $x \in \partial J_{\text{gslope}}(\beta; w)$, then for all $d \in D$

$$\sum_{g=1}^{m} \sqrt{p_g} w_g \|\beta^{(g)} + d^{(g)}\|_2 \geq \sum_{g=1}^{m} \sqrt{p_g} w_g \|\beta^{(g)}\|_2 + x^\top d.$$

Splitting this up into whether the groups are active (whether $g \leq s$):

$$\begin{aligned} \sum_{g=1}^{s} \sqrt{p_g} w_g \|\beta^{(g)} + d^{(g)}\|_2 + \sum_{g=s+1}^{m} \sqrt{p_g} w_g \|d^{(g)}\|_2 \geq & \sum_{g=1}^{s} \sqrt{p_g} w_g \|\beta^{(g)}\|_2 \\ & + \sum_{g=1}^{s} x^{(g)T} d^{(g)} + \sum_{g=s+1}^{m} x^{(g)T} d^{(g)}. \end{aligned} \tag{10}$$

Now, for $g \in \mathcal{G}_{\mathcal{A}}$, define a new set $D_g = \{d \in D : d^{(j)} \equiv \mathbf{0}, j \neq g\}$. Taking $d \in D_g$, Equation 10 becomes

$$\sqrt{p_g} w_g \|\beta^{(g)} + d^{(g)}\|_2 \geq \sqrt{p_g} w_g \|\beta^{(g)}\|_2 + x^{(g)T} d^{(g)}.$$

Since the set $\{d^{(g)} : d \in D_g\}$ is open in $\mathbb{R}^{p_g}$ and contains zero, by Corollary G.1 in Brzyski et al. (2019), it follows that $x^{(g)} \in \partial f_g(b^{(g)})$ for $f_g : \mathbb{R}^{p_g} \to \mathbb{R}, f_g(x) = w_g \sqrt{p_g} \|x\|_2$. Now, for $g \leq s$, f_g is differentiable in $\beta^{(g)}$, giving

$$x^{(g)} = w_g \sqrt{p_g} \frac{\beta^{(g)}}{\|\beta^{(g)}\|_2},$$

proving the result. □

Proof of Proposition 3.2. Suppose we have $\mathcal{B} \neq \varnothing$ after running the algorithm. Then, plugging in $h(\lambda_{k+1}) = ([\nabla f(\hat{\beta}(\lambda_{k+1}))]_{\mathcal{G},-0.5})_{\downarrow}$ gives

$$\operatorname{cumsum}\Big(\big(([\nabla f(\hat{\beta}(\lambda_{k+1}))]_{\mathcal{G},-0.5})_{\downarrow} \big)_{\mathcal{B}} - \lambda_{k+1} w_{\mathcal{B}} \Big) \prec \mathbf{0},$$

so that by the gSLOPE subdifferential (Theorem 3.1) all groups in $\mathcal{B}$ are inactive. This is valid by the KKT conditions (Equation 2), as we know that $-\nabla f(\hat{\beta}(\lambda_{k+1})) \in \partial J_{\text{gslope}}(\mathbf{0}; w)$. Hence, $\mathcal{S}_g(\lambda_{k+1})$ will contain the active set $\mathcal{A}_g(\lambda_{k+1})$. □

Proof of Proposition 3.3. Since $\operatorname{cumsum}(y) \succeq \operatorname{cumsum}(x) \iff y \succeq x$ (Larsson et al., 2020), we only need to show for a group g,

$$|h_g(\lambda_{k+1})| \leq |h_g(\lambda_k)| + \lambda_k w_g - \lambda_{k+1} w_g.$$

Applying the reverse triangle inequality to the Lipschitz assumption gives

$$\begin{aligned} &|h_g(\lambda_{k+1})| - |h_g(\lambda_k)| \leq |h_g(\lambda_{k+1}) - h_g(\lambda_k)| \leq \lambda_k w_g - \lambda_{k+1} w_g \\ \implies &|h_g(\lambda_{k+1})| \leq |h_g(\lambda_k)| + \lambda_k w_g - \lambda_{k+1} w_g, \end{aligned}$$

proving the result. □

A.4 KKT Checks

To check whether a group has been correctly discarded during the screening step, the KKT conditions for gSLOPE are checked. They are given by

$$\begin{aligned} &\mathbf{0} \in \nabla f(\beta) + \lambda \partial J_{\text{gslope}}(\beta; w) \\ \implies &-\nabla f(\beta) \in \lambda \partial J_{\text{gslope}}(\beta; w). \end{aligned}$$

Hence, we are checking whether the gradient of the loss function sits within the set of the gradient of the penalty. As we are only interested in identifying incorrectly discarded groups, we require only to check the subdifferential condition at zero. Hence, a violation occurs if a group is discarded but

$$\begin{aligned} &-\nabla f(\beta) \notin \lambda \partial J_{\text{gslope}}(\mathbf{0}; w_{\mathcal{G}_{\mathcal{Z}}}) \\ \implies &-\nabla f(\beta) \notin \left\{ x \in \mathbb{R}^{\operatorname{card} \mathcal{G}_{\mathcal{Z}}} : [x]_{\mathcal{G}_{\mathcal{Z}},-0.5} \in \partial J_{\text{slope}}(0; \lambda w_{\mathcal{G}_{\mathcal{Z}}}) \right\} \\ \implies &[\nabla f(\beta)]_{\mathcal{G}_{\mathcal{Z}},-0.5} \notin \partial J_{\text{slope}}(0; \lambda w_{\mathcal{G}_{\mathcal{Z}}}) \\ \implies &\operatorname{cumsum}(([\nabla f(\beta)]_{\mathcal{G}_{\mathcal{Z}},-0.5})_{\downarrow} - \lambda w_{\mathcal{G}_{\mathcal{Z}}}) \succ 0. \end{aligned}$$

A.5 Path Start Proof

Proof of Proposition 3.4. The aim is to find the value of λ at which the first group enters the model. When all features are zero, the gSLOPE KKT conditions (Equation 2) are

$$\mathbf{0} \in \nabla f(\mathbf{0}) + \lambda \partial J_{\text{gslope}}(\mathbf{0}; w).$$

This is satisfied when

$$[\nabla f(\mathbf{0})]_{\mathcal{G},-0.5} \in \partial J_{\text{slope}}(\mathbf{0}; \lambda w) \implies \text{cumsum}\big(([\nabla f(\mathbf{0})]_{\mathcal{G},-0.5})_{\downarrow} - \lambda w\big) \preceq \mathbf{0}.$$

Rearranging this gives

$$\lambda \succeq \text{cumsum}\big(([\nabla f(\mathbf{0})]_{\mathcal{G},-0.5})_{\downarrow}\big) \oslash \text{cumsum}(w).$$

Picking the maximum possible λ such that this holds yields

$$\lambda_1 = \max\left\{\text{cumsum}\big(([\nabla f(\mathbf{0})]_{\mathcal{G},-0.5})_{\downarrow}\big) \oslash \text{cumsum}(w)\right\}.$$

This can be verified by noting that $\lambda_1 = J^*_{\text{gslope}}(\nabla f(\mathbf{0}); w)$ (Ndiaye et al., 2016b). Now, $J^*_{\text{gslope}}(x; w) = J^*_{\text{slope}}([x]_{\mathcal{G},-0.5}; w)$ (Brzyski et al., 2019). The dual norm of SLOPE is given by (Negrinho and Martins, 2014)

$$J^*_{\text{slope}}(x; w) = \max\left\{\text{cumsum}(|x|_{\downarrow}) \oslash \text{cumsum}(w)\right\}.$$

Therefore, λ_1 is as before. □

B SPARSE-GROUP SLOPE

B.1 Penalty Weights

The penalty weights for SGS provide variable and group FDR-control simultaneously, under orthogonal designs (Feser and Evangelou, 2023). They are given by (where the indexing corresponds to the sorted variables/groups)

$$v_i^{\max} = \max_{j=1,\dots,m}\left\{\frac{1}{\alpha}F_{\mathcal{N}}^{-1}\left(1 - \frac{q_v i}{2p}\right) - \frac{1}{3\alpha}(1-\alpha)a_j w_j\right\},\ i = 1,\dots,p,$$
$$w_i^{\max} = \max_{j=1,\dots,m}\left\{\frac{F_{\text{FN}}^{-1}(1 - \frac{q_g i}{m}) - \alpha\sum_{k\in\mathcal{G}_j} v_k}{(1-\alpha)p_j}\right\},\ i = 1,\dots,m,$$

where $F_{\chi_{p_j}}$ is the cumulative distribution function of a χ distribution with p_j degrees of freedom, $F_{\mathcal{N}}$ is the cumulative distribution function of a folded Gaussian distribution, and a_j is a quantity that requires estimation. The estimator $\hat{a}_j = \lfloor \alpha p_j \rfloor$ is proposed in Feser and Evangelou (2023). As with gSLOPE (Appendix A.1), a relaxtion is possible, giving the weights

$$v_i^{\text{mean}} = \overline{F}_{\mathcal{N}}^{-1}\left(1 - \frac{q_v i}{2p}\right),\ \text{where } \overline{F}_{\mathcal{N}}(x) := \frac{1}{m}\sum_{j=1}^{m} F_{\mathcal{N}}\left(\alpha x + \frac{1}{3}(1-\alpha)a_j w_j\right), \tag{11}$$

$$w_i^{\text{mean}} = \overline{F}_{\text{FN}}^{-1}\left(1 - \frac{q_g i}{p}\right),\ \text{where } \overline{F}_{\text{FN}}(x) := \frac{1}{m}\sum_{j=1}^{m} F_{\text{FN}}\left((1-\alpha)p_j x + \alpha\sum_{k\in\mathcal{G}_j} v_k\right). \tag{12}$$

In the manuscript, as recommended by Feser and Evangelou (2023) under general settings, the SGS variable mean (Equation 11) and gSLOPE group mean (Equation 9) weights are used for all SGS numerical simulations.

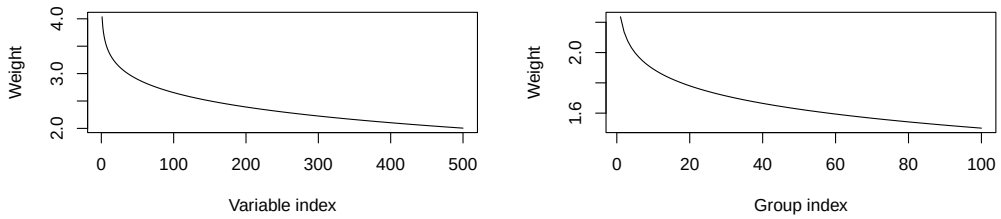


Figure A2: The SGS weights, (v, w), shown for Figure 4 for $p = 500, m = 100, q_v = 0.05, q_g = 0.05, \alpha = 0.95$.

B.2 Derivation of Soft Thresholding Operator

Proof of Lemma 4.1. To determine the form of the quantity $\partial J_{\text{slope}}(\mathbf{0}; v)$, consider that for Equation 7 to be satisfied, the term inside the $[\cdot]$ operator needs to be as small as possible. Now,

$$\partial J_{\text{slope}}(\mathbf{0}; v) = \{y : \text{cumsum}(|y|) \preceq \text{cumsum}(v)\}.$$

Note that $\text{cumsum}(y) \preceq \text{cumsum}(x) \iff y \preceq x$. We consider the cases:

1. $\nabla_i f(\beta) > \lambda\alpha v_i$: choose $y_i = -v_i$.
2. $\nabla_i f(\beta) < -\lambda\alpha v_i$: choose $y_i = v_i$.
3. $\nabla_i f(\beta) \in [-\lambda\alpha v_i, \lambda\alpha v_i]$: choose $y_i = \nabla_i f(\beta)/\lambda\alpha v_i$.

Hence, the term becomes

$$S(\nabla f(\beta), \lambda\alpha v) := \text{sign}(\nabla f(\beta))(|\nabla f(\beta)| - \lambda\alpha v)_+,$$

which is the soft thresholding operator. □

B.3 Theory

Proposition B.1 (Strong group screening rule for SGS)**.** *Let $\tilde{h}(\lambda) := ([S(\nabla f(\beta), \lambda\alpha v)]_{\mathcal{G},-0.5})_\downarrow$. Then taking $c = \tilde{h}(\lambda_{k+1})$ and $\phi = (1-\alpha)\lambda_{k+1} w$ as inputs for Algorithm A1 returns a superset $\mathcal{S}_g(\lambda_{k+1})$ of the active set $\mathcal{A}_g(\lambda_{k+1})$.*

Proof of Proposition B.1. The proof is similar to that of Proposition 3.2. Suppose we have $\mathcal{B} \neq \varnothing$ after running the algorithm. Then,

$$\begin{aligned} &\text{cumsum}(\tilde{h}_{\mathcal{B}}(\lambda_{k+1}) - \lambda_{k+1}(1-\alpha)w_{\mathcal{B}}) \prec \mathbf{0} \\ \implies &\text{cumsum}\Big(\big([S(\nabla f(\beta), \lambda_{k+1}\alpha v)]_{\mathcal{G},-0.5})_\downarrow\big)_{\mathcal{B}} - \lambda_{k+1}(1-\alpha)w_{\mathcal{B}}\Big) \prec \mathbf{0}, \end{aligned}$$

so that by the SGS subdifferential (Equation 7) all groups in $\mathcal{B}$ are inactive. Hence, $\mathcal{S}_g(\lambda_{k+1})$ will contain the active set $\mathcal{A}_g(\lambda_{k+1})$. □

Proof of Proposition 4.2. The proof is identical to that of Proposition 3.3, replacing $h_g(\cdot)$ with $\tilde{h}_g(\cdot)$ and $\lambda_{k+1}w$ by $\lambda_{k+1}(1-\alpha)w$. □

Proposition B.2 (Strong variable screening rule for SGS)**.** *Let $\bar{h}(\lambda) = |(\nabla f(\hat{\beta}(\lambda)))|_\downarrow$. Then taking $c = \bar{h}(\lambda_{k+1})$ and $\phi = \lambda_{k+1}\alpha v$ for only the variables contained in the groups in $\mathcal{A}_g(\lambda_{k+1})$ in Algorithm A1 returns a superset $\mathcal{S}_v(\lambda_{k+1})$ of the active set $\mathcal{A}_v(\lambda_{k+1})$.*

Proof. Suppose we have $\mathcal{B} \neq \varnothing$ after running the algorithm. Then, we have

$$\text{cumsum}(\bar{h}_{\mathcal{B}}(\lambda_{k+1}) - \lambda_{k+1}\alpha v_{\mathcal{B}}) \prec \mathbf{0} \implies \text{cumsum}\Big(\big(|\nabla f(\hat{\beta}(\lambda_{k+1}))|_\downarrow\big)_{\mathcal{B}} - \lambda_{k+1}\alpha v_{\mathcal{B}}\Big) \prec \mathbf{0},$$

so that by the SGS subdifferential for non-zero groups (Equation 8) all variables in $\mathcal{B}$ are inactive. Hence, $\mathcal{S}_v(\lambda_{k+1})$ will contain the active set $\mathcal{A}_v(\lambda_{k+1})$. □

Proof for Proposition 4.3. The proof is identical to that of Proposition 3.3, replacing $h_g(\cdot)$ with $\bar{h}_g(\cdot)$, $\lambda_{k+1}v$ with $\lambda_{k+1}\alpha v$, and considering only variables in the groups contained in $\mathcal{A}_g(\lambda_{k+1})$. □

B.4 KKT Checks

For SGS, the KKT conditions are first checked at the group-level for inactive groups (Appendix B.4.1). Further variable checks are performed for violating groups and variables in active groups (to check whether the variables should also be active) (Appendix B.4.2). The violating variables from these secondary variable checks are added back into $\mathcal{E}_v$.

B.4.1 Group Checks

A group violation occurs if the KKT conditions do not hold at the group-level (Equation 7). That is, a violation occurs if a group is discarded but

$$\text{cumsum}\Big(\left([\nabla f(\beta) + \lambda\alpha\partial J_{\text{slope}}(\mathbf{0}; v)]_{\mathcal{G}_\mathcal{Z}, -0.5}\right)_\downarrow - \lambda(1-\alpha)w_\mathcal{Z}\Big) \succ \mathbf{0}.$$

B.4.2 Variable Checks

For the set of variables in a violating group (from Appendix B.4.1), denoted $\mathcal{G}_{\mathcal{K}_g}$, a variable violation occurs if Equation 8 does not hold. That is, if

$$\nabla_{\mathcal{G}_{\mathcal{K}_g}} f(\beta) \notin \lambda\alpha\partial J_{\text{slope}}(\mathbf{0}; v_{\mathcal{G}_{\mathcal{K}_g}}) \implies \text{cumsum}(|\nabla_{\mathcal{G}_{\mathcal{K}_g}} f(\beta)| - \lambda\alpha v_{\mathcal{G}_{\mathcal{K}_g}}) \succ 0.$$

B.4.3 Alternative KKT Checks

An alternative approach for SGS is to check the KKT conditions directly on the variables. The KKT conditions (Equation 6) can be rewritten as

$$-\nabla f(\beta) - \lambda(1-\alpha)\partial J_{\text{gslope}}(\beta; w) \in \lambda\alpha\partial J_{\text{slope}}(\beta; v).$$

A KKT violation occurs the zero subdifferential conditions are not satisfied

$$\begin{aligned} &-\nabla f(\beta) - \lambda(1-\alpha)\partial J_{\text{gslope}}(\beta; w) \notin \lambda\alpha\partial J_{\text{slope}}(\mathbf{0}; v) \\ \implies &\text{cumsum}\left(|\nabla f(\beta) + \lambda(1-\alpha)\partial J_{\text{gslope}}(\beta; w)|_\downarrow - \lambda\alpha v\right) \succ \mathbf{0}. \end{aligned}$$

Now, the objective is to make the term inside the sorted absolute value operator as small as possible, given that the subdifferential term is unknown. To do this, a similar derivation as in Section B.2 can be used to determine that the term must be the soft thresholding operator, so that a violation occurs if

$$\text{cumsum}\left(|S(\nabla f(\beta), \lambda(1-\alpha)\tau\omega)|_\downarrow - \lambda\alpha v\right) \succ \mathbf{0},$$

where τ and ω are expanded vectors of the group sizes ($\sqrt{p_g}$) and penalty weights (w_g) to p dimensions, so that each variable within the same group is assigned the same value. However, as we have had to approximate the unknown subdifferential term, this check is not exact. In practice, we found that this check was not stringent enough (due to the approximation), leading to violations being missed.

B.5 Path Start Proof

Proof of Proposition 4.4. The aim is to find the value of λ at which the first variable enters the model. When all features are zero, the SGS KKT conditions (Equation 6) are

$$\begin{aligned} &-\nabla f(\mathbf{0}) \in \lambda(1-\alpha)\partial J_{\text{gslope}}(\mathbf{0}; w) + \lambda\alpha\partial J_{\text{slope}}(\mathbf{0}; v) \\ \implies &-\frac{1}{\lambda}\nabla f(\mathbf{0}) - (1-\alpha)\partial J_{\text{gslope}}(\mathbf{0}; w) \in \alpha\partial J_{\text{slope}}(\mathbf{0}; v) \\ \implies &\text{cumsum}\left(\left|-\frac{1}{\lambda}\nabla f(\mathbf{0}) - (1-\alpha)\partial J_{\text{gslope}}(\mathbf{0}; w)\right|_\downarrow - \alpha v\right) \preceq \mathbf{0}. \end{aligned}$$

By the reverse triangle inequality and ordering of the group weights

$$\begin{aligned} &\frac{1}{\lambda}\text{cumsum}\left(|\nabla f(\mathbf{0})|_\downarrow\right) \preceq \text{cumsum}((1-\alpha)|\partial J_{\text{gslope}}(\mathbf{0}; w)| - \alpha v) \\ \implies &\lambda \succeq \text{cumsum}(|\nabla f(\mathbf{0})|_\downarrow) \oslash \text{cumsum}((1-\alpha)|\partial J_{\text{gslope}}(\mathbf{0}; w)| - \alpha v). \end{aligned}$$

Now, note that for $x \in J_{\text{gslope}}(\mathbf{0}; w)$, it holds

$$\text{cumsum}([x]_{\mathcal{G}, -0.5} - w) \preceq \mathbf{0} \implies [x]_{\mathcal{G}, -0.5} \preceq w \implies \|x^{(g)}\|_2 \leq \sqrt{p_g} w_g, \forall g \in \mathcal{G}.$$

This is satisfied at the upper limit at $x = \tau\omega$. Hence,

$$\lambda_1 = \max\left\{\text{cumsum}(|\nabla f(\mathbf{0})|_\downarrow) \oslash \text{cumsum}((1-\alpha)\tau\omega - \alpha v)\right\}.$$

□

C SLOPE ALGORITHM

Algorithm A1 SLOPE subdifferential algorithm from Larsson et al. (2020)

Input: $c \in \mathbb{R}^p, \phi \in \mathbb{R}^p$, where $\phi_1 \geq \cdots \geq \phi_p \geq 0$
$\mathcal{S}, \mathcal{B} \leftarrow \varnothing$
for $i = 1$ **to** p **do**
  $\mathcal{B} \leftarrow \mathcal{B} \cup \{i\}$
  if $\text{cumsum}(c_\mathcal{B} - \phi_\mathcal{B}) \geq 0$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{B}$
    $\mathcal{B} \leftarrow \varnothing$
  end if
end for
Output: $\mathcal{S}$

D SCREEENING RULE FRAMEWORK

D.1 Group SLOPE Algorithm

For the following is performed for $k = 1, \dots, l-1$:

1. Set $\mathcal{E}_g = \mathcal{S}_g(\lambda_{k+1}) \cup \mathcal{A}_g(\lambda_k)$, where $\mathcal{S}_g(\lambda_{k+1})$ is obtained using Proposition 3.3.
2. Compute $\hat{\beta}(\lambda_{k+1})$ by Equation 1 with the gSLOPE norm using only the groups in $\mathcal{E}_g$. For any groups not in $\mathcal{E}_g$, $\hat{\beta}(\lambda_{k+1})$ is set to zero.
3. Check the KKT conditions (Equation 2) for all groups at this solution.
4. If there are no violations, we are done and keep $\hat{\beta}(\lambda_{k+1})$. Otherwise, add the violating groups into $\mathcal{E}$ and return to Step 2.

D.2 SGS Algorithm

For the following is performed for $k = 1, \dots, l-1$:

1. *Group screen step*: Calculate $\mathcal{S}_g(\lambda_{k+1})$ using Proposition 4.2.
2. *Variable screen step*: Set $\mathcal{E}_v = \mathcal{S}_v(\lambda_{k+1}) \cup \mathcal{A}_v(\lambda_k)$, where $\mathcal{S}_v(\lambda_{k+1})$ is obtained using Proposition 4.3 with only the groups in $\mathcal{S}_g(\lambda_{k+1})$.
3. Compute $\hat{\beta}(\lambda_{k+1})$ by Equation 1 with the SGS norm using only the features in $\mathcal{E}_v$. For features not in $\mathcal{E}_v$, $\hat{\beta}(\lambda_{k+1})$ is set to zero.
4. Check the KKT conditions (Equation 6) for all features at this solution.
5. If there are no violations, we are done and keep $\hat{\beta}(\lambda_{k+1})$, otherwise add in the violating variables into $\mathcal{E}_g$ and return to Step 3.

E GROUP-BASED OSCAR

This section provides supplementary materials for extending the proposed screening rules to group-based OSCAR models.

E.1 Penalty Sequence

The gOSCAR and SGO weights are defined by (for a variable $i \in [p]$ and group $g \in [m]$) (Figure A3)

$$v_i = \sigma_1 + \sigma_2(p - i), \; w_g = \sigma_1 + \sigma_3(m - g), \; \sigma_3 = \sigma_1/m. \tag{13}$$

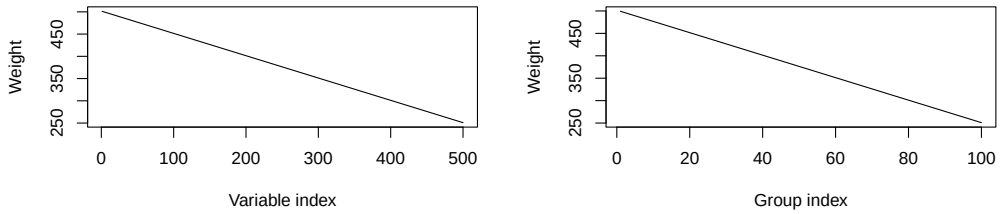


Figure A3: The SGO weights, (v, w), for $p = 500, m = 100, q_v = 0.05, q_g = 0.05, \alpha = 0.95$.

E.2 Results

Observations and conclusions made for the screening rules of gSLOPE and SGS are also found to be true for gOSCAR and SGO (Figures A4 - A8).

Figure A4 illustrates the effectiveness of bi-selection of SGO, similar to the effectiveness observed for SGS. Figures A5 and A6 showcase the efficiency of the screening rules on the proportion of the selected groups/variables. The screening rules are found to be effective across different data characteristics, with the running time of the models significantly decreasing (Figure A7). KKT violations for SGO are more common compared to gOSCAR (Figure A8), due to the additional assumptions made at the second screening layer (as with SGS). Similar to Figure 7, the shape of the increasing number of KKT violations mirrors the log-linear shape of the regularization path.

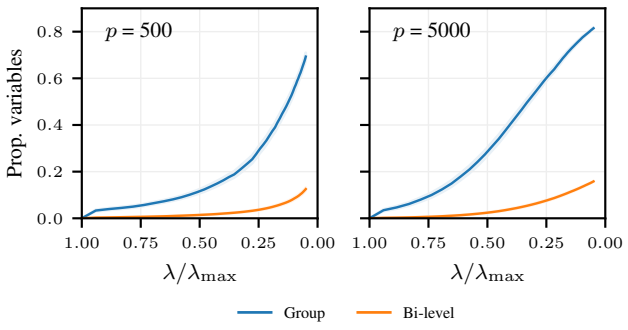


Figure A4: The proportion of variables in $\mathcal{S}_v$ relative to the full input for SGO, shown for group and bi-level screening plotted as function of the regularization path, applied to the synthetic data (Section 6.1). The data are generated under a linear model for $p = 500, 5000$. The results are averaged over 100 repetitions and 95% confidence intervals are shown (the SGO equivalent of Figure 1).

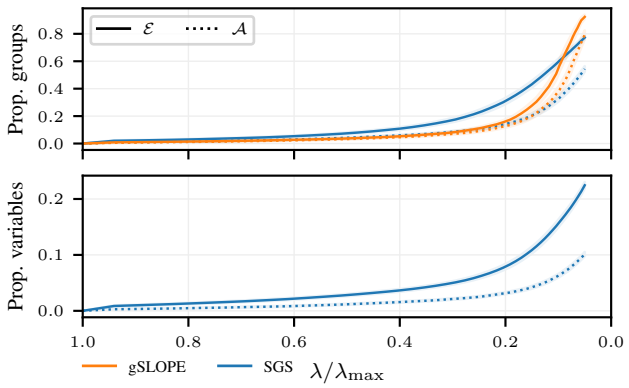


Figure A5: The number of groups/variables in $\mathcal{E}, \mathcal{A}$ for both gOSCAR and SGO as a function of the regularization path for the linear model with $p = 2750, \rho = 0.6, m = 197$. The results are averaged over 100 repetitions, with the shaded regions corresponding to 95% confidence intervals (the gOSCAR/SGO equivalent of Figure 2).

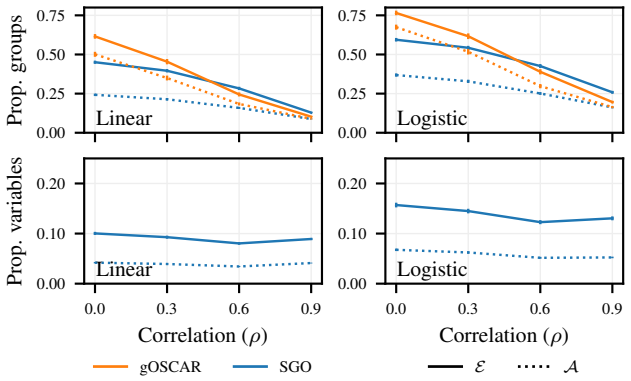


Figure A6: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$, relative to the full input, shown for gOSCAR and SGO. This is shown as a function of the correlation (ρ), averaged over all cases of the input dimension (p), with 100 repetitions for each p, for both linear and logistic models, with standard errors shown (the gOSCAR/SGO equivalent of Figure 3).

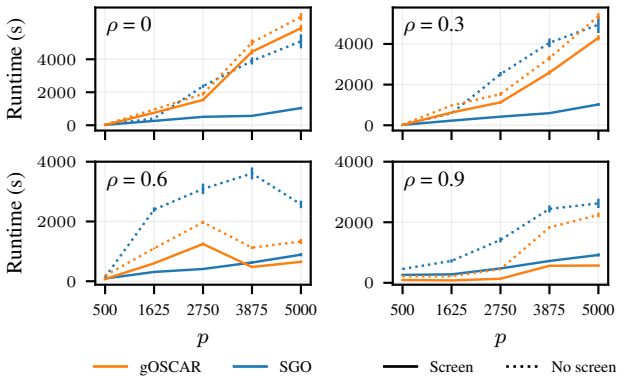


Figure A7: Runtime (in seconds) for fitting 50 models along a path, shown for screening against no screening as a function of p, broken down into different correlation cases, for the linear model. The results are averaged over 100 repetitions, with standard errors shown (the gOSCAR/SGO equivalent of Figure 4).

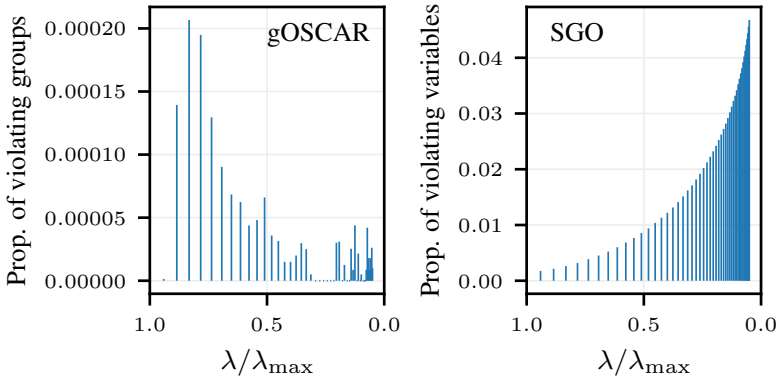


Figure A8: The proportion of KKT violations relative to the full input space, as a function of the regularization path. Group violations for gOSCAR and variable violations for SGS, under linear models, averaged over all cases of p and ρ (the gOSCAR/SGO equivalent of Figure 7).

F RESULTS

F.1 Computational Information

The simulated experiments were executed on a high-performance computing cluster (x86-64 Linux GNU) and the real data analysis was conducted on a Apple Macbook Air (M1, 8GB). Code for all simulations is available in the Supplementary Material. For all models, ATOS was used with the algorithmic parameters given in Table A1.

Table A1: Hyperparameters used for running ATOS in the synthetic and real data studies.

Parameter	Synthetic Data	Real Data
Max iterations	5000	10000
Backtracking	0.7	0.7
Max backtracking iterations	100	100
Convergence tolerance	10^{-5}	10^{-5}
Convergence criteria	$\|x - z\|_2$	$\|x - z\|_2$
Standardization	ℓ_2	ℓ_2
Intercept	Yes for linear	Yes for linear
Warm starts	Yes	Yes

F.2 Solution Optimality

This section presents the accuracy of the models with and without screening, by comparing the ℓ_2 distances observed between the screened and non-screened fitted values.

Synthetic Data For the linear model, the maximum ℓ_2 distances observed between the screened and non-screened fitted values were of order 10^{-6} for gSLOPE and 10^{-9} for SGS (Table A3). Across the different cases, 98000 models were fit in total for each approach (excluding the models for λ_1, where no screening is applied). Of these model fits, there were no instances for gSLOPE where $\mathcal{E}$ was not a superset of $\mathcal{A}$. There was only one instance (out of the 98000) that this occurred for SGS, where $\mathcal{E}$ was missing a single variable contained in $\mathcal{A}$ (which had a non-screen fitted value of $\hat{\beta} = -0.004$).

For the logistic model, the maximum ℓ_2 distances observed between the screened and non-screened fitted values were of order 10^{-8} for gSLOPE and 10^{-9} for SGS (Table A7). Across the different cases, 98000 models were fit in total for each approach (excluding the models for λ_1, where no screening is applied). Of these model fits, there were no instances for gSLOPE or SGS where $\mathcal{E}$ was not a superset of $\mathcal{A}$.

Real Data In the real data analysis, the estimated coefficients with and without screening were very close to each other for both SGS and gSLOPE (Table A10). However, direct comparison is less meaningful here, as the models often failed to converge without screening, therefore not reaching the optimal solution.

F.3 Additional Results from the Simulation Study

Table A2: Variable screening metrics for SGS using linear and logistic models for the simulation study presented in Section 6.1. The number of variables in $\mathcal{A}_v, \mathcal{S}_v, \mathcal{E}_v$, and $\mathcal{K}_v$ are shown, averaged across all 20 cases of the correlation (ρ) and p. Standard errors are shown.

Method	Type	card($\mathcal{A}_v$)	card($\mathcal{S}_v$)	card($\mathcal{E}_v$)	card($\mathcal{K}_v$)
SGS	Linear	179 ± 3	313 ± 5	363 ± 6	51 ± 1
SGS	Logistic	230 ± 3	405 ± 5	472 ± 6	66 ± 1

Table A3: General and group screening metrics for SGS and gSLOPE using linear and logistic models for the simulation study presented in Section 6.1. General metrics: the runtime (in seconds) for screening against no screening, the number of fitting iterations for screening against no screening, and the ℓ_2 distance between the fitted values obtained with screening and no screening. Group screening metrics: the number of groups in $\mathcal{A}_g, \mathcal{S}_g, \mathcal{E}_g$, and $\mathcal{K}_g$. The results are averaged across all 20 cases of the correlation (ρ) and p. Standard errors are shown.

Method	Type	Runtime screen (s)	Runtime no screen (s)	card($\mathcal{A}_g$)	card($\mathcal{S}_g$)	card($\mathcal{E}_g$)	card($\mathcal{K}_g$)	Num it screen	Num it no screen	ℓ_2 dist to no screen
gSLOPE	Linear	1016 ± 21	1623 ± 27	55 ± 1	76 ± 1	76 ± 1	0.006 ± 0.004	333 ± 6	351 ± 6	$2 \times 10^{-6} \pm 1 \times 10^{-6}$
gSLOPE	Logistic	814 ± 8	1409 ± 11	71 ± 1	97 ± 1	97 ± 1	0.014 ± 0.014	78 ± 1	83 ± 1	$1 \times 10^{-8} \pm 1 \times 10^{-8}$
SGS	Linear	735 ± 15	1830 ± 34	61 ± 1	84 ± 1	91 ± 1	-	91 ± 3	708 ± 12	$2 \times 10^{-9} \pm 3 \times 10^{-9}$
SGS	Logistic	407 ± 2	859 ± 6	84 ± 1	107 ± 1	118 ± 1	-	7 ± 0.2	51 ± 0.8	$4 \times 10^{-9} \pm 3 \times 10^{-10}$

F.3.1 Additional Results for the Linear Model

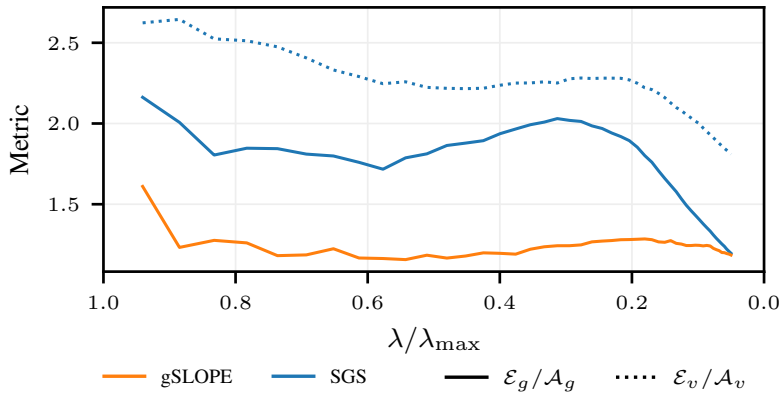
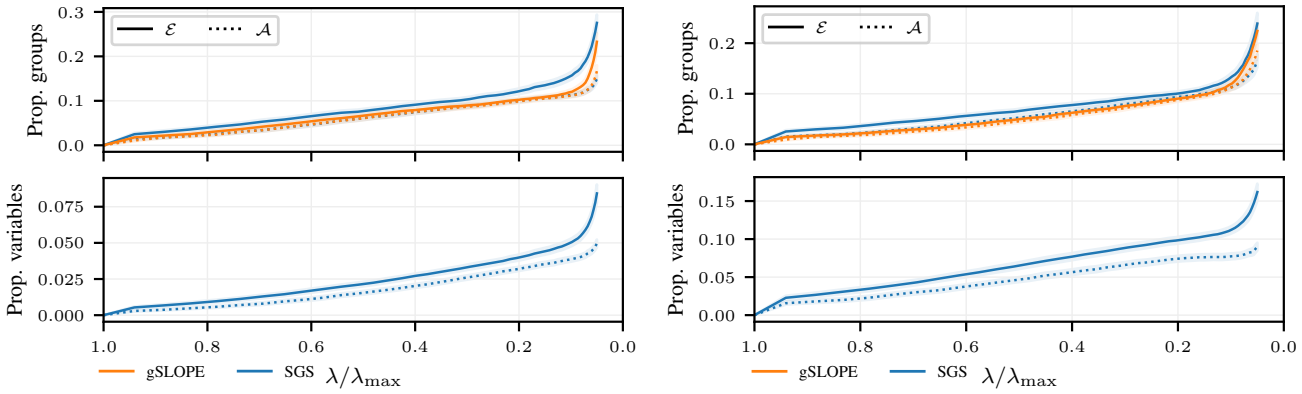
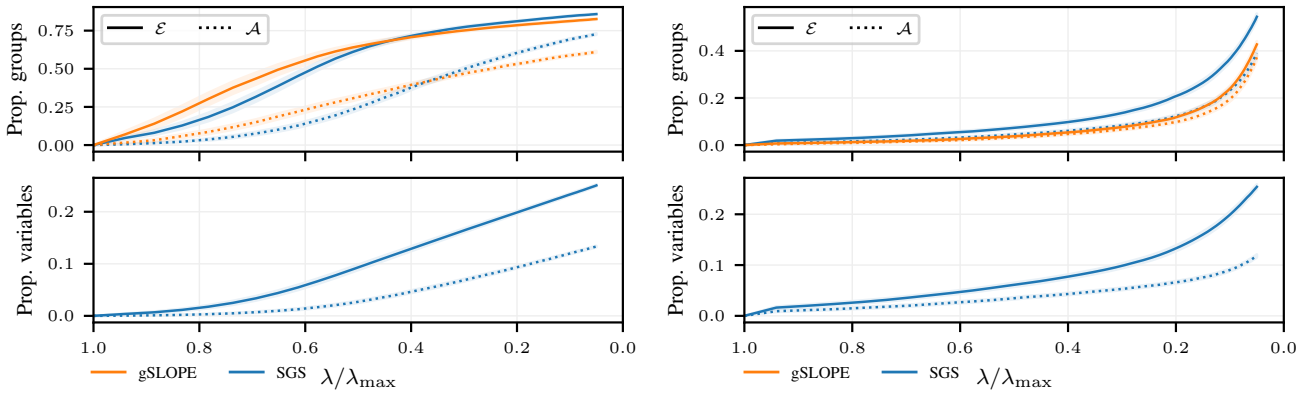


Figure A9: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$, relative to the full input, for gSLOPE and SGS, as a function of the regularization path for the linear model with $p = 2750, \rho = 0.6, m = 197$. The results are averaged over 100 repetitions.



(a) $p = 500, \rho = 0$

(b) $p = 500, \rho = 0.9$



(c) $p = 5000, \rho = 0$

(d) $p = 5000, \rho = 0.9$

Figure A10: The number of groups/variables in $\mathcal{E}, \mathcal{A}$ as a function of the regularization path for the linear model with SGS and gSLOPE, shown for different values of the correlation (ρ) and p. The results are averaged over 100 repetitions, with 95% confidence intervals shown.

Table A4: Variable screening metrics for SGS using a linear model for the simulation study presented in Section 6.1. The number of variables in $\mathcal{A}_v, \mathcal{S}_v, \mathcal{E}_v$, and $\mathcal{K}_v$ are shown. The results are shown for different values of p, averaged across $\rho \in \{0, 0.3, 0.6, 0.9\}$. Standard errors are shown.

Method	p	card($\mathcal{A}_v$)	card($\mathcal{S}_v$)	card($\mathcal{E}_v$)	card($\mathcal{K}_v$)
SGS	500	19 ± 1	24 ± 1	28 ± 1	4 ± 0.2
SGS	1625	83 ± 3	138 ± 5	161 ± 6	22 ± 1
SGS	2750	188 ± 7	316 ± 10	370 ± 12	54 ± 2
SGS	3875	268 ± 9	470 ± 14	548 ± 16	78 ± 3
SGS	5000	334 ± 10	618 ± 17	712 ± 20	95 ± 3

Table A5: General and group screening metrics for SGS and gSLOPE using linear models for the simulation study presented in Section 6.1. General metrics: the runtime (in seconds) for screening against no screening, the number of fitting iterations for screening against no screening, and the ℓ_2 distance between the fitted values obtained with screening and no screening. Group screening metrics: the number of groups in $\mathcal{A}_g, \mathcal{S}_g, \mathcal{E}_g$, and $\mathcal{K}_g$. The results are shown for different values of p, averaged across $\rho \in \{0, 0.3, 0.6, 0.9\}$. Standard errors are shown.

Method	p	Runtime screen (s)	Runtime no screen (s)	card($\mathcal{A}_g$)	card($\mathcal{S}_g$)	card($\mathcal{E}_g$)	card($\mathcal{K}_g$)	Num it screen	Num it no screen	ℓ_2 dist to no screen
gSLOPE	500	89 ± 1	144 ± 1	9 ± 0.2	10 ± 0.3	10 ± 0.3	0.005 ± 0.005	47 ± 1	55 ± 1	$6 \times 10^{-6} \pm 7 \times 10^{-6}$
gSLOPE	1625	231 ± 5	453 ± 5	26 ± 1	36 ± 1	36 ± 1	0.005 ± 0.006	203 ± 6	222 ± 6	$1 \times 10^{-6} \pm 1 \times 10^{-6}$
gSLOPE	2750	1061 ± 15	2296 ± 25	56 ± 2	75 ± 2	75 ± 2	0.004 ± 0.005	270 ± 9	350 ± 9	$5 \times 10^{-7} \pm 7 \times 10^{-7}$
gSLOPE	3875	1765 ± 83	2800 ± 113	82 ± 3	114 ± 3	114 ± 3	0.006 ± 0.008	549 ± 19	546 ± 18	$3 \times 10^{-7} \pm 5 \times 10^{-7}$
gSLOPE	5000	1937 ± 63	2422 ± 66	102 ± 3	147 ± 4	147 ± 4	0.007 ± 0.012	594 ± 21	581 ± 19	$2 \times 10^{-7} \pm 3 \times 10^{-7}$
SGS	500	94 ± 1	133 ± 2	9 ± 0.2	11 ± 0.3	12 ± 0.3	-	28 ± 1	74 ± 3	$5 \times 10^{-10} \pm 4 \times 10^{-10}$
SGS	1625	416 ± 8	1129 ± 19	25 ± 1	37 ± 1	41 ± 1	-	62 ± 7	511 ± 21	$2 \times 10^{-9} \pm 2 \times 10^{-9}$
SGS	2750	639 ± 14	2137 ± 47	62 ± 2	82 ± 2	89 ± 3	-	80 ± 6	791 ± 28	$2 \times 10^{-9} \pm 6 \times 10^{-9}$
SGS	3875	939 ± 31	2862 ± 96	93 ± 3	124 ± 3	136 ± 4	-	112 ± 8	1049 ± 34	$5 \times 10^{-9} \pm 1 \times 10^{-8}$
SGS	5000	1586 ± 66	2891 ± 128	119 ± 4	164 ± 3	180 ± 4	-	171 ± 11	1118 ± 37	$2 \times 10^{-9} \pm 4 \times 10^{-9}$

F.3.2 Additional Results for the Logistic Model

This section presents additional results for the logistic model. Similar trends to the ones observed for the linear model are seen.

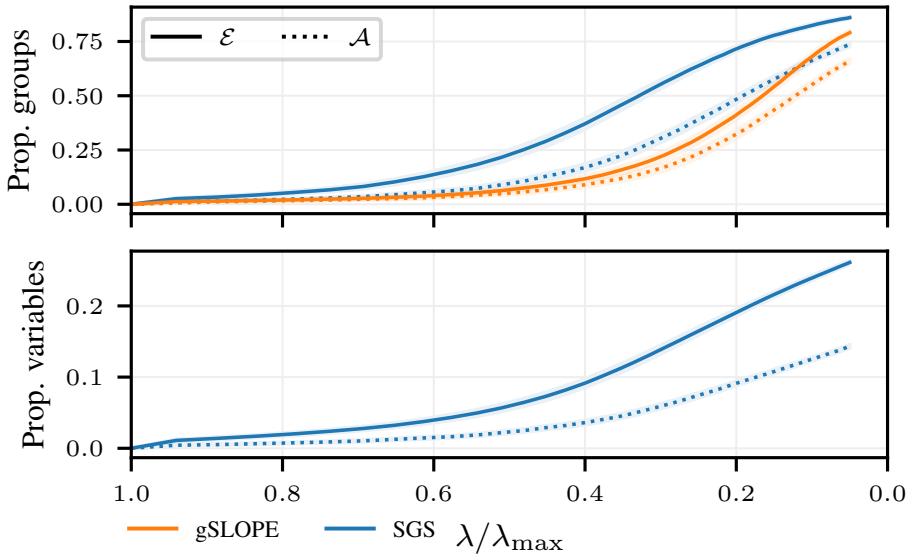
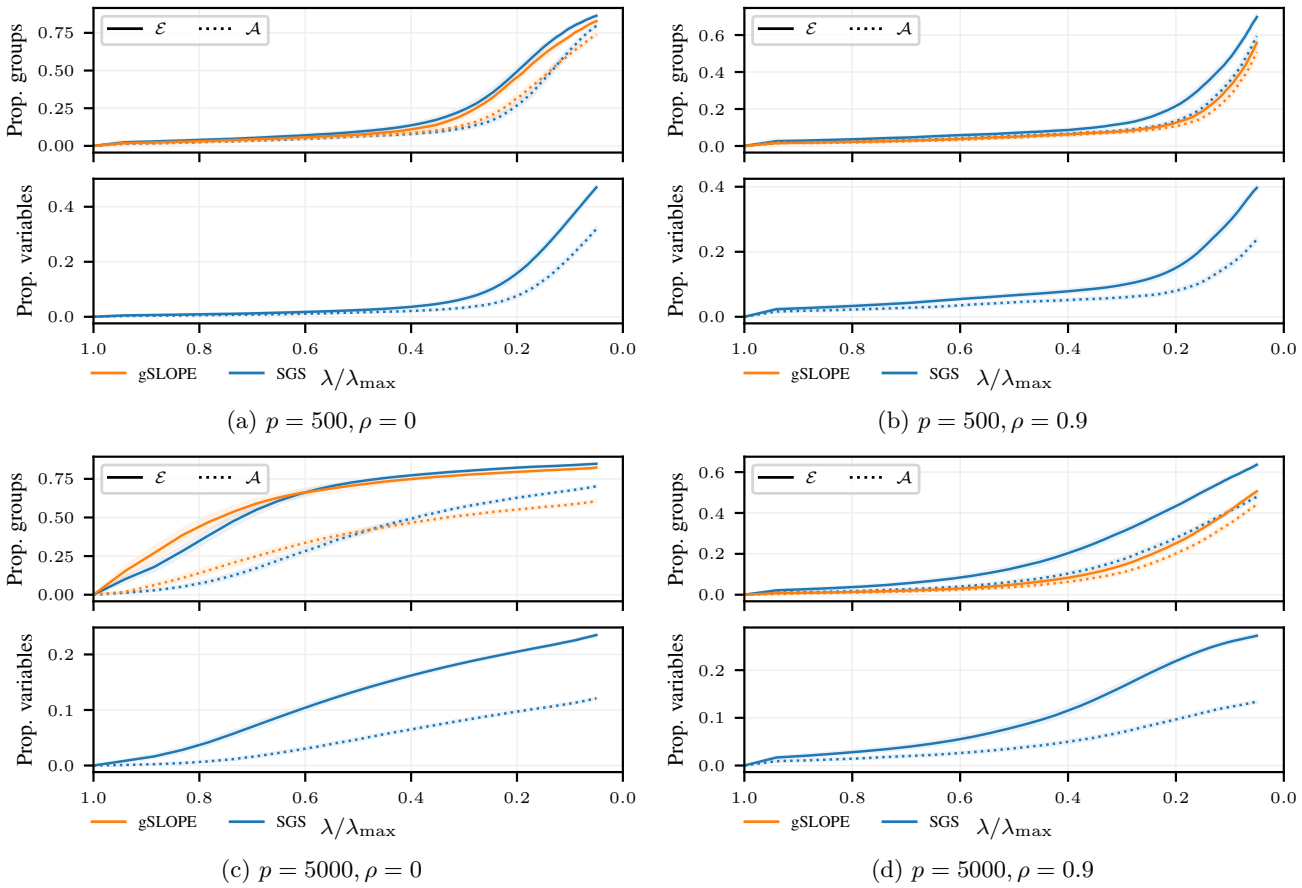


Figure A11: The number of groups/variables in $\mathcal{E}, \mathcal{A}$ as a function of the regularization path for the logistic model with $p = 2750, \rho = 0.6, m = 197$, shown for gSLOPE and SGS. The results are averaged over 100 repetitions, with 95% confidence intervals shown. This figure is the equivalent of Figure 2 for the logistic model.



(a) $p = 500, \rho = 0$

(b) $p = 500, \rho = 0.9$

(c) $p = 5000, \rho = 0$

(d) $p = 5000, \rho = 0.9$

Figure A12: The number of groups/variables in $\mathcal{E}, \mathcal{A}$ as a function of the regularization path for the logistic model with SGS and gSLOPE, shown for different values of the correlation (ρ) and p. The results are averaged over 100 repetitions, with 95% confidence intervals shown. This figure is the equivalent of Figure A10 for the logistic model.

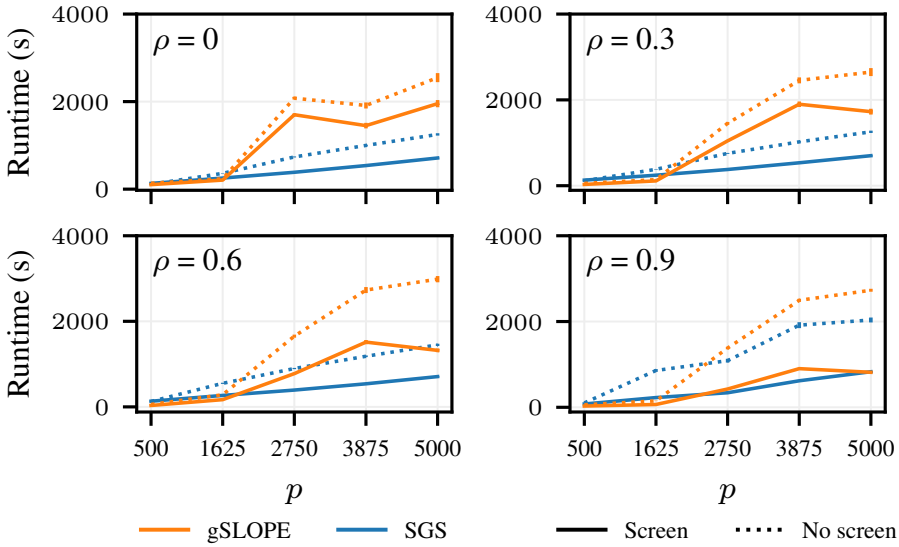


Figure A13: Runtime (in seconds) for screening against no screening as a function of p, broken down into different correlation cases, for the logistic model. The results are averaged over 100 repetitions, with standard errors shown. This figure is the equivalent of Figure 4 for the logistic model.

Table A6: Variable screening metrics for SGS using a logistic model for the simulation study presented in Section 6.1. The number of variables in $\mathcal{A}_v, \mathcal{S}_v, \mathcal{E}_v$, and $\mathcal{K}_v$ are shown. The results are shown for different values of p, averaged across $\rho \in \{0, 0.3, 0.6, 0.9\}$. Standard errors are shown.

Method	p	card($\mathcal{A}_v$)	card($\mathcal{S}_v$)	card($\mathcal{E}_v$)	card($\mathcal{K}_v$)
SGS	500	53 ± 2	71 ± 3	89 ± 4	19 ± 1
SGS	1625	157 ± 5	248 ± 8	291 ± 9	44 ± 1
SGS	2750	247 ± 7	420 ± 11	491 ± 13	71 ± 2
SGS	3875	316 ± 9	571 ± 14	663 ± 16	92 ± 3
SGS	5000	375 ± 10	717 ± 16	824 ± 19	107 ± 3

Table A7: General and group screening metrics for SGS and gSLOPE using logistic models for the simulation study presented in Section 6.1. General metrics: the runtime (in seconds) for screening against no screening, the number of fitting iterations for screening against no screening, and the ℓ_2 distance between the fitted values obtained with screening and no screening. Group screening metrics: the number of groups in $\mathcal{A}_g, \mathcal{S}_g, \mathcal{E}_g$, and $\mathcal{K}_g$. The results are shown for different values of p, averaged across $\rho \in \{0, 0.3, 0.6, 0.9\}$. Standard errors are shown.

Method	p	Runtime screen (s)	Runtime no screen (s)	card($\mathcal{A}_g$)	card($\mathcal{S}_g$)	card($\mathcal{E}_g$)	card($\mathcal{K}_g$)	Num it screen	Num it no screen	ℓ_2 dist to no screen
gSLOPE	500	49 ± 1	76 ± 1	26 ± 1	31 ± 1	31 ± 1	0.003 ± 0.004	31 ± 1	40 ± 1	$4 \times 10^{-8} \pm 5 \times 10^{-8}$
gSLOPE	1625	138 ± 3	203 ± 3	43 ± 1	54 ± 2	54 ± 2	0.004 ± 0.006	78 ± 2	79 ± 1	$1 \times 10^{-8} \pm 5 \times 10^{-9}$
gSLOPE	2750	987 ± 11	1641 ± 16	72 ± 2	95 ± 3	95 ± 3	0.008 ± 0.013	87 ± 2	89 ± 1	$1 \times 10^{-8} \pm 5 \times 10^{-9}$
gSLOPE	3875	1441 ± 26	2398 ± 31	98 ± 3	135 ± 3	135 ± 3	0.031 ± 0.054	95 ± 2	98 ± 1	$7 \times 10^{-9} \pm 3 \times 10^{-9}$
gSLOPE	5000	1454 ± 29	2727 ± 40	118 ± 3	168 ± 4	168 ± 4	0.022 ± 0.041	101 ± 2	109 ± 1	$4 \times 10^{-9} \pm 1 \times 10^{-9}$
SGS	500	118 ± 1	113 ± 1	28 ± 1	33 ± 1	38 ± 1	-	6 ± 0.3	29 ± 1	$8 \times 10^{-9} \pm 7 \times 10^{-10}$
SGS	1625	248 ± 2	538 ± 9	50 ± 2	59 ± 2	64 ± 2	-	7 ± 1	63 ± 3	$5 \times 10^{-9} \pm 1 \times 10^{-9}$
SGS	2750	374 ± 2	868 ± 12	85 ± 3	104 ± 2	115 ± 3	-	7 ± 0.4	57 ± 2	$3 \times 10^{-9} \pm 5 \times 10^{-10}$
SGS	3875	558 ± 4	1280 ± 19	116 ± 3	148 ± 3	164 ± 3	-	8 ± 0.4	54 ± 1	$2 \times 10^{-9} \pm 2 \times 10^{-10}$
SGS	5000	737 ± 5	1498 ± 19	141 ± 4	188 ± 3	209 ± 4	-	8 ± 0.4	54 ± 1	$1 \times 10^{-9} \pm 2 \times 10^{-10}$

F.4 Data Description

- carbotax: Carbotax study of ovarian tumor growth.
  - Response (continuous): Relative tumor volume ($\log_2$ scale).
  - Data matrix: Gene expression measurements. 10000 factors were randomly sampled from a collection of 34964.
  - Grouping structure: Variables are grouped using k-means clustering (Lloyd, 1982).
- scheetz: Gene expression data in the mammalian eye.
  - Response (continuous): Gene expression measurements for the Trim32 gene.
  - Data matrix: Gene expression measurements for other genes.
  - Grouping structure: Variables are grouped using k-means clustering (Lloyd, 1982).
- adenoma: Transcriptome profile data to identify formation of colorectal adenomas.
  - Response (binary): Binary labels for whether sample came from adenoma or normal mucosa.
  - Data matrix: Transcriptome profile measurements.
  - Grouping structure: Genes are assigned to pathways (groups) using the C3 regulatory target gene sets.[1]
- cancer: Breast cancer patients treated with tamoxifen for 5 years.
  - Response (binary): Binary labels classifying whether the cancer had recurred.
  - Data matrix: Gene expression data.
  - Grouping structure: Genes are assigned to pathways (groups) using the C3 regulatory target gene sets.[1]
- celiac: Gene expression data of primary leucocytes to classify celiac disease.
  - Response (binary): Binary labels as to whether a patient has celiac disease.
  - Data matrix: Gene expression measurements from the primary leucocytes.
  - Grouping structure: Genes are assigned to pathways (groups) using the C3 regulatory target gene sets.[1]
- colitis: Blood cells data for classifying whether a patient has colitis.
  - Response (binary): Binary labels classifying whether a patient has colitis.
  - Data matrix: Gene expression measurements.
  - Grouping structure: Genes are assigned to pathways (groups) using the C3 regulatory target gene sets.[1]
- tumour: Gene expression data of pancreative cancer samples to identify tumorous tissue.
  - Response (binary): Binary labels indicating if a sample is from tumour tissue.
  - Data matrix: Gene expression measurements.
  - Grouping structure: Genes are assigned to pathways (groups) using the C3 regulatory target gene sets.[1]

Table A8: Dataset information for the six datasets used in the real data analysis.

Dataset	p	n	m	Group sizes	Type	Source
carbotax	10000	101	100	$[1, 126]$	Linear	Koussounadis et al. (2014)[2]
scheetz	18975	120	379	$[1, 165]$	Linear	Scheetz et al. (2006)[2]
adenoma	17661	64	1849	$[1, 646]$	Logistic	Sabates-Bellver et al. (2007)[3]
cancer	7057	60	1277	$[1, 292]$	Logistic	Ma et al. (2004)[3]
celiac	14294	132	1666	$[1, 570]$	Logistic	Heap et al. (2009)[3]
colitis	11999	127	1528	$[1, 497]$	Logistic	Burczynski et al. (2006)[3]
tumour	17661	52	1849	$[1, 646]$	Logistic	Pei et al. (2009); Ellsworth et al. (2013); Li et al. (2016)[3]

[1] `gsea-msigdb.org/gsea/msigdb/human/collections.jsp`. Accessed 08/2024.
[2] downloaded from `https://iowabiostat.github.io/data-sets/`. Accessed 08/2024.
[3] downloaded from `https://www.ncbi.nlm.nih.gov/`. Accessed 09/2024.

F.5 Additional Results from the Real Data Analysis

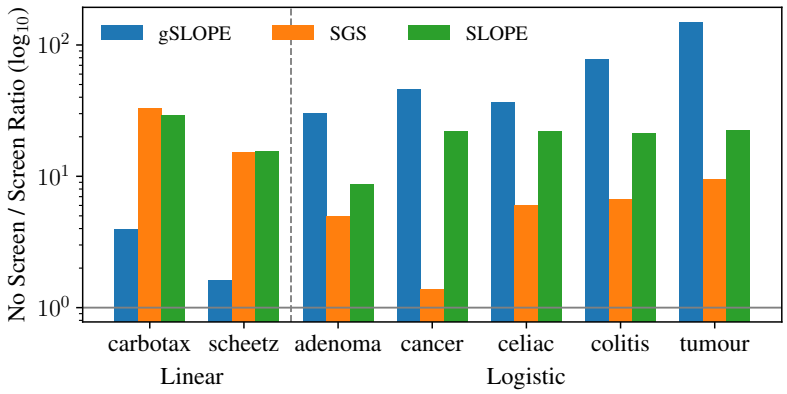


Figure A14: The ratio of no screen time to screen time (↑) of SLOPE, gSLOPE, and SGS applied to the real datasets, for fitting 100 path models, split into response type. The horizontal grey line represents no screening improvement.

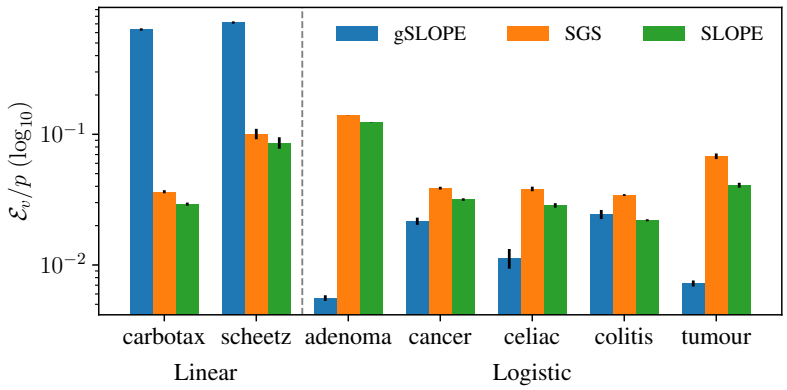


Figure A15: The ratio of the fitting set ($\mathcal{E}_v$) to the input dimensionality (p) (↓) of SLOPE, gSLOPE, and SGS applied to the real datasets, for fitting 100 path models, split into response type.

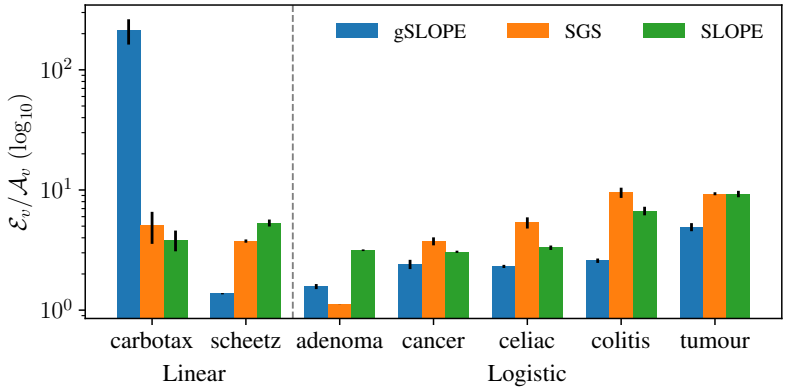


Figure A16: The ratio of the fitting set ($\mathcal{E}_v$) to the active set ($\mathcal{A}_v$) (↓) of SLOPE, gSLOPE, and SGS applied to the real datasets, for fitting 100 path models, split into response type.

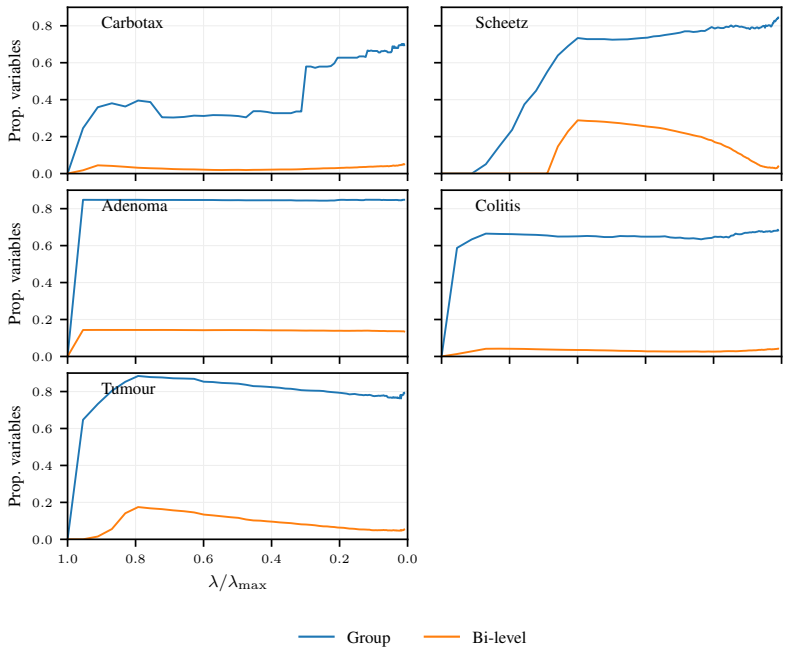


Figure A17: The proportion of variables in $\mathcal{S}_v$ relative to p for group-only and bi-level screening applied to SGS, plotted along the regularization path for the carbotax, scheetz, adenoma, colitis, and tumour datasets.

Table A9: Variable screening metrics for SGS applied to real data in Section 6.2. The number of variables in $\mathcal{A}_v, \mathcal{S}_v, \mathcal{E}_v$, and $\mathcal{K}_v$ are shown. The results are averaged across the nine pathway collections, with standard errors shown.

Method	Dataset	card($\mathcal{A}_v$)	card($\mathcal{S}_v$)	card($\mathcal{E}_v$)	card($\mathcal{K}_v$)
SGS	carbotax	144 ± 6	361 ± 10	364 ± 10	2 ± 0.50
SGS	scheetz	612 ± 74	1908 ± 174	1915 ± 175	7 ± 1.28
SGS	adenoma	2174 ± 23	2451 ± 4	2463 ± 5	12 ± 0.22
SGS	cancer	86 ± 3	273 ± 7	273 ± 7	1 ± 0.05
SGS	celiac	175 ± 10	545 ± 22	546 ± 22	1 ± 0.07
SGS	colitis	68 ± 4	409 ± 7	411 ± 7	2 ± 0.18
SGS	tumour	128 ± 5	1189 ± 59	1200 ± 59	12 ± 0.54

Table A10: General and group screening metrics for SGS and gSLOPE applied to real data in Section 6.2. General metrics: the runtime (in seconds) for screening against no screening, the number of fitting iterations for screening against no screening (with the number of occasions of failed convergence given in brackets), and the ℓ_2 distance between the fitted values obtained with screening and no screening. Group screening metrics: the number of groups in $\mathcal{A}_g, \mathcal{S}_g, \mathcal{E}_g$, and $\mathcal{K}_g$. The results are averaged across the nine pathway collections, with standard errors shown.

Method	Dataset	Runtime screen (s)	Runtime no screen (s)	card($\mathcal{A}_g$)	card($\mathcal{S}_g$)	card($\mathcal{E}_g$)	card($\mathcal{K}_g$)	Num it screen (num failed)	Num it no screen (num failed)	ℓ_2 dist to no screen
gSLOPE	carbotax	865	3456	71 ± 3	124 ± 2	124 ± 2	0.05 ± 0.05	$2367 \pm 176(1)$	$6433 \pm 303(10)$	$1 \times 10^{-9} \pm 2 \times 10^{-10}$
gSLOPE	scheetz	4731	7735	153 ± 6	212 ± 6	212 ± 6	0 ± 0	$6252 \pm 313(20)$	$7183 \pm 332(38)$	$8 \times 10^{-10} \pm 8 \times 10^{-10}$
gSLOPE	adenoma	415	12610	71 ± 6	77 ± 4	77 ± 4	0 ± 0	$3956 \pm 378(14)$	$7864 \pm 394(78)$	$9 \times 10^{-6} \pm 1 \times 10^{-6}$
gSLOPE	cancer	109	5003	46 ± 3	81 ± 5	81 ± 5	0 ± 0	$764 \pm 74(0)$	$6250 \pm 404(40)$	$4 \times 10^{-7} \pm 8 \times 10^{-8}$
gSLOPE	celiac	205	7530	34 ± 75	62 ± 9	62 ± 9	0 ± 0	$834 \pm 177(0)$	$4764 \pm 428(27)$	$4 \times 10^{-6} \pm 1 \times 10^{-5}$
gSLOPE	colitis	148	11567	69 ± 5	114 ± 8	114 ± 8	0 ± 0	$749 \pm 62(0)$	$9090 \pm 206(72)$	$4 \times 10^{-6} \pm 7 \times 10^{-7}$
gSLOPE	tumour	222	33405	24 ± 1	83 ± 4	83 ± 4	0 ± 0	$892 \pm 67(0)$	$9224 \pm 218(86)$	$4 \times 10^{-7} \pm 5 \times 10^{-8}$
SGS	carbotax	63	2067	40 ± 1	138 ± 1	119 ± 1	-	$168 \pm 60(0)$	$3572 \pm 350(3)$	$8 \times 10^{-10} \pm 3 \times 10^{-10}$
SGS	scheetz	358	5476	105 ± 4	203 ± 5	186 ± 6	-	$10 \pm 4(0)$	$5039 \pm 360(28)$	$7 \times 10^{-8} \pm 2 \times 10^{-8}$
SGS	adenoma	1234	6143	601 ± 6	789 ± 1	735 ± 1	-	$1172 \pm 69(0)$	$4224 \pm 338(20)$	$6 \times 10^{-7} \pm 5 \times 10^{-8}$
SGS	cancer	151	211	70 ± 2	295 ± 3	182 ± 4	-	$124 \pm 15(0)$	$147 \pm 14(0)$	$2 \times 10^{-8} \pm 2 \times 10^{-9}$
SGS	celiac	339	2035	120 ± 7	460 ± 10	288 ± 10	-	$55 \pm 12(0)$	$1259 \pm 112(0)$	$9 \times 10^{-5} \pm 2 \times 10^{-5}$
SGS	colitis	271	1821	52 ± 3	368 ± 3	230 ± 3	-	$54 \pm 10(0)$	$1411 \pm 188(0)$	$1 \times 10^{-6} \pm 3 \times 10^{-7}$
SGS	tumour	781	7451	94 ± 3	733 ± 10	479 ± 14	-	$33 \pm 7(0)$	$6006 \pm 287(23)$	$7 \times 10^{-7} \pm 2 \times 10^{-7}$