OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Dual Feature Reduction for the Sparse-Group Lasso and its Adaptive Variant

Fabio Feser FF120@IMPERIAL.AC.UK
Marina Evangelou M.EVANGELOU@IMPERIAL.AC.UK
Department of Mathematics, Imperial College London, UK

Abstract

The sparse-group lasso (SGL) performs both variable and group selection. It has found widespread use in many fields, due to its sparse-group penalty, which allows it to utilize grouping information and shrink inactive variables in active groups. However, SGL can be computationally expensive, due to the added shrinkage complexity. This paper introduces a feature reduction approach for SGL and the adaptive SGL, Dual Feature Reduction (DFR), which applies strong screening rules to reduce the input space before optimization. DFR applies two layers of screening and is based on dual norms. Through synthetic and real numerical studies, it is shown that DFR is the state-of-the-art screening rule for SGL by drastically reducing the computational cost under many different scenarios, outperforming other existing methods.

1. Introduction

High-dimensional datasets, where the number of features (p) is far greater than the number of observations (n) in a matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, are becoming increasingly common with the increased rate of data collection. To handle this, *shrinkage methods*, such as the lasso [39], elastic-net [49], and SLOPE [3] have been proposed and found increased use in the machine learning community [1, 20, 25, 38]. These methods shrink estimates towards zero during optimization, enabling *variable selection*, to identify which features, $\beta \in \mathbb{R}^p$, have an association with the response $y \in \mathbb{R}^n$.

In genetics, these methods help identify genes associated to disease outcomes. As genes are naturally found in groups (pathways), *group selection* approaches have been proposed, such as the group lasso [46], group SLOPE [4], and group SCAD [13]. Applying only group shrinkage can harm convergence and prediction, as noise variables in active groups are retained [11, 37].

This limitation led to the development of sparse-group models, such as the Sparse-group Lasso (SGL) [37] and Sparse-group SLOPE (SGS) [11]. These models apply shrinkage on both variables and groups to yield concurrent variable and group selection. SGL has found increased popularity in applications in the machine learning [41, 45] and healthcare [10, 31, 37] communities. It has consistently outperformed the lasso and group lasso in selection and prediction tasks [11, 37].

Suppose the variables sit in a grouping structure, with disjoint sets of groups $\mathcal{G}_1, \dots, \mathcal{G}_m$ of sizes $p_1, \dots, p_m$. Then, SGL is a convex combination of the lasso and group lasso [37]:

$$\hat{\beta}_{\text{sgl}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\arg\min} \{ f(\beta) + \lambda \|\beta\|_{\text{sgl}} \}, \text{ where } \|\beta\|_{\text{sgl}} = \alpha \|\beta\|_1 + (1-\alpha) \sum_{g=1}^{m} \sqrt{p_g} \|\beta^{(g)}\|_2, \quad (1)$$

such that f is a differentiable and convex loss function, $\lambda > 0$ defines the level of shrinkage,

$\beta^{(g)} \in \mathbb{R}^{p_g}$ is the vector of coefficients in group g, and $\alpha \in [0, 1]$. SGL has been extended to have adaptive shrinkage through the adaptive sparse-group lasso (aSGL) [24, 32] (Appendix B.1).

1.1. Feature reduction approaches for the sparse-group lasso

The benefits of SGL come with increased computational cost, due to the additional shrinkage and the tuning of two hyper-parameters. Typically, α is set subjectively (Simon et al. [37] suggest $\alpha = 0.95$) and λ is tuned using cross-validation along a path $\lambda_1 \geq \ldots \geq \lambda_l \geq 0$. Feature reduction techniques, including screening rules, help ease the cost by discarding inactive features before optimization.

Feature reduction techniques are either *exact* or *heuristic*. Exact methods strictly discard only inactive features but are conservative, while heuristic methods discard more features at the risk of violations. These violations are countered by checking the Karush–Kuhn–Tucker (KKT) optimality conditions [17] and adding any offending features back into the optimization. Heuristic rules discard significantly more variables than exact rules, providing large computational savings [40].

Most exact methods follow the seminal Safe Feature Elimination (SAFE) framework [7], but other exact methods include the dome test [44] and Dual Polytope Projections (DPP) [43]. The strong rule by Tibshirani et al. [40] provides a framework for applying heuristic reduction with single separable norms, which has been extended to non-separable [19] and sparse-group norms [12]. Other examples include Sure Independence Screening (SIS) [9] and the Hessian rule [18].

An exact reduction method for SGL, called GAP safe rules, was proposed by Ndiaye et al. [27] using the SAFE framework. GAP safe uses the duality gap to create feasible regions for active variables and applies reduction on the groups and variables. Other reduction methods for SGL include Two-layer Feature Reduction (TLFre) (exact) [42], though it was shown not to be exact [26], and sparsegl (heuristic) [23], which applies only group-level reduction. Additional speed-up attempts include using approximate bounds for inactive conditions [15] and a heuristic screening rule limited to multi-response Cox modeling [22].

1.2. Contributions

In this manuscript, we propose a new dual feature reduction method for SGL and aSGL, *Dual Feature Reduction* (DFR), which is based on the strong rule [40] and the bi-level framework for SGS [12]. DFR introduces the first bi-level strong (heuristic) screening rules for SGL and the first screening rules for aSGL. It applies two layers of screening, discarding inactive groups and inactive variables within active groups. Reducing input dimensionality before optimization enables broader tuning regimes, such as concurrent tuning of λ and α. The computational efficiency of DFR increases the accessibility of SGL and aSGL models, encouraging wider adoption across fields.

The GAP safe rules for SGL require computation of safe regions, which includes a radius and center, and the dual norm. In contrast, DFR only needs the dual norm, making it significantly less expensive, as evidenced by our results. Applied to synthetic and real data, DFR consistently delivers robust computational and input dimensionality improvements, outperforming other methods.

2. Theory

2.1. Problem statement

SGL is fit along a path of parameters $\lambda_1 \geq \ldots \geq \lambda_l \geq 0$. The objective is to use the solution at λ_k to generate a set of *candidate variables* $\mathcal{C}_v(\lambda_{k+1}) \subset [p] := \{1, \ldots, p\}$, that is a superset of the

(unknown) set of active variables, $\mathcal{A}_v(\lambda_{k+1}) := \{i \in [p] : \hat{\beta}_i(\lambda_{k+1}) \neq 0\}$. The optimization at λ_{k+1} (Equation 1) is then performed using only $\mathcal{C}_v(\lambda_{k+1})$, leading to large computational savings.

DFR starts by first generating a candidate group set (Section 2.2.1), which is then used to construct a candidate variable set (Section 2.2.2). DFR requires evaluating the dual norm of SGL, $\|z\|^*_{\text{sgl}} := \sup\{z^\top x : \|x\|_{\text{sgl}} \leq 1\}$, and this is found through the ϵ-norm [27] (Definition 3):

$$\|\beta\|_{\text{sgl}} = \sum_{g=1}^{m}(\alpha + (1-\alpha)\sqrt{p_g})\|\beta^{(g)}\|^*_{\epsilon_g} = \sum_{g=1}^{m}\tau_g\|\beta^{(g)}\|^*_{\epsilon_g}, \text{ where } \tau_g = \alpha + (1-\alpha)\sqrt{p_g}. \quad (2)$$

2.2. Dual feature reduction

DFR is first derived for SGL (Sections 2.2.1 and 2.2.2) and is subsequently expanded to aSGL (Appendix B.2), where DFR-aSGL requires the norm to be rewritten so that it can also be expressed as the ϵ-norm. The screening rules for DFR are summarised for reference in Table A1.

2.2.1. GROUP REDUCTION

To generate a candidate group set, the KKT stationarity conditions [17] are used, providing conditions for an inactive group. For SGL, for a group g at λ_{k+1} (using Equations 1 and 2), we have

$$\mathbf{0} \in \nabla_g f(\hat{\beta}(\lambda_{k+1})) + \tau_g\lambda_{k+1}\Theta^0_{g,k+1}, \text{ where } \Theta^0_{g,k+1} := \partial\|0\|^*_{\epsilon_g} = \left\{x \in \mathbb{R}^{p_g} : \|x\|_{\epsilon_g} \leq 1\right\} \quad (3)$$

is the subgradient of the dual norm of the ϵ-norm for an inactive group (at zero) [36]. Applying the ϵ-norm, the subgradient can be canceled out, so the KKT conditions can be written as

$$-\nabla_g f(\hat{\beta}(\lambda_{k+1})) \in \tau_g\lambda_{k+1}\Theta^0_{g,k+1} \overset{\|\cdot\|_{\epsilon_g}}{\iff} \|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g} = \tau_g\lambda_{k+1}\|\Theta^0_{g,k+1}\|_{\epsilon_g} \leq \tau_g\lambda_{k+1}. \quad (4)$$

If the gradient were available, it would be possible to identify the support at λ_{k+1} (Proposition 4). However, as this is not possible in practice, an approximation that allows for screening, using the gradient at λ_k and a Lipschitz assumption, is constructed instead, described in Proposition 1 (the derivation is provided in Appendix A.1.1).

Proposition 1 (DFR-SGL group screening) *For any $\lambda_{k+1}, k \in [l-1]$, assuming that*

$$\|\nabla_g f(\hat{\beta}(\lambda_{k+1})) - \nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g} \leq \tau_g|\lambda_{k+1} - \lambda_k|,$$

the candidate set $\mathcal{C}_g(\lambda_{k+1}) = \{g \in [m] : \|\nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g} > \tau_g(2\lambda_{k+1} - \lambda_k)\}$ is a superset of the set of active groups for SGL. That is, $\mathcal{A}_g(\lambda_{k+1}) \subset \mathcal{C}_g(\lambda_{k+1})$.

2.2.2. VARIABLE REDUCTION

Further reduction is possible by applying a second layer of screening to the variables in the candidate groups. For an inactive variable, $i \notin \mathcal{A}_v(\lambda_{k+1})$, the KKT conditions are (the subgradient of the ℓ_2 norm vanishes for an inactive variable)

$$\mathbf{0} \in \nabla_i f(\hat{\beta}(\lambda_{k+1})) + \lambda_{k+1}\alpha\Phi^0_{i,k+1}, \text{ where } \Phi^0_{i,k+1} = \{x \in \mathbb{R} : |x| \leq 1\} \quad (5)$$

is the subgradients of the ℓ_1 norm at zero (leading to Proposition 5). The derivation is similar to that for the group screening (it is described in Appendix A.1.2), culminating in Proposition 2. To derive Equation 5, knowledge of $\mathcal{A}_g(\lambda_{k+1})$ is required, but this is unknown. By Proposition 1, $\mathcal{A}_g(\lambda_{k+1}) \subset \mathcal{C}_g(\lambda_{k+1})$, and so the candidate set is used in practice for applying Proposition 2.

Proposition 2 (DFR-SGL variable screening) *For any $\lambda_{k+1}, k \in [l-1]$, assuming that*

$$|\nabla_i f(\hat{\beta}(\lambda_{k+1})) - \nabla_i f(\hat{\beta}(\lambda_k))| \leq \alpha(\lambda_k - \lambda_{k+1}),$$

the candidate set $\mathcal{C}_v(\lambda_{k+1}) = \{i \in \mathcal{G}_g \text{ for } g \in \mathcal{A}_g(\lambda_{k+1}) : |\nabla_i f(\hat{\beta}(\lambda_k))| > \alpha(2\lambda_{k+1} - \lambda_k)\}$ is a superset of the set of active variables for SGL. That is, $\mathcal{A}_v(\lambda_{k+1}) \subset \mathcal{C}_v(\lambda_{k+1})$.

2.3. Algorithm

The DFR algorithm (Algorithm A1) is based on the strong sparse-group screening framework, proposed by Feser and Evangelou [12]. The algorithm has the following key steps (for λ_{k+1}):

1. Group screening: find $\mathcal{C}_g(\lambda_{k+1})$ using Proposition 1.
2. Variable screening: find $\mathcal{C}_v(\lambda_{k+1})$ using Proposition 2 for $i \in \mathcal{G}_g, g \in \mathcal{C}_g(\lambda_{k+1})$.
3. Compute $\hat{\beta}_{\mathcal{O}_v}(\lambda_{k+1})$ using the *optimization set* $\mathcal{O}_v = \mathcal{C}_v(\lambda_{k+1}) \cup \mathcal{A}_v(\lambda_k)$, with the active set included due to models being nested. Perform KKT checks to identify any violations (Appendix A.2) and add offending variables into $\mathcal{O}_v$. Repeat this step until no violations.

The two main computational costs of Algorithm A1 are calculating the solution and the ϵ-norm. The former depends on the fitting algorithm, with proximal and descent algorithms typically having complexities of $O(tp^2)$, for t iterations [48]. The latter has a worst-case cost of $O(p_g \log p_g)$ [27].

3. Numerical results

This section evaluates the efficiency and robustness of DFR using synthetic and real data with varying characteristics. Two metrics are used to measure reductions in dimensionality and cost:

- *Improvement factor = no screen time / screen time*, which measures the reduction in computational fitting time due to screening.
- *Input proportion* = $\mathcal{O}_v/p$, which measures the proportion of the input space used for fitting.

DFR is compared with the existing SGL screening rules sparsegl [23] and GAP safe [27] (see Appendix C for descriptions). Table A1 summarizes all the rules considered. For DFR and sparsegl, optimization is performed using the Adaptive Three Operator Splitting (ATOS) [29] algorithm, but DFR works with any fitting algorithm. In Sections 3.1 and 3.2 we present the main results, with additional commentary and results presented in Appendices D and E.

3.1. Synthetic data analysis

The data for this section is generated using the linear model $y = \mathbf{X}\beta + \epsilon$, where $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \in \mathbb{R}^{200\times 1000}$, $\beta \sim \mathcal{N}(0, 4)$, and $\epsilon \sim \mathcal{N}(0, 1)$. Correlation in $\mathbf{X}$ is applied within each group, with $\Sigma_{i,j} = \rho = 0.3$ for i and j in the same group. Variables are grouped into $m = 22$ uneven groups of sizes in $[3, 100]$, with a 0.2 active group proportion and a 0.2 active proportion for variables in an active group. Models were fit along a 50-length path from λ_1 (chosen to generate the null model, see Appendices A.4 and B.2.3) to $\lambda_{50} = 0.1\lambda_1$. Each simulation was repeated 100 times, with results averaged across the repeats, unless stated otherwise. Detailed setup information is in Appendix D.2.

Comparison to GAP safe. The improvement factor for DFR is significantly superior to that of the GAP safe rules (Figure 1). Although the input proportion of DFR and GAP safe are of similar levels (Figure A2), the cost of calculating safe regions appears to nullify any gain in dimensionality reduction. This shows that the two reduction approaches (heuristic vs exact) arrive at very similar results (the screened sets), but DFR achieves this with significantly greater computational efficiency.

DFR, through bi-level screening, shows a tangible benefit under increasing dimensionality (Figure 1). This is further illustrated in the analysis of interaction data (Appendix D.3.4), where DFR also provided large computational savings, making it more feasible for SGL and aSGL to be applied to problems such as gene-gene and gene-environment interaction detection [6, 47].

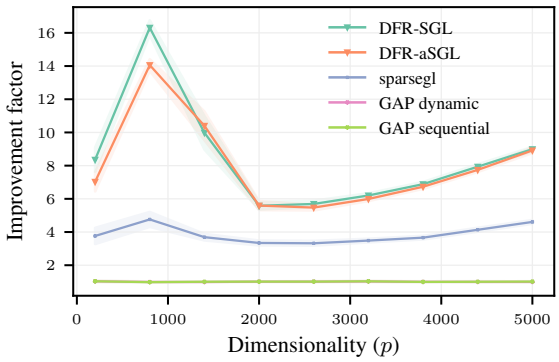

Figure 1: The improvement factor for strong against safe rules, applied to synthetic data under even groups of sizes 20, as a function of p, with 95% confidence intervals.

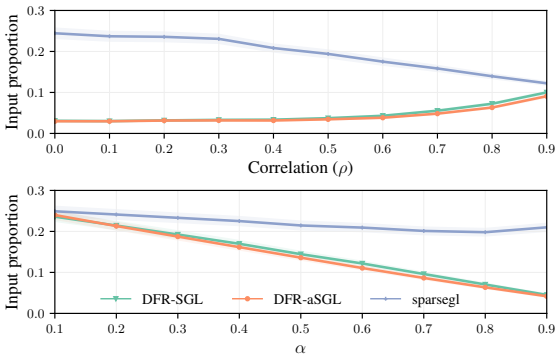

Figure 2: The input proportion for the screening methods applied to synthetic data, as a function of the data correlation (top) and α (below), with 95% confidence intervals.

Robustness. DFR is found to be robust under various scenarios. Under all considered values of group correlation in $\mathbf{X}$, DFR is more effective at reducing the input space when compared to sparsegl, especially under minor correlation (Figures 2 and A5). Under higher correlation, the models become less sparse, resulting in reduced screening importance. Across different α values, DFR also efficiently reduces the input space (Figure 2), with the screening efficiency decreasing linearly, showing that DFR can be used to tune a hyperparameter grid (λ, α). As $\alpha \to 0$, SGL is forced to select more variables in a group, limiting reduction potential. In such scenarios, the second layer of screening is not as important, as shown by the similar performances of DFR and sparsegl.

DFR was also found to be robust under varying signal sparsity and strength, and varying weight hyperparameters for aSGL (Appendix D.3.3). Additionally, DFR provided strong benefits when applied to cross-validation (Appendix D.3.5) and logistic models (Appendix D.4).

3.2. Real data analysis

The efficiency of DFR is further evaluated through the analysis of six real datasets with different characteristics, including continuous and binary response variables, and low- and high-dimensional data (see Appendix E for details). Models were fit along a 100-length path, terminating at $0.2\lambda_1$.

For all datasets considered, DFR outperforms sparsegl for improving computational cost and reducing input proportion (Figure 3), keeping the proportion low across the full path (Figure A13). DFR-aSGL in particular was found to be very effective for the *scheetz* and *adenoma* datasets. As

sparsegl only screens groups, sparsegl is forced to fit with full groups, which is a limitation when there are large groups present (see Table A39). Through bi-level screening, DFR overcomes this restriction.

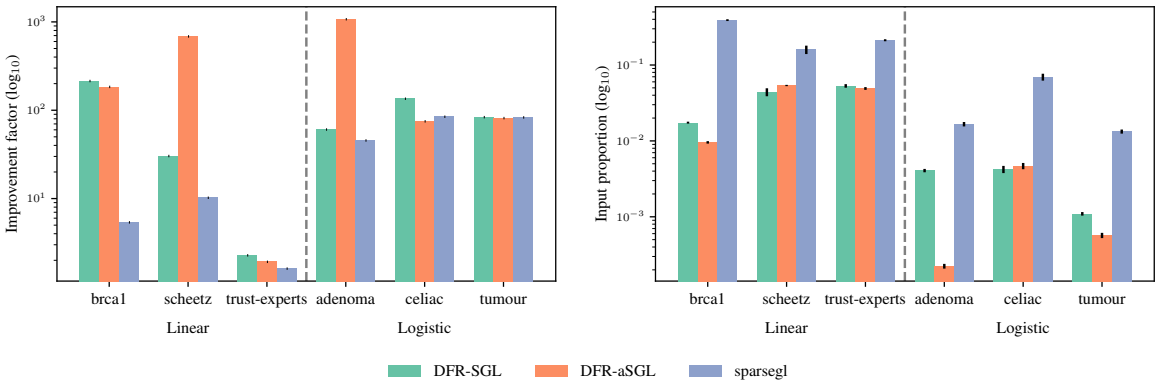

Figure 3: The improvement factor (high is best) and input proportion (low is best) ($\log_{10}$ scale) of the screening methods applied to six real datasets, split by model type.

4. Discussion

In this paper, a new feature reduction approach for the sparse-group lasso and adaptive sparse-group lasso is introduced, called *Dual Feature Reduction* (DFR). DFR establishes the first bi-level strong screening rules for SGL and the first screening rules for aSGL.

DFR applies two layers of reduction using the dual norm of SGL with strong screening rules to efficiently reduce the input dimensionality for optimization and is computationally simpler than the GAP safe rules. By discarding variables that would have been inactive at the optimal solution, DFR achieves significant computational savings, allowing the SGL family of models to be more efficiently implemented and applied to larger and more complex datasets. This gain comes at no cost, as the optimal solution is still achieved (Appendices D.3.6, D.4, and E.3). In fact, by reducing the input dimensionality, instances were observed where DFR helped SGL overcome convergence issues that would have occurred otherwise (Table A42).

DFR proved robust across different data and model parameters, achieving drastic feature reduction under all scenarios tested. This consistently translated into large computational savings across both real and synthetic data. DFR also outperformed other screening approaches for SGL, under all considered situations, showing the benefit of applying two layers of strong screening.

Limitations. Several assumptions were required to perform two layers of feature reduction for DFR. For both variable and group screening, Propositions 1 and 2 use Lipschitz assumptions which are consistent with the general strong framework [40]. Any breach of assumptions are guarded against using KKT checks. Only a single KKT violation occurred for SGL across all our simulations and only very infrequently for aSGL (Appendix D.3.1). However, there may be scenarios that we did not consider where the assumptions break down. This is a limitation of any strong screening rule, although DFR, in particular, carries additional assumptions over other strong rules, which are necessary when applying the second layer of screening.

Disclosure of Funding

This work was supported by EPSRC through the Modern Statistics and Statistical Machine Learning (StatML) CDT programme, grant no. EP/S023151/1.

References

[1] Ahmed Alaoui and Michael W Mahoney. Fast Randomized Kernel Ridge Regression with Statistical Guarantees. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[2] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, Cham, 2017. ISBN 978-3-319-48310-8. doi: 10.1007/978-3-319-48311-5.

[3] 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.

[4] 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.

[5] Oleg Burdakov. A new vector norm for nonlinear curve fitting and some other optimization problems. In *Mathematische Optimierung — Theorie und Anwendungen*, volume 33 of *Int. Wiss. Kolloq. Fortragsreihe*, pages 15–17, 1988.

[6] Gina M D'Angelo, DC Rao, and C Charles Gu. Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *BMC Proceedings*, 3(S7):S62, 2009. ISSN 1753-6561. doi: 10.1186/1753-6561-3-S7-S62.

[7] 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.

[8] 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.

[9] 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.

[10] Kuangnan Fang, Xiaoyan Wang, Shengwei Zhang, Jianping Zhu, and Shuangge Ma. Bi-level variable selection via adaptive sparse group Lasso. *Journal of Statistical Computation and Simulation*, 85(13):2750–2760, 2015. ISSN 15635163. doi: 10.1080/00949655.2014.938241.

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

[12] Fabio Feser and Marina Evangelou. Strong screening rules for group-based SLOPE models. *arXiv preprint arXiv:2405.15357*, 2024.

[13] Xiao Guo, Hai Zhang, Yao Wang, and Jiang-Lun Wu. Model selection and estimation in high dimensional regression models with group SCAD. *Statistics & Probability Letters*, 103:86–92, 2015. ISSN 01677152. doi: 10.1016/j.spl.2015.04.017.

[14] 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.

[15] Yasutoshi Ida, Yasuhiro Fujiwara, and Hisashi Kashima. Fast Sparse Group Lasso. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[16] National Cancer Institute. The cancer genome atlas program. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

[17] 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.

[18] Johan Larsson and Jonas Wallin. The Hessian screening rule. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2024. ISBN 9781713871088.

[19] 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.

[20] Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. LassoNet: A Neural Network with Feature Sparsity. *Journal of Machine Learning Research*, 22(127):1–29, 2021.

[21] 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.

[22] Ruilin Li, Christopher Chang, Johanne M Justesen, Yosuke Tanigawa, Junyang Qian, Trevor Hastie, Manuel A Rivas, and Robert Tibshirani. Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. *Biostatistics*, 23(2):522–540, 2022. ISSN 1465-4644. doi: 10.1093/biostatistics/kxaa038.

[23] 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.

[24] Alvaro Mendez-Civieta, M. Carmen Aguilera-Morillo, and Rosa E. Lillo. Adaptive sparse group LASSO in quantile regression. *Advances in Data Analysis and Classification*, 15(3): 547–573, 2021. ISSN 1862-5347. doi: 10.1007/s11634-020-00413-8.

[25] Tom Michoel. Analytic solution and stationary phase approximation for the Bayesian lasso and elastic net. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[26] Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. GAP Safe screening rules for sparse multi-task and multi-class models. *Advances in Neural Information Processing Systems*, 2015-January:811–819, 2015. ISSN 10495258.

[27] 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., 2016.

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

[29] 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.

[30] 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.

[31] Jie Peng, Ji Zhu, Anna Bergamaschi, Wonshik Han, Dong-Young Noh, Jonathan R. Pollack, and Pei Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The Annals of Applied Statistics*, 4 (1), 2010. ISSN 1932-6157. doi: 10.1214/09-AOAS271.

[32] Benjamin Poignard. Asymptotic theory of the adaptive Sparse Group Lasso. *Annals of the Institute of Statistical Mathematics*, 72(1):297–328, 2020. ISSN 0020-3157. doi: 10.1007/s10463-018-0692-7.

[33] 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.

[34] Joshua A. Salomon, Alex Reinhart, Alyssa Bilinski, Eu Jing Chua, Wichada La Motte-Kerr, Minttu M. Rönn, Marissa B. Reitsma, Katherine A. Morris, Sarah LaRocca, Tamer H. Farag, Frauke Kreuter, Roni Rosenfeld, and Ryan J. Tibshirani. The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination. *Proceedings of the National Academy of Sciences*, 118 (51), 2021. ISSN 0027-8424. doi: 10.1073/pnas.2111454118.

[35] 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.

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

[37] 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.

[38] Ryan Thompson, Amir Dezfouli, and Robert Kohn. The Contextual Lasso: Sparse Linear Models via Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 19940–19961. Curran Associates, Inc., 2023.

[39] 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.

[40] 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.

[41] M. Vidyasagar. Machine learning methods in the computational biology of cancer. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 470(2167): 20140081, 2014. ISSN 1364-5021. doi: 10.1098/rspa.2014.0081.

[42] 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.

[43] 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.

[44] 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.

[45] Dani Yogatama and Noah A. Smith. Linguistic Structured Sparsity in Text Categorization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–796, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1074.

[46] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1): 49–67, 2006. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00532.x.

[47] Natalia Zemlianskaia, W. James Gauderman, and Juan Pablo Lewinger. A Scalable Hierarchical Lasso for Gene–Environment Interactions. *Journal of Computational and Graphical Statistics*, 31(4):1091–1103, 2022. ISSN 1061-8600. doi: 10.1080/10618600.2022.2039161.

[48] 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.

[49] Hui Zou and Trevor Hastie. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x.

Appendix

Appendix A. Sparse-group lasso

A.1. Theory

Definition 3 (ϵ-norm) *The ϵ-norm, $\|x\|_{\epsilon_g}$, applied to SGL is defined as the unique nonnegative solution q of the equation [5]*

$$\sum_{i=1}^{p_g}(|x_i| - (1-\epsilon_g)q)_+^2 = (\epsilon_g q)^2, \quad \textit{where} \quad \epsilon_g = \frac{\tau_g - \alpha}{\tau_g}. \tag{6}$$

Using Definition 3 and Ndiaye et al. [27], the dual norm of SGL applied to a group g can be formulated as

$$\|\xi^{(g)}\|_{\text{sgl}}^* = \max_{g=1,\dots,m} \tau_g^{-1}\|\xi^{(g)}\|_{\epsilon_g}. \tag{7}$$

A.1.1. GROUP REDUCTION

From Equation 4 we obtain

$$\|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g} \leq \tau_g \lambda_{k+1}. \tag{8}$$

Now if the gradient at λ_{k+1} were available, screening would be possible (Proposition 4). Instead, we seek an approximation $\mathcal{M}_g$ such that

$$\|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g} \leq \mathcal{M}_g. \tag{9}$$

Then, the screening rule tests whether $\mathcal{M}_g \leq \tau_g \lambda_{k+1}$. If this is found to be true, it can be concluded that Equation 4 holds and so the group must be inactive. An approximation can be found by assuming that the gradient is a Lipschitz function of λ_{k+1} with respect to the ϵ-norm,

$$\|\nabla_g f(\hat{\beta}(\lambda_{k+1})) - \nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g} \leq \tau_g |\lambda_{k+1} - \lambda_k|, \tag{10}$$

which is a similar assumption to the lasso strong rule [40]. Using the reverse triangle inequality, we have

$$\|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g} \leq \|\nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g} + \tau_g(\lambda_k - \lambda_{k+1}) =: \mathcal{M}_g, \tag{11}$$

yielding a suitable approximation $\mathcal{M}_g$. Therefore, the strong group screening rule for SGL can be formulated by plugging the approximation from Equation 11 into Equation 9: discard a group g if

$$\|\nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g} + \tau_g(\lambda_k - \lambda_{k+1}) \leq \tau_g \lambda_{k+1} \iff \|\nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g} \leq \tau_g(2\lambda_{k+1} - \lambda_k), \tag{12}$$

which leads to Proposition 1.

Proof [Proof of Proposition 1] To prove the candidate set is a superset of the active set, we need to prove that for any $g \in [m]$ and $k \in [l-1]$, $g \in \mathcal{A}_g(\lambda_{k+1}) \implies g \in \mathcal{C}_g(\lambda_{k+1})$. We instead prove the contrapositive: $g \notin \mathcal{C}_g(\lambda_{k+1}) \implies g \notin \mathcal{A}_g(\lambda_{k+1})$. First, we rewrite the Lipschitz assumption as (using the reverse triangle inequality)

$$\begin{aligned}\|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g} - \|\nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g} \leq \|\nabla_g f(\hat{\beta}(\lambda_{k+1})) - \nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g} \leq \tau_g|\lambda_{k+1} - \lambda_k| \\ \implies \|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g} \leq \|\nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g} + \tau_g|\lambda_{k+1} - \lambda_k|.\end{aligned} \tag{13}$$

Now, as $g \notin \mathcal{C}_g(\lambda_{k+1})$,

$$\|\nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g} \leq \tau_g(2\lambda_{k+1} - \lambda_k).$$

Plugging this into Equation 13 yields

$$\begin{aligned}
&\|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g} \leq \tau_g(2\lambda_{k+1} - \lambda_k) + \tau_g|\lambda_{k+1} - \lambda_k| \\
\implies &\|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g} \leq \tau_g\lambda_{k+1} \\
\implies &-\nabla_g f(\hat{\beta}(\lambda_{k+1})) \in \tau_g\lambda_{k+1}\Theta^0_{g,k+1}, && \text{as } \Theta^0_{g,k+1} = \left\{x \in \mathbb{R}^{p_g} : \|x\|_{\epsilon_g} \leq 1\right\} \\
\implies &\mathbf{0} \in \nabla_g f(\hat{\beta}(\lambda_{k+1})) + \tau_g\lambda_{k+1}\Theta^0_{g,k+1} \\
\implies &g \notin \mathcal{A}_g(\lambda_{k+1}), && \text{by the KKT conditions (Equation 3).}
\end{aligned}$$

■

Proposition 4 (Theoretical SGL group screening) *For any $\lambda_{k+1}, k \in [l-1]$, the candidate set $\mathcal{C}_g(\lambda_{k+1}) = \{g \in [m] : \|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g} > \tau_g\lambda_{k+1}\}$ recovers the exact support of the active groups for SGL. That is, $\mathcal{C}_g(\lambda_{k+1}) = \mathcal{A}_g(\lambda_{k+1}) := \{g \in [m] : \|\hat{\beta}^{(g)}(\lambda_{k+1})\|_2 \neq 0\}$.*

Proof [Proof of Proposition 4] To prove the two sets are equivalent, we need to prove that for any $g \in [m]$ and $k \in [l-1]$, $g \in \mathcal{A}_g(\lambda_{k+1}) \iff g \in \mathcal{C}_g(\lambda_{k+1})$. We instead prove the contrapositive: $g \notin \mathcal{C}_g(\lambda_{k+1}) \iff g \notin \mathcal{A}_g(\lambda_{k+1})$. So,

$$\begin{aligned}
g \notin \mathcal{C}_g(\lambda_{k+1}) &\iff \|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g} \leq \tau_g\lambda_{k+1}, && \text{by definition of the candidate set} \\
&\iff -\nabla_g f(\hat{\beta}(\lambda_{k+1})) \in \tau_g\lambda_{k+1}\Theta^0_{g,k+1}, && \text{as } \Theta^0_{g,k+1} = \left\{x \in \mathbb{R}^{p_g} : \|x\|_{\epsilon_g} \leq 1\right\} \\
&\iff \mathbf{0} \in \nabla_g f(\hat{\beta}(\lambda_{k+1})) + \tau_g\lambda_{k+1}\Theta^0_{g,k+1} \\
&\iff g \notin \mathcal{A}_g(\lambda_{k+1}), && \text{by the KKT conditions (Equation 3).}
\end{aligned}$$

■

A.1.2. VARIABLE REDUCTION

Applying the absolute value to Equation 5 leads to

$$-\nabla_i f(\hat{\beta}(\lambda_{k+1})) \in \lambda_{k+1}\alpha\Phi^0_{i,k+1} \overset{|\cdot|}{\iff} |\nabla_i f(\hat{\beta}(\lambda_{k+1}))| \leq \lambda_{k+1}\alpha. \tag{14}$$

As before, if we had access to the gradient we could screen via Proposition 5. The scenario is similar to the strong screening rule for the lasso [40], scaled by α. Therefore, using the Lipschitz assumption

$$|\nabla_i f(\hat{\beta}(\lambda_{k+1})) - \nabla_i f(\hat{\beta}(\lambda_k))| \leq \alpha(\lambda_k - \lambda_{k+1}),$$

yields the strong variable screening rule for SGL: discard a variable j in an active group g if

$$|\nabla_i f(\hat{\beta}(\lambda_k))| \leq \alpha(2\lambda_{k+1} - \lambda_k), \tag{15}$$

which leads to Proposition 2.

Proof [Proof of Proposition 2] The proof strategy is similar to that of Proposition 1. To prove the candidate set is a superset of the active set, we need to prove that for any $i \in \mathcal{G}_g$ such that $g \in \mathcal{A}_g$, and $k \in [l-1]$, $i \in \mathcal{A}_v(\lambda_{k+1}) \implies i \in \mathcal{C}_v(\lambda_{k+1})$. We instead prove the contrapositive: $i \notin \mathcal{C}_v(\lambda_{k+1}) \implies i \notin \mathcal{A}_v(\lambda_{k+1})$. First, we rewrite the Lipschitz assumption as (using the reverse triangle inequality)

$$|\nabla_i f(\hat{\beta}(\lambda_{k+1}))| \leq \|\nabla_i f(\hat{\beta}(\lambda_k))| + \alpha|\lambda_{k+1} - \lambda_k|. \tag{16}$$

Now, as $i \notin \mathcal{C}_v(\lambda_{k+1})$,

$$|\nabla_i f(\hat{\beta}(\lambda_k))| \leq \alpha(2\lambda_{k+1} - \lambda_k).$$

Plugging this into Equation 16 yields

$$\begin{aligned}
&|\nabla_i f(\hat{\beta}(\lambda_{k+1}))| \leq \alpha\lambda_{k+1} \\
\implies &-\nabla_i f(\hat{\beta}(\lambda_{k+1})) \in \alpha\lambda_{k+1}\Phi^0_{i,k+1}, \text{ as } \Phi^0_{i,k+1} = \{x \in \mathbb{R} : |x| \leq 1\} \\
\implies &\mathbf{0} \in \nabla_i f(\hat{\beta}(\lambda_{k+1})) + \alpha\lambda_{k+1}\Phi^0_{i,k+1} \\
\implies &i \notin \mathcal{A}_v(\lambda_{k+1}), \qquad \text{by the KKT conditions (Equation 3).}
\end{aligned}$$

■

Proposition 5 (Theoretical SGL variable screening) *For any $\lambda_{k+1}, k \in [l-1]$, the candidate set $\mathcal{C}_v(\lambda_{k+1}) = \{i \in \mathcal{G}_g \text{ for } g \in \mathcal{A}_g(\lambda_{k+1}) : |\nabla_i f(\hat{\beta}(\lambda_{k+1}))| > \lambda_{k+1}\alpha\}$ recovers the exact support of the active variables for SGL. That is, $\mathcal{C}_v(\lambda_{k+1}) = \mathcal{A}_v(\lambda_{k+1})$.*

Proof [Proof of Proposition 5] The proof strategy is similar to that of Proposition 4. To prove the two sets are equivalent, we need to prove that for any $i \in \mathcal{G}_g$ such that $g \in \mathcal{A}_g$, and $k \in [l-1]$, $i \in \mathcal{A}_v(\lambda_{k+1}) \iff i \in \mathcal{C}_v(\lambda_{k+1})$. We instead prove the contrapositive: $i \notin \mathcal{C}_v(\lambda_{k+1}) \iff i \notin \mathcal{A}_v(\lambda_{k+1})$. So,

$$\begin{aligned}
i \notin \mathcal{C}_v(\lambda_{k+1}) &\iff |\nabla_i f(\hat{\beta}(\lambda_{k+1}))| \leq \lambda_{k+1}\alpha, && \text{by definition of the candidate set} \\
&\iff -\nabla_v f(\hat{\beta}(\lambda_{k+1})) \in \lambda_{k+1}\alpha\Phi^0_{i,k+1}, && \text{as } \Phi^0_{i,k+1} = \{x \in \mathbb{R} : |x| \leq 1\}, \\
& && \text{for } i \in \mathcal{G}_g, g \in \mathcal{A}_g(\lambda_{k+1}) \\
&\iff \mathbf{0} \in \nabla_v f(\hat{\beta}(\lambda_{k+1})) + \lambda_{k+1}\alpha\Phi^0_{i,k+1} \\
&\iff i \notin \mathcal{A}_v(\lambda_{k+1}), && \text{by the KKT conditions (Equation 5).}
\end{aligned}$$

■

A.2. Karush–Kuhn–Tucker (KKT) checks

The screening rules of DFR use several Lipschitz assumptions (Propositions 1 and 2), as well as approximating the group active set by the group candidate set for the variable screening step (Section 2.2.2). When these assumptions fail, the screening rules can make mistakes and incorrectly exclude active variables. To protect against this, the KKT conditions are checked for each variable after screening, and violations are added back into the optimization.

To check whether a variable $i \in \mathcal{G}_g$ has been correctly discarded, the KKT optimality conditions are checked. Equation 5 describes the condition under which a variable $i \in \mathcal{G}_g$ is inactive and can be rewritten as, for a general variable (by the definition of $\Phi^0_{i,k+1}$)

$$|\nabla_i f(\hat{\beta}(\lambda_{k+1})) + \lambda_{k+1}(1-\alpha)\Psi^{(g)}_{i,k+1}| \leq \lambda_{k+1}\alpha, \tag{17}$$

where $\Psi^{(g)}_{k+1} = \{x \in \mathbb{R}^{\sqrt{p_g}} : \|x\|_2 \leq 1\}$ is the subgradient of the ℓ_2 norm. To satisfy Equation 17, the unknown subdifferential, $\Psi^{(g)}_{i,k+1}$, is taken to be its minimum possible value. For $x \in \Psi^{(g)}_{k+1}$, we have that

$$\begin{aligned} \|x\|_2 \leq 1 &\implies \sqrt{p_g}\|x\|_2 \leq \sqrt{p_g} \\ &\implies \|x\|_1 \leq \sqrt{p_g} \text{ by the inequality } \|x\|_1 \leq \sqrt{p_g}\|x\|_2 \\ &\implies |x_i| \leq \sqrt{p_g}. \end{aligned}$$

Hence, the values in the subdifferential are bound by $\sqrt{p_g}$. We consider the following scenarios for Equation 17:

1. $\nabla_i f(\hat{\beta}(\lambda_{k+1})) > \lambda_{k+1}(1-\alpha)\sqrt{p_g}$: choose $x_i = -\sqrt{p_g}$.
2. $\nabla_i f(\hat{\beta}(\lambda_{k+1})) < -\lambda_{k+1}(1-\alpha)\sqrt{p_g}$: choose $x_i = \sqrt{p_g}$.
3. $\nabla_i f(\hat{\beta}(\lambda_{k+1})) \in [-\lambda_{k+1}(1-\alpha)\sqrt{p_g}, \lambda_{k+1}(1-\alpha)\sqrt{p_g}]$: choose $y_i = \frac{\nabla_i f(\hat{\beta}(\lambda_{k+1}))}{\lambda_{k+1}(1-\alpha)\sqrt{p_g}}$.

We can now rewrite Equation 17 using the soft-thresholding operator as

$$|S(\nabla_i f(\hat{\beta}(\lambda_{k+1})), \lambda_{k+1}(1-\alpha)\sqrt{p_g})| \leq \lambda_{k+1}\alpha, \tag{18}$$

where $S(a,b) = \text{sign}(a)(|a| - b)_+$. Therefore, a KKT violation occurs for variable $i \in \mathcal{G}_g$ if Equation 17 does not hold. A violating variable is added back into the optimization procedure (see Section 2.3). A similar derivation can be found in Simon et al. [37] to derive conditions to check whether a group is active for SGL.

A.3. Algorithm

Algorithm A1 Dual Feature Reduction (DFR) for SGL

Input: $(\lambda_1, \dots, \lambda_l) \in \mathbb{R}^l$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, $\alpha \in [0, 1]$
compute $\hat{\beta}(\lambda_1)$ using Equation 1
for $k = 1$ **to** $l - 1$ **do**
  $\mathcal{C}_g(\lambda_{k+1}) \leftarrow$ candidate groups from Proposition 1
  $\mathcal{C}_v(\lambda_{k+1}) \leftarrow$ candidate variables from Proposition 2 for $i \in \mathcal{G}_g, g \in \mathcal{C}_g(\lambda_{k+1})$, and $i \notin \mathcal{A}_v(\lambda_k)$
  $\mathcal{O}_v \leftarrow \mathcal{C}_v(\lambda_{k+1}) \cup \mathcal{A}_v(\lambda_k)$ ▶ Optimization set
  compute $\hat{\beta}_i(\lambda_{k+1}), i \in \mathcal{O}_v$, using Equation 1
  $\mathcal{K}_v \leftarrow$ variable KKT violations for $i \notin \mathcal{O}_v$, using Equation 18 ▶ KKT check
  while $\text{card}(\mathcal{K}_v) > 0$ **do**
    $\mathcal{O}_v \leftarrow \mathcal{O}_v \cup \mathcal{K}_v$ ▶ Optimization set
    compute $\hat{\beta}_i(\lambda_{k+1}), i \in \mathcal{O}_v$, using Equation 1
    $\mathcal{K}_v \leftarrow$ variable KKT violations for $i \notin \mathcal{O}_v$ using Equation 18 ▶ KKT check
  end while
end for
Output: $\hat{\beta}_{\text{sgl}}(\lambda_1), \dots, \hat{\beta}_{\text{sgl}}(\lambda_l) \in \mathbb{R}^p$

A.4. Path start

When fitting SGL along a path of values, $\lambda_1 \geq \dots \geq \lambda_l \geq 0$, λ_1 is often chosen to be the exact point at which the first predictor becomes non-zero. By Ndiaye et al. [28] and using the dual norm from Equation 7, this value is given by

$$\lambda_1 = \|\nabla f(0)\|^*_{\text{sgl}} = \max_{g=1,\dots,m} \tau_g^{-1} \|\nabla_g f(0)\|_{\epsilon_g}.$$

A.5. Reduction to (adaptive) lasso and (adaptive) group lasso

Under $\alpha = 1$, SGL reduces to the lasso. In this case, no group screening occurs and the variable screening rule reduces to the lasso strong rule [40]:

$$|\nabla_i f(\hat{\beta}(\lambda_k))| \leq 2\lambda_{k+1} - \lambda_k.$$

Under $\alpha = 0$, SGL reduces to the group lasso. Under this scenario, the group variable screening reduces to the group lasso strong rule [40]:

$$\|\nabla_g f(\hat{\beta}(\lambda_k))\|_2 \leq \sqrt{p_g}(2\lambda_{k+1} - \lambda_k).$$

and no variable screening is performed. For aSGL (Appendix B), the rules reduce to the adaptive lasso and adaptive group lasso:

$$\text{Adaptive lasso: } |\nabla_i f(\hat{\beta}(\lambda_k))| \leq v_i(2\lambda_{k+1} - \lambda_k) \implies \hat{\beta}_i(\lambda_{k+1}) = 0.$$
$$\text{Adaptive group lasso: } \|\nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon'_g} \leq w_g\sqrt{p_g}(2\lambda_{k+1} - \lambda_k) \implies \hat{\beta}^{(g)}(\lambda_{k+1}) \equiv \mathbf{0},$$

where $\epsilon'_g = 1$ in the ϵ-norm (Definition 3 and Equation 20).

Appendix B. Adaptive sparse-group lasso

B.1. Definition

The *Adaptive Sparse-group Lasso* (aSGL) applies adaptive shrinkage in a sparse-group setting, achieving the oracle property in a double-asymptotic framework, and has the norm [24, 32]

$$\|\beta\|_{\text{asgl}} = \alpha \sum_{i=1}^{p} v_i |\beta_i| + (1-\alpha) \sum_{g=1}^{m} w_g \sqrt{p_g} \|\beta^{(g)}\|_2, \tag{19}$$

where v_i and w_g are adaptive weights. aSGL has a less straightforward, but nonetheless useful, connection to the ϵ-norm, which allows for screening rules to be derived. The aSGL norm can be rewritten as (with the derivation given in Appendix B.2.1)

$$\|\beta\|_{\text{asgl}} = \sum_{g=1}^{m} \gamma_g \|\beta^{(g)}\|_{\epsilon_g'}^*, \text{ where} \tag{20}$$

$$\gamma_g = \alpha \|v^{(g)}\|_1 - \frac{\alpha}{\|\hat{\beta}^{(g)}\|_1} \sum_{i,j \in \mathcal{G}_g, i \neq j} v_j |\hat{\beta}_i| + (1-\alpha) w_g \sqrt{p_g}, \quad \epsilon_g' = \gamma_g^{-1} (1-\alpha) w_g \sqrt{p_g}.$$

Using similar calculations as for SGL, the strong screening rules and KKT checks for aSGL are derived in Appendix B.2. Algorithm A1 is also applicable for aSGL, using the corresponding aSGL equations as replacement (Algorithm A2). The choice of adaptive weights is described in Appendix B.3.

B.2. Theory

B.2.1. DERIVATION OF THE CONNECTION TO ϵ-NORM

The aim is to link the aSGL norm (Equation 19) to the ϵ-norm, in a similar way to SGL:

$$\|\beta\|_{\text{sgl}} = \sum_{g=1}^{m} (\alpha + (1-\alpha)\sqrt{p_g}) \|\beta^{(g)}\|_{\epsilon_g}^*.$$

Splitting up the adaptive lasso term in Equation 19 yields

$$\begin{aligned}
\alpha \sum_{i=1}^{p} v_i |\beta_i| &= \alpha \sum_{g=1}^{m} \sum_{i \in \mathcal{G}_g} v_i |\beta_i| \\
&= \alpha \sum_{g=1}^{m} \left(\sum_{j \in \mathcal{G}_g} v_j \sum_{i \in \mathcal{G}_g} |\beta_i| - \sum_{i,j \in \mathcal{G}_g, i \neq j} v_j |\beta_i| \right) \\
&= \alpha \sum_{g=1}^{m} \left(\sum_{j \in \mathcal{G}_g} v_j \sum_{i \in \mathcal{G}_g} |\beta_i| - \frac{\sum_{i,j \in \mathcal{G}_g, i \neq j} v_j |\beta_i|}{\sum_{i \in \mathcal{G}_g} |\beta_i|} \sum_{i \in \mathcal{G}_g} |\beta_i| \right) \\
&= \alpha \sum_{g=1}^{m} \sum_{i \in \mathcal{G}_g} |\beta_i| \left(\sum_{j \in \mathcal{G}_g} v_j - \frac{\sum_{i,j \in \mathcal{G}_g, i \neq j} v_j |\beta_i|}{\sum_{i \in \mathcal{G}_g} |\beta_i|} \right) \\
&= \alpha \sum_{g=1}^{m} \|\beta^{(g)}\|_1 \left(\|v^{(g)}\|_1 - \frac{\sum_{i,j \in \mathcal{G}_g, i \neq j} v_j |\beta_i|}{\|\beta^{(g)}\|_1} \right).
\end{aligned}$$

Hence

$$\begin{aligned}\|\beta\|_{\text{asgl}} &= \alpha \sum_{i=1}^{p} v_i|\beta_i| + (1-\alpha)\sum_{g=1}^{m} w_g\sqrt{p_g}\|\beta^{(g)}\|_2 \\ &= \sum_{g=1}^{m}\left[\left(\|v^{(g)}\|_1 - \frac{\sum_{i,j\in\mathcal{G}_g, i\neq j} v_j|\beta_i|}{\|\beta^{(g)}\|_1}\right)\alpha\|\beta^{(g)}\|_1 + (1-\alpha)w_g\sqrt{p_g}\|\beta^{(g)}\|_2\right]. \end{aligned} \tag{21}$$

Further, setting

$$\gamma_g = \alpha\|v^{(g)}\|_1 - \frac{\alpha\sum_{i,j\in\mathcal{G}_g, i\neq j} v_j|\beta_i|}{\|\beta^{(g)}\|_1} + (1-\alpha)w_g\sqrt{p_g},$$

simplifies Equation 21 to

$$\|\beta\|_{\text{asgl}} = \sum_{g=1}^{m}\gamma_g\left[\left(\frac{\gamma_g - (1-\alpha)w_g\sqrt{p_g}}{\gamma_g}\right)\|\beta^{(g)}\|_1 + \left(\frac{(1-\alpha)w_g\sqrt{p_g}}{\gamma_g}\right)\|\beta^{(g)}\|_2\right]. \tag{22}$$

Further, setting

$$\epsilon_g' = \frac{(1-\alpha)w_g\sqrt{p_g}}{\gamma_g},$$

allows Equation 22 to be written in terms of the ϵ-norm

$$\|\beta\|_{\text{asgl}} = \sum_{g=1}^{m}\gamma_g\left[(1-\epsilon_g')\|\beta^{(g)}\|_1 + \epsilon_g'\|\beta^{(g)}\|_2\right] = \sum_{g=1}^{m}\gamma_g\|\beta^{(g)}\|_{\epsilon_g'}^{*}.$$

B.2.2. PROPERTIES OF THE DERIVATION TO THE ϵ-NORM

An important aspect to note is that under $\beta^{(g)} \equiv \mathbf{0}$ for a group g, the middle term in γ_g becomes

$$\lim_{\beta^{(g)}\to 0}\left(\frac{\alpha\sum_{i,j\in\mathcal{G}_g, i\neq j} v_j|\beta_i|}{\|\beta^{(g)}\|_1}\right) = \frac{\alpha(p_g-1)}{p_g}\sum_{i=1}^{p_g} v_i,$$

so that γ_g still exists. This can be observed by using L'Hôpital's rule and noting that for $i \in \mathcal{G}_g$,

$$\frac{\partial}{\partial\beta_i}\sum_{i\neq j} v_j|\beta_i| = \sum_{i\neq j} v_j, \quad \frac{\partial}{\partial\beta_i}\|\beta^{(g)}\|_1 = 1.$$

To see how this reduces to SGL under $v \equiv 1$ and $w \equiv 1$, note that

$$\begin{aligned}\gamma_g &= \alpha\left(p_g - \frac{\sum_{i,j\in\mathcal{G}_g, i\neq j} v_j|\beta_i|}{\|\beta^{(g)}\|_1}\right) + (1-\alpha)\sqrt{p_g} \\ &= \alpha\left(p_g - \frac{(p_g-1)\|\beta^{(g)}\|_1}{\|\beta^{(g)}\|_1}\right) + (1-\alpha)\sqrt{p_g} \\ &= \alpha + (1-\alpha)\sqrt{p_g} = \tau_g.\end{aligned}$$

To understand the cross summation term, note that for the summation we are summing over each β term $p_g - 1$ times, as the matching indices are removed, that is (for simplicity of notation, we consider $\mathcal{G}_1$ so that the indexing here is reset from 1)

$$\begin{aligned}\sum_{i,j\in\mathcal{G}_1, i\neq j} v_j|\beta_i| &= |\beta_1|v_2 + \ldots + |\beta_1|v_{p_1} + |\beta_2|v_1 + \ldots + |\beta_2|v_{p_1} + \ldots + |\beta_{p_1}|v_{p_1-1} \\ &= (p_1-1)|\beta_1| + \ldots + (p_1-1)|\beta_{p_1}|, \text{ by setting } v_j = 1, \forall j \in \mathcal{G}_1, \text{ for SGL} \\ &= (p_1-1)\sum_{i\in\mathcal{G}_1}|\beta_i| = (p_1-1)\|\beta^{(g)}\|_1.\end{aligned}$$

Hence, ϵ_g' also reduces to ϵ_g.

B.2.3. PATH START

To find the path start for adaptive SGL, the dual norm can not be used, as γ_g is undefined for $\beta \equiv 0$. A derivation can instead be found using a similar approach to that in Simon et al. [37], where the point is found by solving the piecewise quadratic, for each $g \in \mathcal{G}$

$$\left\|S\left(X^{(g)\top}y/n, \lambda_g v^{(g)}\alpha\right)\right\|_2^2 - p_g w_g^2(1-\alpha)^2\lambda_g^2 = 0.$$

Then, choosing $\lambda_1 = \max_g \lambda_g$ gives the path start point.

B.2.4. GROUP SCREENING

To derive the group screening rule for aSGL, we compare the formulations of SGL and aSGL in terms of the ϵ-norm (Equations 2 and 20):

$$\|\beta\|_{\text{sgl}} = \sum_{g=1}^m \tau_g \|\beta^{(g)}\|_{\epsilon_g}^*, \quad \|\beta\|_{\text{asgl}} = \sum_{g=1}^m \gamma_g \|\beta^{(g)}\|_{\epsilon_g'}^*. \tag{23}$$

Therefore, the derivation for the group screening rule for aSGL is identical to that of SGL (Section 2.2.1) replacing τ_g with γ_g and $\|\cdot\|_{\epsilon_g}$ with $\|\cdot\|_{\epsilon_g'}$. The group screening rule is given by: discard a group g if

$$\|\nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g'} \leq \gamma_g(2\lambda_{k+1} - \lambda_k), \tag{24}$$

and is formalized in Propositions 6 and 7.

Proposition 6 (Theoretical aSGL group screening) *For any $\lambda_{k+1}, k \in [l-1]$, the candidate set $\mathcal{C}_g(\lambda_{k+1}) = \{g \in [m] : \|\nabla_g f(\hat{\beta}(\lambda_{k+1}))\|_{\epsilon_g'} > \gamma_g\lambda_{k+1}\}$ recovers the exact support of the active groups for aSGL. That is, $\mathcal{C}_g(\lambda_{k+1}) = \mathcal{A}_g(\lambda_{k+1})$.*

Proof The proof is identical to that of Proposition 4 replacing τ_g with γ_g and $\|\cdot\|_{\epsilon_g}$ with $\|\cdot\|_{\epsilon_g'}$ (Appendix A.1.1). ■

Proposition 7 (DFR-aSGL group screening) *For any $\lambda_{k+1}, k \in [l-1]$, assuming that*

$$\|\nabla_g f(\hat{\beta}(\lambda_{k+1})) - \nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g'} \leq \gamma_g|\lambda_{k+1} - \lambda_k|,$$

the candidate set $\mathcal{C}_g(\lambda_{k+1}) = \{g \in [m] : \|\nabla_g f(\hat{\beta}(\lambda_k))\|_{\epsilon_g'} > \gamma_g(2\lambda_{k+1} - \lambda_k)\}$ is a superset of the set of active groups for aSGL. That is, $\mathcal{A}_g(\lambda_{k+1}) \subset \mathcal{C}_g(\lambda_{k+1})$.

Proof The proof is identical to that of Proposition 1 replacing τ_g with γ_g and $\|\cdot\|_{\epsilon_g}$ with $\|\cdot\|_{\epsilon'_g}$ (Appendix A.1.1). ■

B.2.5. VARIABLE SCREENING

The construction of the variable screening rule for aSGL is very similar to that of SGL (Section 2.2.2). The key difference is that the KKT stationary conditions for aSGL for an inactive variable in an active group are given by (in comparison to Equation 14 for SGL)

$$-\nabla_i f(\hat{\beta}(\lambda_{k+1})) \in \lambda_{k+1}\alpha v_i \Phi^0_{i,k+1}.$$

Therefore, the derivation of the rule is identical, replacing α with αv_i. The variable screening rule is given by: discard a variable i if

$$|\nabla_i f(\hat{\beta}(\lambda_k))| \leq \alpha v_i(2\lambda_{k+1} - \lambda_k), \tag{25}$$

and is formalized in Propositions 8 and 9.

Proposition 8 (Theoretical aSGL variable screening) *For any $\lambda_{k+1}, k \in [l-1]$, the candidate set $\mathcal{C}_v(\lambda_{k+1}) = \{i \in \mathcal{G}_g \text{ for } g \in \mathcal{A}_g(\lambda_{k+1}) : |\nabla_i f(\hat{\beta}(\lambda_{k+1}))| > \lambda_{k+1}\alpha v_i\}$ recovers the exact support of the active variables for aSGL, that is, $\mathcal{C}_v(\lambda_{k+1}) = \mathcal{A}_v(\lambda_{k+1})$.*

Proof The proof is identical to that of Proposition 5 replacing α with αv_i (Appendix A.1.2). ■

Proposition 9 (DFR-aSGL variable screening) *For any $\lambda_{k+1}, k \in [l-1]$, assuming that*

$$|\nabla_i f(\hat{\beta}(\lambda_{k+1})) - \nabla_i f(\hat{\beta}(\lambda_k))| \leq \alpha v_i(\lambda_k - \lambda_{k+1}),$$

the candidate set $\mathcal{C}_v(\lambda_{k+1}) = \{i \in \mathcal{G}_g \text{ for } g \in \mathcal{A}_g(\lambda_{k+1}) : |\nabla_i f(\hat{\beta}(\lambda_k))| > \alpha v_i(2\lambda_{k+1} - \lambda_k)\}$ is a superset of the set of active variables for aSGL, that is, $\mathcal{A}_v(\lambda_{k+1}) \subset \mathcal{C}_v(\lambda_{k+1})$.

Proof The proof is identical to that of Proposition 2 replacing α with αv_i (Appendix A.1.2). ■

B.2.6. KKT CHECKS

The KKT checks for aSGL are also similar to those for SGL (Appendix A.2): a KKT violation occurs for a variable $i \in \mathcal{G}_g$ if

$$|S(\nabla_i f(\hat{\beta}(\lambda_{k+1})), \lambda_{k+1}(1-\alpha)w_g\sqrt{p_g})| \leq \lambda_{k+1} v_i \alpha. \tag{26}$$

B.3. Choice of adaptive weights

The adaptive weights are chosen according to Mendez-Civieta et al. [24] as

$$v_i = \frac{1}{|q_{1i}|^{\gamma_1}}, w_g = \frac{1}{\|q_1^{(g)}\|_2^{\gamma_2}},$$

where q_1 is the first principal component from performing principal component analysis on $\mathbf{X}$ and γ_1, γ_2 are chosen by the user, often in the range $[0, 2]$. The weights are shown in Figure A1.

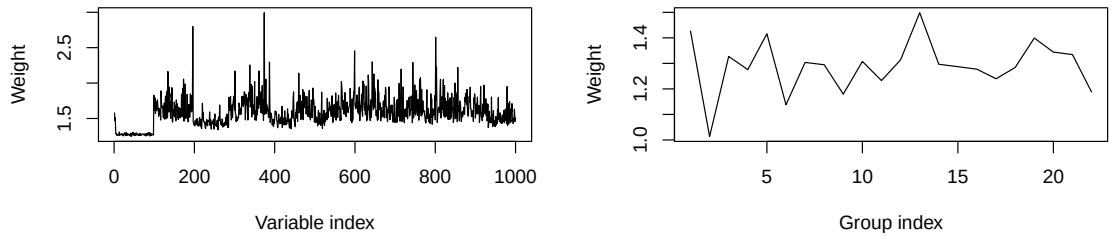

Figure A1: The aSGL weights, (v, w), for $p = 1000, n = 200, m = 22, \rho = 0.3, \gamma_1 = \gamma_2 = 0.1$, and $\alpha = 0.95$.

B.4. Algorithm

Algorithm A2 Dual Feature Reduction (DFR) for aSGL

Input: $(\lambda_1, \dots, \lambda_l) \in \mathbb{R}^l$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$, $\alpha \in [0, 1]$
compute $\hat{\beta}(\lambda_1)$ using Equation 1, replacing the SGL norm with Equation 19
for $k = 1$ **to** $l - 1$ **do**
  $\mathcal{C}_g(\lambda_{k+1}) \leftarrow$ candidate groups from Proposition 7
  $\mathcal{C}_v(\lambda_{k+1}) \leftarrow$ candidate variables from Proposition 9 for $i \in \mathcal{G}_g, g \in \mathcal{C}_g(\lambda_{k+1})$, and $i \notin \mathcal{A}_v(\lambda_k)$
  $\mathcal{O}_v \leftarrow \mathcal{C}_v(\lambda_{k+1}) \cup \mathcal{A}_v(\lambda_k)$ ▶ Optimization set
  compute $\hat{\beta}_i(\lambda_{k+1}), i \in \mathcal{O}_v$, using Equation 1, replacing the SGL norm with Equation 19
  $\mathcal{K}_v \leftarrow$ variable KKT violations for $i \notin \mathcal{O}_v$, using Equation 26 ▶ KKT check
  while $\text{card}(\mathcal{K}_v) > 0$ **do**
    $\mathcal{O}_v \leftarrow \mathcal{O}_v \cup \mathcal{K}_v$ ▶ Optimization set
    compute $\hat{\beta}_i(\lambda_{k+1}), i \in \mathcal{O}_v$, using Equation 1, replacing the SGL norm with Equation 19
    $\mathcal{K}_v \leftarrow$ variable KKT violations for $i \notin \mathcal{O}_v$, using Equation 26 ▶ KKT check
  end while
end for
Output: $\hat{\beta}_{\text{asgl}}(\lambda_1), \dots, \hat{\beta}_{\text{asgl}}(\lambda_l) \in \mathbb{R}^p$

Appendix C. Competitive feature reduction approaches

Table A1: A summary of the four screening rules for SGL considered in this manuscript.

	Rules (discard if true)	
Method	Variable	Group
DFR-aSGL	$\lvert\nabla_i f(\hat{\beta}(\lambda_k))\rvert \leq \alpha v_i(2\lambda_{k+1} - \lambda_k)$	$\lVert\nabla_g f(\hat{\beta}(\lambda_k))\rVert_{\epsilon'_g} \leq \gamma_g(2\lambda_{k+1} - \lambda_k)$
DFR-SGL	$\lvert\nabla_i f(\hat{\beta}(\lambda_k))\rvert \leq \alpha(2\lambda_{k+1} - \lambda_k)$	$\lVert\nabla_g f(\hat{\beta}(\lambda_k))\rVert_{\epsilon_g} \leq \tau_g(2\lambda_{k+1} - \lambda_k)$
sparsegl	-	$\lVert S(\nabla_g f(\hat{\beta}(\lambda_k)), \lambda_k\alpha)\rVert_2 \leq w_g(1-\alpha)(2\lambda_{k+1} - \lambda_k)$
GAP safe	$\lvert X_i^\top \Theta_c\rvert + r\lVert X_i\rVert_2 < \tau$	$\mathcal{T}_g < (1-\alpha)\sqrt{p_g}$

C.1. Sparsegl

Sparsegl is a screening rule proposed by Liang et al. [23] and performs a single layer of group screening. The rule is based on the strong screening framework [40] and the first order condition derived in Simon et al. [37], i.e. that a group g is inactive if

$$\lVert S(\nabla_g f(\hat{\beta}(\lambda_{k+1})), \lambda_{k+1}\alpha)\rVert_2 \leq \sqrt{p_g}(1-\alpha)\lambda_{k+1}. \tag{27}$$

As the gradient at $k+1$ is not available, the following Lipschitz assumption on the ℓ_2 norm is used:

$$\lVert S(\nabla_g f(\hat{\beta}(\lambda_{k+1})), \lambda_{k+1}\alpha) - S(\nabla_g f(\hat{\beta}(\lambda_k)), \lambda_k\alpha)\rVert_2 \leq w_g(1-\alpha)\lvert\lambda_{k+1} - \lambda_k\rvert. \tag{28}$$

This leads to the sparsegl screening rule (via the triangle inequality): discard a group g if

$$\lVert S(\nabla_g f(\hat{\beta}(\lambda_k)), \lambda_k\alpha)\rVert_2 \leq w_g(1-\alpha)(2\lambda_{k+1} - \lambda_k). \tag{29}$$

This screening rule uses a different Lipschitz assumption (Equation 28), which leads to a different group-level rule. The DFR group Lipschitz assumption (Equation 10) is more consistent with the work of Tibshirani et al. [40], as the assumption is with regards to the dual norm of the full SGL norm, rather than just the group lasso component.

C.2. GAP safe

An exact feature reduction method for SGL was proposed in Ndiaye et al. [27] under linear regression. The approach makes use of the sub-differential inclusion equation of Fermat's rule [2]:

$$\mathbf{X}^\top \hat{\Theta}^{(\lambda, \lVert\cdot\rVert_{\text{sgl}})} \in \partial \lVert\cdot\rVert_{\text{sgl}}(\hat{\beta}^{(\lambda, \lVert\cdot\rVert_{\text{sgl}})}), \tag{30}$$

where $\hat{\Theta}$ is the solution to the dual formulation of Equation 1. Using this, exact (theoretical) rules are derived to determine which variables and groups are inactive at the optimal solution. The rules are theoretical as they rely on $\hat{\Theta}^{\lambda, \lVert\cdot\rVert_{\text{sgl}}}$, which is not available in practice. Instead, a safe region is constructed that contains the optimal dual solution; in Ndiaye et al. [27] it is taken as a sphere, but other regions can also be used (such as domes). Due to these safe regions, the reduction is generally more conservative.

The safe sphere is defined as $B(\Theta_c, r)$ with center Θ_c and radius r. An ideal region would be such that r is small and the center is close to $\hat{\Theta}^{\lambda,\|\cdot\|_{\text{sgl}}}$. Using this safe region, the GAP safe rules at λ_{k+1} are derived as, for a variable i and group g,

$$\text{Variable screening: } |X_i^\top \Theta_c| + r\|X_i\|_2 < \tau \implies \hat{\beta}_i(\lambda_{k+1}) = 0, \tag{31}$$

$$\text{Group screening: } \mathcal{T}_g < (1-\alpha)\sqrt{p_g} \implies \hat{\beta}^{(g)}(\lambda_{k+1}) = 0, \tag{32}$$

where

$$\mathcal{T}_g = \begin{cases} \|S(X_g^\top \Theta_c, \alpha)\| + r\|X_g\|, & \text{if } \|X_g^\top \Theta_c\|_\infty > \alpha, \\ (\|X_g^\top \Theta_c\|_\infty + r\|X_g\| - \alpha)_+, & \text{otherwise.} \end{cases} \tag{33}$$

The center Θ_c and the radius r are derived using the duality gap and are calculated at iteration t in an iterative algorithm as

$$\Theta_t(\beta_{(t)}) = \frac{y - \mathbf{X}\beta_{(t)}}{\max(\lambda_{k+1}, \|X^\top (y - \mathbf{X}\beta_{(t)})\|^*_{\text{sgl}})}, \quad r_t(\beta_{(t)}, \Theta_t) = \sqrt{\frac{2P_{\lambda_{k+1},\alpha}(\beta_{(t)}) - D_{\lambda_{k+1}}(\Theta_t)}{\lambda_{k+1}^2}}, \tag{34}$$

where $P_{\lambda,\alpha}$ and D_λ are the primal and dual objectives, and $\beta_{(t)}$ is the primal value at iteration t. The radius and center are expensive to evaluate, so are calculated every 10 iterations [27].

The above formulation combines both dynamic and sequential screening. The method can also be implemented using just sequential screening, in which the primal values used in the calculation of the center and radius are from λ_k.

For both the GAP safe rules and DFR, theoretically it would be possible to exactly identify the active sets, but both instead require approximations. While GAP safe has different implementations, we present the best-performing versions in our studies.

Appendix D. Synthetic data analysis

This section complements Section 3.1 by providing further information about the simulation set-up and additional results generated for the synthetic data. Additional tables and figures are provided that further showcase the effectiveness of DFR. In particular, results from applying screening to interaction detection (Appendix D.3.4) and cross-validation (Appendix D.3.5) are presented. All synthetic results, from the main text and the appendix, are repeated using a logistic model (Appendix D.4).

D.1. Metrics

The following metrics are shown in the tables in the appendix:

- $\mathcal{A}_v, \mathcal{A}_g$: the number of active variables/groups.
- $\mathcal{C}_v, \mathcal{C}_g$: the number of variables/groups in the candidate sets.
- $\mathcal{O}_c, \mathcal{O}_g$: the number of variables/groups used in the optimization process.
- $\mathcal{K}_c, \mathcal{K}_g$: the number of variable/group KKT violations. DFR only checks for group violations and sparsegl only checks for variable violations.
- $\mathcal{O}_v \,/\, \mathcal{A}_v$ and $\mathcal{O}_g \,/\, \mathcal{A}_g$: the proportion of variables/groups used in the optimization against the number active. Defines how efficient the rules are. A low number is preferred.
- $\mathcal{O}_v \,/\, p$ and $\mathcal{O}_g \,/\, m$: the variable/group input proportion, as defined in Section 3.
- ℓ_2 distance to no screen: ℓ_2 from the fitted values obtained with screening to without.

D.2. Set up

Table A2: Default model, data, and algorithm parameters for the synthetic and real data analyses.

Category	Parameter	Values	
		Synthetic	**Real**
Model			
	α	0.95	0.95
	$\gamma_1 = \gamma_2$ (aSGL only)	0.1	0.1
	Path length (l)	50	100
	Path termination (λ_l)	$0.1\lambda_1$	$0.2\lambda_1$
	Path shape	Log-linear	Log-linear
Data			
	p	1000	-
	n	200	-
	m	22	-
	Group sizes	$[3, 100]$	-
	β	$\mathcal{N}(0, 4)$	-
	Variable sparsity	0.2	-
	Group sparsity	0.2	-
	ρ	0.3	-
	ϵ	$\mathcal{N}(0, 1)$	-
Algorithm (ATOS)			
	Max iterations	5000	10000
	Backtracking	0.7	0.7
	Max backtracking iterations	100	100
	Convergence tol	10^{-5}	10^{-5}
	Standardization	ℓ_2	ℓ_2
	Intercept	Yes for linear	Yes for linear
	Warm starts	Yes	Yes

D.3. Additional results and commentary

D.3.1. KKT VIOLATIONS

KKT violations for DFR are very rare. Across all experiments, DFR-SGL had only a single KKT violation (Table A12). Violations were more common for DFR-aSGL and sparsegl, but were still rare. In the experiment in which DFR-SGL had its only violation (Figure 2 (below)), DFR-aSGL had a violation every 1739 fits, and sparsegl had one every 53900 fits. Note that sparsegl violations refer to group violations and DFR-aSGL to variable ones. It is more likely for there to be a variable violation, given that $p \geq m$. In some instances, sparsegl demonstrates more efficient group-level screening (Table A8), while on other occasions DFR-SGL is more efficient (Table A11). However, the elevated number of KKT violations for sparsegl suggests that the Lipschitz assumption of DFR is more robust.

D.3.2. COMPARISON TO GAP SAFE

In Figure 1, a spike is observed around $p = 800$. In this case, the groups were fixed at size 20, and so for each value of p the group sizes, as a proportion of the input, are different. It is possible that $p = 800$ represents a *sweet spot* where the grouping structure favors bi-level screening. For small p, the grouping structure would dominate, so that discarding one of the few discrete chunks of groups would have a large influence on computational savings. On the other hand, as p becomes large, the grouping structure becomes less important, as there are many small groups.

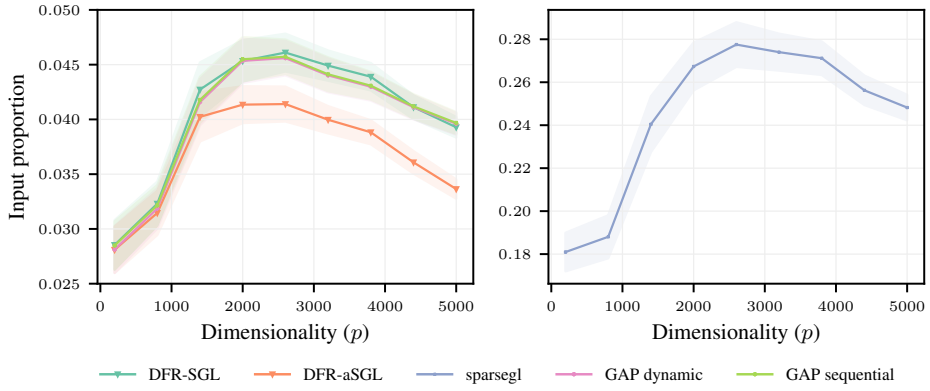

Figure A2: The input proportion for strong against safe rules, applied to synthetic data, as a function of the dimensionality (p), with 95% confidence intervals. sparsegl has been separated into the right plot, using a different y-scale, so that the narrow differences between the other methods can be observed.

D.3.3. ROBUSTNESS

A clear benefit of DFR over sparsegl is observed under very sparse signals (Figures A3 and A4). It is clear that screening rules have an increasing impact as the signal becomes sparser. However, when the signal saturates, screening approaches perform similarly, as their effectiveness is reduced. DFR is further found to be relatively unaffected by the strength of the signal and provides a benefit regardless of the strength (Figures A3 and A4).

DFR-aSGL was also found to be robust under different values of γ_1 and γ_2 (Figure A6).

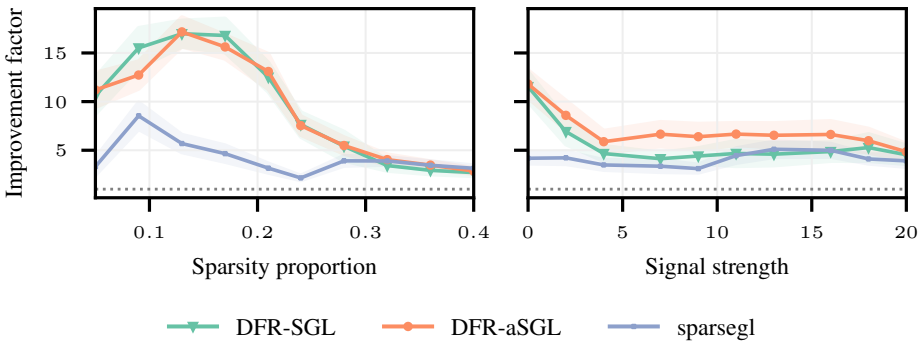

Figure A3: The improvement factor for the screening methods applied to synthetic data, as a function of the data sparsity proportion (left) and signal strength (right), with 95% confidence intervals.

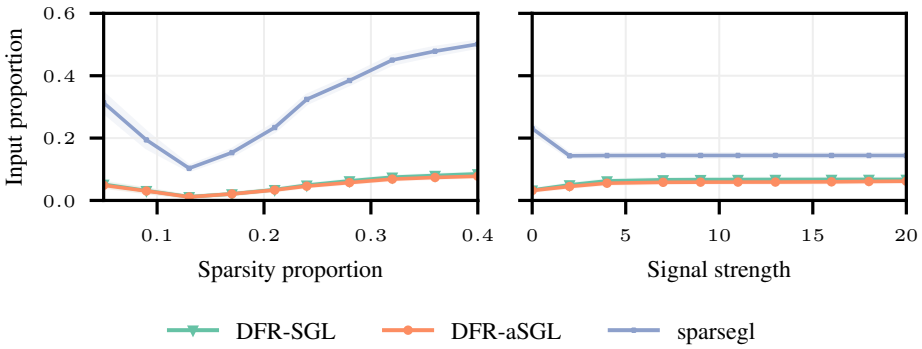

Figure A4: The input proportion for the screening methods applied to synthetic data, under the linear model, as a function of the data sparsity proportion (left) and signal strength (right), with 95% confidence intervals.

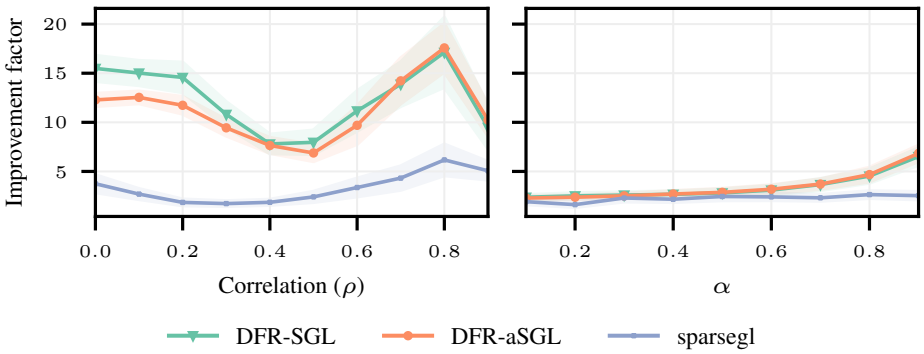

Figure A5: The improvement factor for the screening methods applied to synthetic data, under the linear model, as a function of the data correlation (left) and α (right), with 95% confidence intervals.

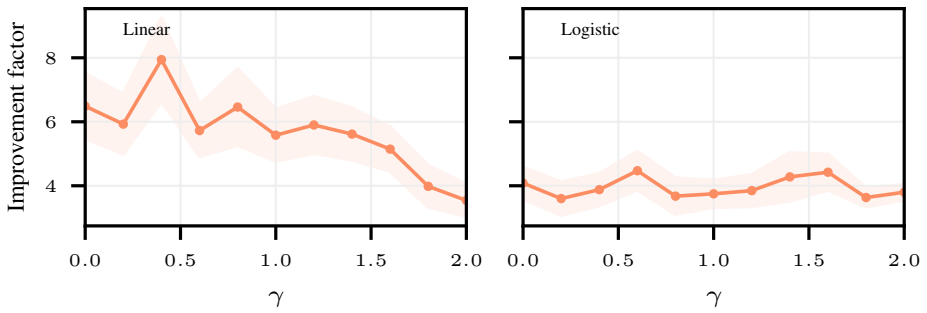

Figure A6: Robustness of DFR-aSGL under different $\gamma_1 = \gamma_2$ values for the weights, shown for linear (left) and logistic (right) models, with 95% confidence intervals. The data was generated using the parameters in Table A2.

D.3.4. INTERACTION DETECTION

The setup of the interaction data is as described in Section 3.1 for the marginal effects. For the interactions, within each group, all possible interactions of order 2 and 3 were generated. DFR is able to provide large computational savings when fitting interactions, especially compared to sparsegl, which under order 3 interactions provides only a marginal benefit (Table A3 and Figure A7). These savings make it more feasible for sparse-group models to be used in interaction detection problems. Such challenges are frequently seen in the field of genetics, where gene-gene and gene-environment relationships are useful discoveries [6, 47].

Table A3: The improvement factor for the strong rules applied to synthetic interaction data, with standard errors. The parameters of the synthetic data were set as $p = 400, n = 80$, and $m = 52$ groups of sizes in $[3, 15]$. The interaction input dimensionality was $p_{O_2} = 2111, p_{O_3} = 7338$ for orders 2 and 3, with no interaction hierarchy imposed. An active proportion of 0.3 was used (using the same signal as the marginal effects).

	Interaction	
Method	Order 2	Order 3
DFR-aSGL	137.3 ± 12.0	54.0 ± 10.7
DFR-SGL	44.3 ± 2.4	23.6 ± 3.1
sparsegl	7.4 ± 0.9	1.2 ± 0.3

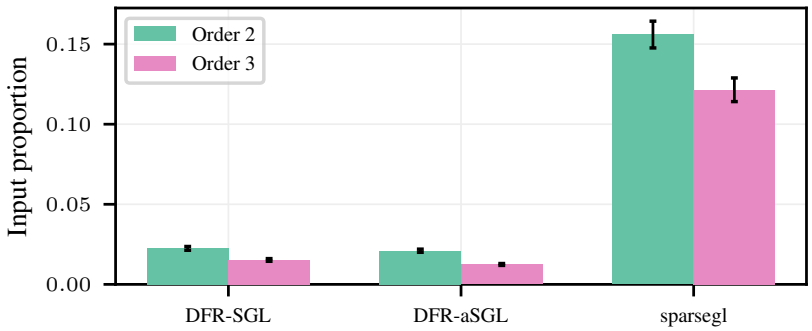

Figure A7: The input proportion for the strong rules applied to synthetic interaction data, under the linear model, with standard errors. The parameters of the synthetic data were set as $p = 400, n = 80$, and $m = 52$ groups of sizes in [3, 15]. The interaction input dimensionality was $p_{O_2} = 2111, p_{O_3} = 7338$ for orders 2 and 3, with no interaction hierarchy imposed.

D.3.5. CROSS-VALIDATION

Cross-validation (CV) is an important tool for tuning λ. However, due to its cost, α is often set manually, rather than included in a grid optimization scheme. Using DFR with 10-fold CV yielded computationally gains (Table A4) that enable future tuning schemes for SGL to consider both α and λ, and aSGL to include the weight hyperparameters γ_1, γ_2.

Table A4: The improvement factor for the screening methods applied to synthetic data, under the linear and logistic models, with cross-validation (CV), with standard errors. The data was generated using the parameters in Table A2.

Method	Linear	Logistic
DFR-aSGL	3.9 ± 0.2	2.3 ± 0.1
DFR-SGL	4.2 ± 0.3	2.6 ± 0.1
sparsegl	2.0 ± 0.2	2.1 ± 0.1

D.3.6. RESULTS TABLES FOR THE LINEAR MODEL

Table A5: Group screening metrics corresponding to the GAP safe comparison simulation (Figures 1 and A2), averaged over all simulation iterations, dimensionality (p) cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g \,/\, \mathcal{A}_g$	$\mathcal{O}_g \,/\, m$
DFR-ASGL	28.61 ± 0.13	36.92 ± 0.17	36.88 ± 0.17	–	1.2488 ± 0.0011	$0.1475 \pm 7 \times 10^{-4}$
DFR-SGL	28.25 ± 0.13	38.95 ± 0.18	38.95 ± 0.18	–	1.3353 ± 0.0015	$0.1558 \pm 7 \times 10^{-4}$
SPARSEGL	28.25 ± 0.13	33.64 ± 0.16	33.64 ± 0.16	$2 \times 10^{-5} \pm 2 \times 10^{-5}$	$1.142 \pm 7 \times 10^{-4}$	$0.1346 \pm 6 \times 10^{-4}$
GAP SEQUENTIAL	28.70 ± 0.13	39.00 ± 0.18	39.00 ± 0.18	–	1.3177 ± 0.0014	$0.1560 \pm 7 \times 10^{-4}$
GAP DYNAMIC	28.70 ± 0.13	39.00 ± 0.18	39.00 ± 0.18	–	1.3177 ± 0.0014	$0.1560 \pm 7 \times 10^{-4}$

Table A6: Variable screening metrics corresponding to the GAP safe comparison simulation (Figures 1 and A2), averaged over all simulation iterations, dimensionality (p) cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v \,/\, \mathcal{A}_v$	$\mathcal{O}_v \,/\, p$
DFR-ASGL	64.18 ± 0.30	36.31 ± 0.18	97.76 ± 0.47	0.1251 ± 0.0017	1.4624 ± 0.0017	$0.0196 \pm 9 \times 10^{-5}$
DFR-SGL	68.29 ± 0.32	44.70 ± 0.22	109.98 ± 0.52	0 ± 0	1.5514 ± 0.0022	$0.0220 \pm 1 \times 10^{-4}$
SPARSEGL	68.29 ± 0.32	672.82 ± 3.21	672.82 ± 3.21	–	10.1974 ± 0.0174	$0.1346 \pm 6 \times 10^{-4}$
GAP SEQUENTIAL	69.84 ± 0.33	109.36 ± 0.52	109.36 ± 0.52	–	1.5101 ± 0.0019	$0.0219 \pm 1 \times 10^{-4}$
GAP DYNAMIC	69.84 ± 0.33	109.07 ± 0.52	109.07 ± 0.52	–	1.4786 ± 0.0018	$0.0218 \pm 1 \times 10^{-4}$

Table A7: Model fitting metrics corresponding to the GAP safe comparison simulation (Figures 1 and A2), averaged over all simulation iterations, dimensionality (p) cases, and path points, shown with standard errors.

	TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	136.53 ± 4.53	19.07 ± 0.57	7.99 ± 0.11	404.82 ± 5.85	292.99 ± 6.89	$1 \times 10^{-7} \pm 3 \times 10^{-9}$	0 ± 0	0 ± 0
DFR-SGL	129.84 ± 4.2	17.86 ± 0.52	8.44 ± 0.13	395.27 ± 5.57	309.99 ± 7.05	$1 \times 10^{-10} \pm 7 \times 10^{-12}$	0 ± 0	0 ± 0
SPARSEGL	129.84 ± 4.2	34.13 ± 1.04	3.86 ± 0.05	395.27 ± 5.57	387.65 ± 5.90	$9 \times 10^{-11} \pm 6 \times 10^{-12}$	0 ± 0	0 ± 0
GAP SEQUENTIAL	0.77 ± 0.02	0.77 ± 0.02	1.01 ± 0.01	–	–	0 ± 0	–	–
GAP DYNAMIC	0.77 ± 0.02	0.80 ± 0.03	0.99 ± 0.02	–	–	$3 \times 10^{-18} \pm 2 \times 10^{-18}$	–	–

Table A8: Group screening metrics corresponding to the correlation simulation (Figures 2 and A5 (left)), averaged over all simulation iterations, correlation cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g$ / $\mathcal{A}_g$	$\mathcal{O}_g$ / m
DFR-ASGL	3.79 ± 0.01	4.31 ± 0.02	4.31 ± 0.02	–	$1.1085 \pm 1 \times 10^{-3}$	$0.1958 \pm 7 \times 10^{-4}$
DFR-SGL	3.67 ± 0.01	4.31 ± 0.02	4.31 ± 0.02	–	1.1356 ± 0.0012	$0.1959 \pm 7 \times 10^{-4}$
SPARSEGL	3.67 ± 0.01	3.98 ± 0.01	3.98 ± 0.01	$1 \times 10^{-4} \pm 5 \times 10^{-5}$	$1.0571 \pm 7 \times 10^{-4}$	$0.1809 \pm 7 \times 10^{-4}$

Table A9: Variable screening metrics corresponding to the correlation simulation (Figures 2 and A5 (left)), averaged over all simulation iterations, correlation cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v$ / $\mathcal{A}_v$	$\mathcal{O}_v$ / p
DFR-ASGL	26.37 ± 0.09	17.39 ± 0.08	42.92 ± 0.16	$0.0088 \pm 4 \times 10^{-4}$	1.6214 ± 0.0026	$0.0429 \pm 2 \times 10^{-4}$
DFR-SGL	29.36 ± 0.11	18.25 ± 0.08	46.77 ± 0.17	0 ± 0	1.6025 ± 0.003	$0.0468 \pm 2 \times 10^{-4}$
SPARSEGL	29.36 ± 0.11	194.56 ± 0.78	194.56 ± 0.78	–	9.033 ± 0.0325	$0.1946 \pm 8 \times 10^{-4}$

Table A10: Model fitting metrics corresponding to the correlation simulation (Figures 2 and A5 (left)), averaged over all simulation iterations, correlation cases, and path points, shown with standard errors.

	TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	251.4 ± 7.9	29.4 ± 1.2	11.2 ± 0.3	825.6 ± 32.0	300.6 ± 14.6	$1 \times 10^{-8} \pm 2 \times 10^{-9}$	0 ± 0	0 ± 0
DFR-SGL	268.1 ± 10.0	35.4 ± 1.8	12.3 ± 0.3	801.6 ± 30.5	334.1 ± 16.0	$5 \times 10^{-10} \pm 4 \times 10^{-11}$	0 ± 0	0 ± 0
SPARSEGL	268.1 ± 10.0	104.8 ± 3.1	3.3 ± 0.1	801.6 ± 30.5	624.9 ± 25.8	$1 \times 10^{-10} \pm 9 \times 10^{-12}$	0 ± 0	0 ± 0

Table A11: Group screening metrics corresponding to the α simulation (Figures 2 and A5 (right)), averaged over all simulation iterations, α cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g \,/\, \mathcal{A}_g$	$\mathcal{O}_g \,/\, m$
DFR-ASGL	3.40 ± 0.01	4.61 ± 0.02	4.61 ± 0.02	–	1.1680 ± 0.0012	$0.2094 \pm 8 \times 10^{-4}$
DFR-SGL	3.99 ± 0.01	5.37 ± 0.02	5.37 ± 0.02	–	1.1819 ± 0.0012	$0.2441 \pm 9 \times 10^{-4}$
SPARSEGL	3.99 ± 0.01	6.64 ± 0.03	6.64 ± 0.03	$2 \times 10^{-5} \pm 2 \times 10^{-5}$	1.7826 ± 0.0123	0.302 ± 0.0013

Table A12: Variable screening metrics corresponding to the α simulation (Figures 2 and A5 (right)), averaged over all simulation iterations, α cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v \,/\, \mathcal{A}_v$	$\mathcal{O}_v \,/\, p$
DFR-ASGL	91.66 ± 0.44	61.06 ± 0.64	149.44 ± 0.80	$6 \times 10^{-4} \pm 1 \times 10^{-4}$	1.3676 ± 0.0025	$0.1494 \pm 8 \times 10^{-4}$
DFR-SGL	96.34 ± 0.44	61.15 ± 0.62	154.12 ± 0.80	$2 \times 10^{-5} \pm 2 \times 10^{-5}$	1.3546 ± 0.0022	$0.1541 \pm 8 \times 10^{-4}$
SPARSEGL	96.42 ± 0.44	305.94 ± 1.29	305.94 ± 1.29	–	20.3134 ± 0.4232	0.3059 ± 0.0013

Table A13: Model fitting metrics corresponding to the α simulation (Figures 2 and A5 (right)), averaged over all simulation iterations, α cases, and path points, shown with standard errors.

	TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	87.8 ± 1.3	33.8 ± 0.8	3.9 ± 0.1	179.6 ± 2.6	144.8 ± 2.7	$5 \times 10^{-4} \pm 4 \times 10^{-6}$	0 ± 0	0 ± 0
DFR-SGL	99.4 ± 1.1	35.5 ± 0.8	5.3 ± 0.2	213.6 ± 2.9	153.9 ± 2.5	$4 \times 10^{-4} \pm 3 \times 10^{-6}$	0 ± 0	0 ± 0
SPARSEGL	99.4 ± 1.1	75.7 ± 2.3	2.5 ± 0.1	213.6 ± 2.9	195.3 ± 3.2	$4 \times 10^{-4} \pm 3 \times 10^{-6}$	0 ± 0	0 ± 0

Table A14: Group screening metrics corresponding to the sparsity proportion simulation (Figures A3 and A4 (left)), averaged over all simulation iterations, sparsity proportion cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g \,/\, \mathcal{A}_g$	$\mathcal{O}_g \,/\, m$
DFR-ASGL	5.91 ± 0.03	6.62 ± 0.03	6.63 ± 0.03	–	1.1187 ± 0.0011	0.3013 ± 0.0013
DFR-SGL	5.94 ± 0.03	6.84 ± 0.03	6.84 ± 0.03	–	1.1464 ± 0.0012	0.3107 ± 0.0013
SPARSEGL	5.94 ± 0.03	6.37 ± 0.03	6.37 ± 0.03	$8 \times 10^{-5} \pm 4 \times 10^{-5}$	$1.0607 \pm 7 \times 10^{-4}$	0.2898 ± 0.0013

Table A15: Variable screening metrics corresponding to the sparsity proportion simulation (Figures A3 and A4 (left)), averaged over all simulation iterations, sparsity proportion cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v \,/\, \mathcal{A}_v$	$\mathcal{O}_v \,/\, p$
DFR-ASGL	34.62 ± 0.18	13.69 ± 0.06	46.71 ± 0.23	$0.0162 \pm 6 \times 10^{-4}$	1.3806 ± 0.002	$0.0467 \pm 2 \times 10^{-4}$
DFR-SGL	36.51 ± 0.19	15.3 ± 0.07	50.16 ± 0.25	0 ± 0	1.4017 ± 0.0022	$0.0502 \pm 2 \times 10^{-4}$
SPARSEGL	36.52 ± 0.19	313.66 ± 1.39	313.66 ± 1.39	–	12.2926 ± 0.0393	0.3137 ± 0.0014

Table A16: Model fitting metrics corresponding to the sparsity proportion simulation (Figures A3 and A4 (left)), averaged over all simulation iterations, sparsity proportion cases, and path points, shown with standard errors.

	TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	424.8 ± 4.6	98.0 ± 2.9	9.3 ± 0.3	248.6 ± 2.6	118.4 ± 3.2	$5 \times 10^{-8} \pm 6 \times 10^{-9}$	0 ± 0	0 ± 0
DFR-SGL	407.5 ± 4.7	106.9 ± 3.2	9.4 ± 0.3	243.1 ± 2.5	132.4 ± 3.4	$4 \times 10^{-10} \pm 2 \times 10^{-11}$	0 ± 0	0 ± 0
SPARSEGL	407.5 ± 4.7	162.0 ± 4.0	4.2 ± 0.1	243.1 ± 2.5	230.9 ± 3.0	$2 \times 10^{-10} \pm 1 \times 10^{-11}$	0 ± 0	0 ± 0

Table A17: Group screening metrics corresponding to the signal strength simulation (Figures A3 and A4 (right)), averaged over all simulation iterations, signal strength cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g$ / $\mathcal{A}_g$	$\mathcal{O}_g$ / m
DFR-ASGL	2.91 ± 0.01	3.04 ± 0.01	3.05 ± 0.01	–	$1.0437 \pm 7 \times 10^{-4}$	$0.1384 \pm 4 \times 10^{-4}$
DFR-SGL	2.94 ± 0.01	3.12 ± 0.01	3.12 ± 0.01	–	$1.0667 \pm 1 \times 10^{-3}$	$0.1417 \pm 4 \times 10^{-4}$
SPARSEGL	2.94 ± 0.01	3.01 ± 0.01	3.01 ± 0.01	$1 \times 10^{-4} \pm 5 \times 10^{-5}$	$1.0221 \pm 5 \times 10^{-4}$	$0.1369 \pm 3 \times 10^{-4}$

Table A18: Variable screening metrics corresponding to the signal strength simulation (Figures A3 and A4 (right)), averaged over all simulation iterations, signal strength cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v$ / $\mathcal{A}_v$	$\mathcal{O}_v$ / p
DFR-ASGL	38.52 ± 0.1	17.46 ± 0.04	54.86 ± 0.13	$0.0045 \pm 3 \times 10^{-4}$	1.4829 ± 0.0014	$0.0549 \pm 1 \times 10^{-4}$
DFR-SGL	42.35 ± 0.11	20.3 ± 0.05	61.48 ± 0.14	0 ± 0	1.504 ± 0.0015	$0.0615 \pm 1 \times 10^{-4}$
SPARSEGL	42.35 ± 0.11	152.75 ± 0.4	152.75 ± 0.4	–	4.4183 ± 0.016	$0.1528 \pm 4 \times 10^{-4}$

Table A19: Model fitting metrics corresponding to the signal strength simulation (Figures A3 and A4 (right)), averaged over all simulation iterations, signal strength cases, and path points, shown with standard errors.

	TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	444.2 ± 4.8	113.6 ± 2.4	7.0 ± 0.2	292.8 ± 1.9	268.2 ± 3.1	$4 \times 10^{-6} \pm 3 \times 10^{-7}$	0 ± 0	0 ± 0
DFR-SGL	181.3 ± 2.1	58.1 ± 1.2	5.6 ± 0.2	276.8 ± 1.4	255.6 ± 2.6	$5 \times 10^{-11} \pm 7 \times 10^{-12}$	0 ± 0	0 ± 0
SPARSEGL	181.3 ± 2.1	67.7 ± 1.4	4.1 ± 0.1	276.8 ± 1.4	275.6 ± 1.5	$4 \times 10^{-11} \pm 6 \times 10^{-12}$	0 ± 0	0 ± 0

Table A20: Group screening metrics corresponding to the interaction simulation (Table A3 and Figure A7), averaged all simulation iterations and path points, shown with standard errors.

		CARDINALITY				INPUT PROPORTION	
METHOD	TYPE	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g$ / $\mathcal{A}_g$	$\mathcal{O}_g$ / m
DFR-ASGL	ORDER 2	11.3 ± 0.1	14.9 ± 0.2	14.9 ± 0.2	–	1.291 ± 0.004	0.143 ± 0.002
DFR-SGL	ORDER 2	10.9 ± 0.1	15.7 ± 0.2	15.7 ± 0.2	–	1.391 ± 0.005	0.151 ± 0.002
SPARSEGL	ORDER 2	10.9 ± 0.1	13.2 ± 0.2	13.2 ± 0.2	0 ± 0	1.166 ± 0.003	0.127 ± 0.002
DFR-ASGL	ORDER 3	7.9 ± 0.1	11.0 ± 0.2	10.9 ± 0.2	–	1.340 ± 0.005	0.105 ± 0.002
DFR-SGL	ORDER 3	8.1 ± 0.1	12.4 ± 0.2	12.4 ± 0.2	–	1.468 ± 0.006	0.119 ± 0.002
SPARSEGL	ORDER 3	8.1 ± 0.1	10.2 ± 0.2	10.2 ± 0.2	$4 \times 10^{-4} \pm 3 \times 10^{-4}$	1.203 ± 0.003	0.098 ± 0.001

Table A21: Variable screening metrics corresponding to the interaction simulation (Table A3 and Figure A7), averaged all simulation iterations and path points, shown with standard errors.

		CARDINALITY				INPUT PROPORTION	
METHOD	TYPE	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v$ / $\mathcal{A}_v$	$\mathcal{O}_v$ / p
DFR-ASGL	ORDER 2	26.5 ± 0.4	19.1 ± 0.3	44.4 ± 0.6	0.071 ± 0.004	1.617 ± 0.007	0.021 ± 0.000
DFR-SGL	ORDER 2	25.9 ± 0.4	22.8 ± 0.3	47.5 ± 0.7	0 ± 0	1.739 ± 0.009	0.023 ± 0.000
SPARSEGL	ORDER 2	25.9 ± 0.4	329.2 ± 5.0	329.2 ± 5.0	–	11.771 ± 0.097	0.156 ± 0.002
DFR-ASGL	ORDER 3	41.8 ± 0.7	51.8 ± 0.8	91.2 ± 1.5	0.047 ± 0.003	2.278 ± 0.038	0.012 ± 0.000
DFR-SGL	ORDER 3	46.2 ± 0.8	68.4 ± 1.2	111.9 ± 1.9	0 ± 0	2.712 ± 0.052	0.015 ± 0.000
SPARSEGL	ORDER 3	46.2 ± 0.8	891.5 ± 16.5	891.5 ± 16.5	–	17.886 ± 0.295	0.122 ± 0.002

Table A22: Model fitting metrics corresponding to the interaction simulation (Table A3 and Figure A7), averaged all simulation iterations and path points, shown with standard errors.

		TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	TYPE	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	ORDER 2	12041 ± 411	174 ± 13	137 ± 12	1333 ± 39	137 ± 6	$2 \times 10^{-8} \pm 2 \times 10^{-9}$	0 ± 0	0 ± 0
DFR-SGL	ORDER 2	987 ± 48	32 ± 3	44 ± 2	1271 ± 38	140 ± 6	$7 \times 10^{-11} \pm 1 \times 10^{-11}$	0 ± 0	0 ± 0
SPARSEGL	ORDER 2	987 ± 48	319 ± 29	7 ± 1	1271 ± 38	658 ± 53	$2 \times 10^{-11} \pm 1 \times 10^{-11}$	0 ± 0	0 ± 0
DFR-ASGL	ORDER 3	12209 ± 491	877 ± 65	54 ± 11	1003 ± 33	615 ± 30	$1 \times 10^{-8} \pm 2 \times 10^{-9}$	0 ± 0	0 ± 0
DFR-SGL	ORDER 3	1265 ± 52	103 ± 7	24 ± 3	1089 ± 49	677 ± 32	$5 \times 10^{-11} \pm 6 \times 10^{-12}$	0 ± 0	0 ± 0
SPARSEGL	ORDER 3	1265 ± 52	2381 ± 170	1 ± 0.3	1089 ± 49	782 ± 33	$3 \times 10^{-11} \pm 4 \times 10^{-12}$	0 ± 0	0 ± 0

D.4. Results for the logistic model

The data input components $\mathbf{X}$, β, and ϵ for the logistic model were generated as for the linear models. The class probabilities for the response were calculated using $\sigma(\mathbf{X}\beta + \epsilon)$, where σ is the sigmoid function.

D.4.1. ROBUSTNESS

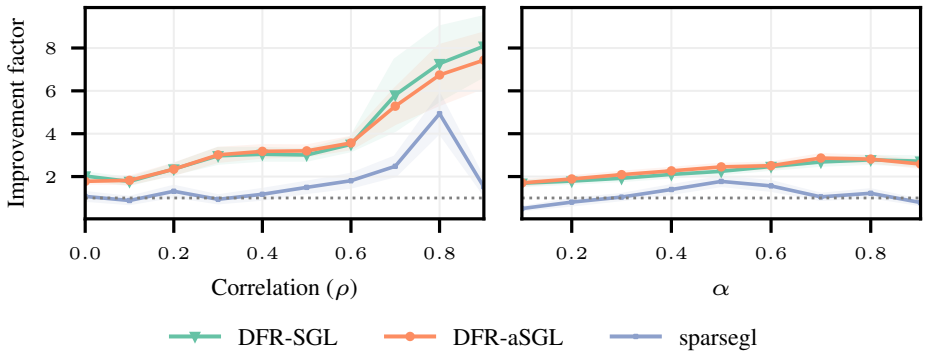

Figure A8: The improvement factor for the screening methods applied to synthetic data, under the logistic model, as a function of the data correlation (left) and α (right), with 95% confidence intervals.

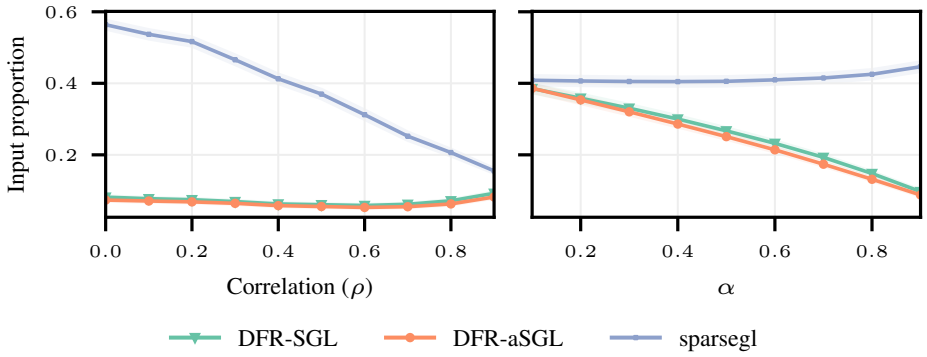

Figure A9: The input proportion for the screening methods applied to synthetic data, under the logistic model, as a function of the data correlation (left) and α (right), with 95% confidence intervals.

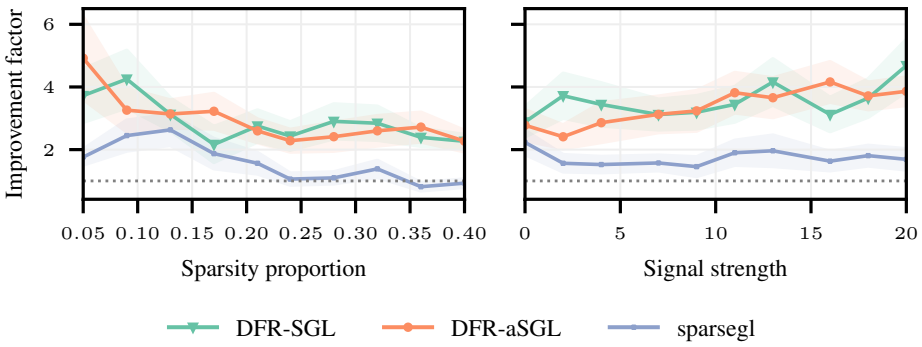

Figure A10: The improvement factor for the screening methods applied to synthetic data, under the logistic model, as a function of the data sparsity proportion (left) and signal strength (right), with 95% confidence intervals.

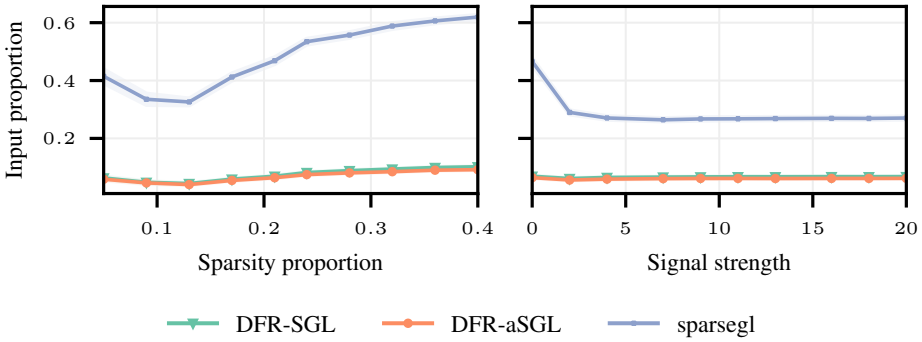

Figure A11: The input proportion for the screening methods applied to synthetic data, under the logistic model, as a function of the data sparsity proportion (left) and signal strength (right), with 95% confidence intervals.

D.4.2. INTERACTION DETECTION

Table A23: The improvement factor for the strong rules applied to synthetic interaction data, under the logistic model, with standard errors. For the interaction data, the parameters of the synthetic data were set as $p = 400, n = 80$, and $m = 52$ groups of sizes in $[3, 15]$. The interaction input dimensionality was $p_{O_2} = 2111, p_{O_3} = 7338$ for orders 2 and 3, with no interaction hierarchy imposed.

	Interaction	
Method	Order 2	Order 3
DFR-aSGL	6.7 ± 0.4	12.2 ± 0.4
DFR-SGL	5.8 ± 0.2	8.3 ± 0.4
sparsegl	1.0 ± 0.1	2.1 ± 0.3

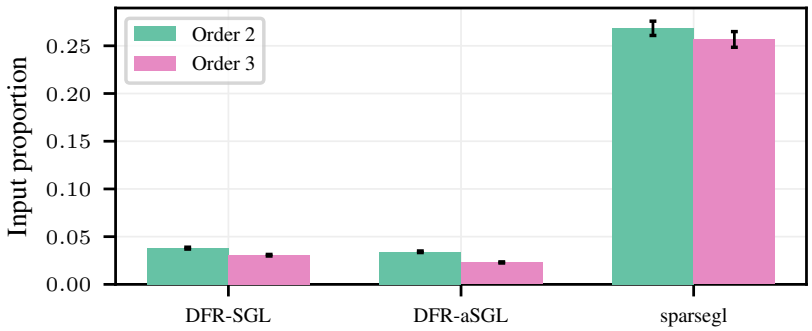

Figure A12: The input proportion for the strong rules applied to synthetic interaction data, under the logistic model, with standard errors. The parameters of the synthetic data were set as $p = 400, n = 80$, and $m = 52$ groups of sizes in [3, 15]. The interaction input dimensionality was $p_{O_2} = 2111, p_{O_3} = 7338$ for orders 2 and 3, with no interaction hierarchy imposed.

D.4.3. RESULTS TABLES FOR THE LOGISTIC MODEL

Table A24: Group screening metrics corresponding to the correlation simulation (Figures A8 and A9 (left)), averaged over all simulation iterations, correlation cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_g \downarrow$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g \,/\, \mathcal{A}_g$	$\mathcal{O}_g \,/\, m$
DFR-ASGL	7.46 ± 0.03	7.90 ± 0.03	7.91 ± 0.03	–	1.0379 ± 0.0012	0.3594 ± 0.0013
DFR-SGL	7.58 ± 0.03	8.20 ± 0.03	8.20 ± 0.03	–	1.0529 ± 0.0013	0.3726 ± 0.0014
SPARSEGL	7.58 ± 0.03	7.64 ± 0.03	7.64 ± 0.03	$2 \times 10^{-4} \pm 6 \times 10^{-5}$	$0.967 \pm 1 \times 10^{-3}$	0.3473 ± 0.0013

Table A25: Variable screening metrics corresponding the correlation simulation (Figures A8 and A9 (left)), averaged over all simulation iterations, correlation cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v \,/\, \mathcal{A}_v$	$\mathcal{O}_v \,/\, p$
DFR-ASGL	41.78 ± 0.15	24.28 ± 0.07	64.35 ± 0.20	$0.0165 \pm 6 \times 10^{-4}$	1.6487 ± 0.0022	$0.0643 \pm 2 \times 10^{-4}$
DFR-SGL	45.89 ± 0.16	27.01 ± 0.08	71.04 ± 0.22	0 ± 0	1.6442 ± 0.0023	$0.071 \pm 2 \times 10^{-4}$
SPARSEGL	45.89 ± 0.16	379.16 ± 1.50	379.17 ± 1.50	–	9.3997 ± 0.0307	0.3792 ± 0.0015

Table A26: Model fitting metrics corresponding to the correlation simulation (Figures A8 and A9 (left)), averaged over all simulation iterations, correlation cases, and path points, shown with standard errors.

	TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	86.0 ± 3.4	21.7 ± 0.4	3.8 ± 0.1	132.5 ± 6.1	73.3 ± 1.8	$1 \times 10^{-9} \pm 5 \times 10^{-11}$	0 ± 0	0 ± 0
DFR-SGL	97.4 ± 4.2	24.4 ± 0.5	4.0 ± 0.2	145.3 ± 6.5	79.9 ± 2.1	$2 \times 10^{-10} \pm 2 \times 10^{-12}$	0 ± 0	0 ± 0
SPARSEGL	97.4 ± 4.2	66.2 ± 1.7	1.8 ± 0.1	145.3 ± 6.5	132.2 ± 4.8	$3 \times 10^{-11} \pm 7 \times 10^{-13}$	0 ± 0	0 ± 0

Table A27: Group screening metrics corresponding to the α simulation (Figures A8 and A9 (right)), averaged over all simulation iterations, α cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g \,/\, \mathcal{A}_g$	$\mathcal{O}_g \,/\, m$
DFR-ASGL	8.41 ± 0.03	7.86 ± 0.03	7.86 ± 0.03	–	0.8905 ± 0.0013	0.3571 ± 0.0013
DFR-SGL	9.73 ± 0.03	9.30 ± 0.03	9.31 ± 0.03	–	0.9128 ± 0.0012	0.4230 ± 0.0013
SPARSEGL	9.74 ± 0.03	10.02 ± 0.03	10.02 ± 0.03	0 ± 0	1.3212 ± 0.011	0.4557 ± 0.0014

Table A28: Variable screening metrics corresponding to the α simulation (Figures A8 and A9 (right)), averaged over all simulation iterations, α cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v \,/\, \mathcal{A}_v$	$\mathcal{O}_v \,/\, p$
DFR-ASGL	179.90 ± 0.77	65.84 ± 0.31	238.58 ± 0.99	$9 \times 10^{-4} \pm 1 \times 10^{-4}$	1.3751 ± 0.0025	$0.2386 \pm 1 \times 10^{-3}$
DFR-SGL	186.76 ± 0.79	72.25 ± 0.34	251.18 ± 1.03	$4 \times 10^{-5} \pm 3 \times 10^{-5}$	1.3891 ± 0.0026	0.2512 ± 0.0010
SPARSEGL	186.86 ± 0.79	467.06 ± 1.51	467.07 ± 1.51	–	17.1921 ± 0.3909	0.4671 ± 0.0015

Table A29: Model fitting metrics corresponding to the α simulation (Figures A8 and A9 (right)), averaged over all simulation iterations, α cases, and path points, shown with standard errors.

	TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	101.5 ± 2.1	51.9 ± 1.3	2.2 ± 0.02	152.7 ± 2.5	112.5 ± 2.1	$1 \times 10^{-10} \pm 2 \times 10^{-11}$	0 ± 0	0 ± 0
DFR-SGL	107.8 ± 1.9	55.1 ± 1.4	2.7 ± 0.05	186.8 ± 2.8	123.8 ± 1.9	$2 \times 10^{-10} \pm 4 \times 10^{-11}$	0 ± 0	0 ± 0
SPARSEGL	107.8 ± 1.9	152.7 ± 4.2	1.0 ± 0.02	186.8 ± 2.8	169.5 ± 2.9	$4 \times 10^{-11} \pm 6 \times 10^{-13}$	0 ± 0	0 ± 0

Table A30: Group screening metrics corresponding to the sparsity proportion simulation (Figures A10 and A11 (left)), averaged over all simulation iterations, sparsity proportion cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g \,/\, \mathcal{A}_g$	$\mathcal{O}_g \,/\, m$
DFR-ASGL	9.09 ± 0.03	9.87 ± 0.03	9.88 ± 0.03	–	1.1033 ± 0.0011	0.4489 ± 0.0015
DFR-SGL	9.39 ± 0.03	10.41 ± 0.03	10.41 ± 0.03	–	1.1348 ± 0.0013	0.4733 ± 0.0016
SPARSEGL	9.39 ± 0.03	9.75 ± 0.03	9.75 ± 0.03	0 ± 0	$1.0344 \pm 8 \times 10^{-4}$	0.4433 ± 0.0015

Table A31: Variable screening metrics corresponding to the sparsity proportion simulation (Figures A10 and A11 (left)), averaged over all simulation iterations, sparsity proportion cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v \,/\, \mathcal{A}_v$	$\mathcal{O}_v \,/\, p$
DFR-ASGL	49.61 ± 0.2	21.32 ± 0.07	68.74 ± 0.26	$0.0211 \pm 7 \times 10^{-4}$	1.4654 ± 0.0021	$0.0687 \pm 3 \times 10^{-4}$
DFR-SGL	53.19 ± 0.21	24.26 ± 0.08	75.07 ± 0.28	0 ± 0	1.4990 ± 0.0022	$0.0751 \pm 3 \times 10^{-4}$
SPARSEGL	53.19 ± 0.21	486.18 ± 1.66	486.18 ± 1.66	–	12.231 ± 0.0381	0.4862 ± 0.0017

Table A32: Model fitting metrics corresponding to the sparsity proportion simulation (Figures A10 and A11 (left)), averaged over all simulation iterations, sparsity proportion cases, and path points, shown with standard errors.

	TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	41.4 ± 0.8	21.2 ± 0.4	2.9 ± 0.1	57.9 ± 0.4	40.6 ± 0.4	$8 \times 10^{-10} \pm 4 \times 10^{-11}$	0 ± 0	0 ± 0
DFR-SGL	48.6 ± 1.0	24.5 ± 0.5	2.9 ± 0.1	64.0 ± 0.4	43.9 ± 0.4	$1 \times 10^{-10} \pm 3 \times 10^{-12}$	0 ± 0	0 ± 0
SPARSEGL	48.6 ± 1.0	46.2 ± 0.9	1.6 ± 0.1	64.0 ± 0.4	63.6 ± 0.4	$5 \times 10^{-11} \pm 2 \times 10^{-12}$	0 ± 0	0 ± 0

Table A33: Group screening metrics corresponding to the signal strength simulation (Figures A10 and A11 (right)), averaged over all simulation iterations, signal strength cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g$ / $\mathcal{A}_g$	$\mathcal{O}_g$ / m
DFR-ASGL	5.77 ± 0.02	6.04 ± 0.02	6.04 ± 0.02	–	1.0242 ± 0.0012	0.2746 ± 0.001
DFR-SGL	5.80 ± 0.02	6.29 ± 0.02	6.29 ± 0.02	–	1.0564 ± 0.0014	0.2857 ± 0.0011
SPARSEGL	5.80 ± 0.02	5.71 ± 0.02	5.71 ± 0.02	$2 \times 10^{-5} \pm 2 \times 10^{-5}$	$0.9562 \pm 1 \times 10^{-3}$	$0.2594 \pm 1 \times 10^{-3}$

Table A34: Variable screening metrics corresponding to the signal strength simulation (Figures A10 and A11 (right)), averaged over all simulation iterations, signal strength cases, and path points, shown with standard errors.

	CARDINALITY				INPUT PROPORTION	
METHOD	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v$ / $\mathcal{A}_v$	$\mathcal{O}_v$ / p
DFR-ASGL	42.22 ± 0.12	20.54 ± 0.05	61.10 ± 0.15	$0.0142 \pm 5 \times 10^{-4}$	1.5138 ± 0.0014	$0.0611 \pm 2 \times 10^{-4}$
DFR-SGL	45.32 ± 0.12	23.19 ± 0.06	66.78 ± 0.16	0 ± 0	1.5277 ± 0.0014	$0.0668 \pm 2 \times 10^{-4}$
SPARSEGL	45.33 ± 0.12	290.29 ± 1.07	290.29 ± 1.07	–	6.2306 ± 0.0152	0.2903 ± 0.0011

Table A35: Model fitting metrics corresponding to the signal strength simulation (Figures A10 and A11 (right)), averaged over all simulation iterations, signal strength cases, and path points, shown with standard errors.

	TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	50.9 ± 1.0	21.8 ± 0.5	3.4 ± 0.1	68.2 ± 0.6	51.0 ± 0.4	$1 \times 10^{-9} \pm 5 \times 10^{-11}$	0 ± 0	0 ± 0
DFR-SGL	46.0 ± 0.8	20.0 ± 0.4	3.5 ± 0.1	76.3 ± 0.7	56.2 ± 0.5	$8 \times 10^{-11} \pm 1 \times 10^{-12}$	0 ± 0	0 ± 0
SPARSEGL	46.0 ± 0.8	38.8 ± 0.7	1.7 ± 0.1	76.3 ± 0.7	63.5 ± 0.5	$8 \times 10^{-11} \pm 9 \times 10^{-13}$	0 ± 0	0 ± 0

Table A36: Group screening metrics corresponding to the interaction simulation (Table A23 and Figure A12), averaged over all simulation iterations and path points, shown with standard errors.

		CARDINALITY				INPUT PROPORTION	
METHOD	TYPE	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g$ / $\mathcal{A}_g$	$\mathcal{O}_g$ / m
DFR-ASGL	ORDER 2	17.6 ± 0.2	23.1 ± 0.2	23.1 ± 0.2	–	1.320 ± 0.004	0.222 ± 0.002
DFR-SGL	ORDER 2	17.8 ± 0.2	25.0 ± 0.2	25.0 ± 0.2	–	1.419 ± 0.005	0.240 ± 0.002
SPARSEGL	ORDER 2	17.8 ± 0.2	20.9 ± 0.2	20.9 ± 0.2	$2 \times 10^{-4} \pm 2 \times 10^{-4}$	1.159 ± 0.003	0.201 ± 0.002
DFR-ASGL	ORDER 3	14.1 ± 0.2	19.9 ± 0.2	19.8 ± 0.2	–	1.396 ± 0.005	0.190 ± 0.002
DFR-SGL	ORDER 3	15.3 ± 0.2	23.7 ± 0.3	23.7 ± 0.3	–	1.525 ± 0.006	0.228 ± 0.002
SPARSEGL	ORDER 3	15.3 ± 0.2	19.3 ± 0.2	19.3 ± 0.2	$2 \times 10^{-4} \pm 2 \times 10^{-4}$	1.215 ± 0.003	0.185 ± 0.002

Table A37: Variable screening metrics corresponding to the interaction simulation (Table A23 and Figure A12), averaged over all simulation iterations and path points, shown with standard errors.

		CARDINALITY				INPUT PROPORTION	
METHOD	TYPE	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v$ / $\mathcal{A}_v$	$\mathcal{O}_v$ / p
DFR-ASGL	ORDER 2	40.6 ± 0.4	33.1 ± 0.3	72.1 ± 0.7	0.081 ± 0.004	1.773 ± 0.007	0.034 ± 0.0003
DFR-SGL	ORDER 2	40.8 ± 0.4	41.0 ± 0.4	80.1 ± 0.8	0 ± 0	1.951 ± 0.009	0.038 ± 0.0004
SPARSEGL	ORDER 2	40.8 ± 0.4	566.4 ± 6.2	566.4 ± 6.2	–	13.026 ± 0.075	0.268 ± 0.0030
DFR-ASGL	ORDER 3	74.1 ± 1.0	97.7 ± 1.2	168.7 ± 2.0	0.058 ± 0.004	2.405 ± 0.025	0.023 ± 0.0003
DFR-SGL	ORDER 3	82.7 ± 1.1	145.3 ± 1.8	224.2 ± 2.8	0 ± 0	2.829 ± 0.025	0.031 ± 0.0004
SPARSEGL	ORDER 3	82.8 ± 1.1	1883.4 ± 25.2	1883.4 ± 25.2	–	20.204 ± 0.184	0.257 ± 0.0030

Table A38: Model fitting metrics corresponding to the interaction simulation (Table A23 and Figure A12), averaged over all simulation iterations and path points, shown with standard errors.

		TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	TYPE	NO SCREEN (S)	SCREEN (S)	IMPROVEMENT FACTOR	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	ORDER 2	357 ± 22	68 ± 5	7 ± 0.4	306 ± 15	150 ± 4	$2 \times 10^{-9} \pm 1 \times 10^{-10}$	0 ± 0	0 ± 0
DFR-SGL	ORDER 2	253 ± 13	45 ± 2	6 ± 0.2	326 ± 18	166 ± 5	$7 \times 10^{-10} \pm 1 \times 10^{-11}$	0 ± 0	0 ± 0
SPARSEGL	ORDER 2	253 ± 13	367 ± 30	1 ± 0.1	326 ± 18	197 ± 12	$1 \times 10^{-9} \pm 2 \times 10^{-11}$	0 ± 0	0 ± 0
DFR-ASGL	ORDER 3	2425 ± 198	208 ± 10	12 ± 1.0	835 ± 57	358 ± 12	$2 \times 10^{-9} \pm 8 \times 10^{-11}$	0 ± 0	0 ± 0
DFR-SGL	ORDER 3	1077 ± 91	134 ± 8	8 ± 0.4	844 ± 62	451 ± 19	$5 \times 10^{-10} \pm 1 \times 10^{-11}$	0 ± 0	0 ± 0
SPARSEGL	ORDER 3	1077 ± 91	1167 ± 156	2 ± 0.3	844 ± 62	561 ± 43	$5 \times 10^{-10} \pm 2 \times 10^{-11}$	0 ± 0	0 ± 0

Appendix E. Real data analysis

E.1. Data description

- brca1: Gene expression data for breast cancer tissue samples.
  - Response (continuous): Gene expression measurements for the BRCA1 gene.
  - Data matrix: Gene expression measurements for the other genes.
  - Grouping structure: Variables are grouped via singular value decomposition.
- scheetz: Gene expression data in the mammalian eye.
  - Response (continuous): Gene expression measurements for the Trim32 gene.
  - Data matrix: Gene expression measurements for the other genes.
  - Grouping structure: Variables are grouped via singular value decomposition.
- trust-experts: Survey response data as to how much participants trust *experts* (e.g. doctors, nurses, scientists) to provide COVID-19 news and information.
  - Response (continuous): The trust level of each participant.
  - Data matrix: Contingency table including factors about participants (e.g. age, gender, ethnicity).
  - Grouping structure: The factor levels grouped into their original factors.
- adenoma: Transcriptome profile data to identify the formation of colorectal adenomas, which are the predominate cause of colorectal cancers.
  - Response (binary): Labels classifying whether the sample came from an adenoma or normal mucosa.
  - Data matrix: Transcriptome profile measurements.
  - Grouping structure: Genes were assigned to pathways from all nine gene sets on the Molecular Signature Database.
- celiac: Gene expression data of primary leucocytes to identify celiac disease.
  - Response (binary): Labels classifying patients into whether they have celiac disease.
  - Data matrix: Gene expression measurements from the primary leucocytes.
  - Grouping structure: Genes were assigned to pathways from all nine gene sets on the Molecular Signature Database.
- tumour: Gene expression data of pancreative cancer samples to identify tumorous tissue.
  - Response (binary): Labels classifying whether sample is from normal of tumour tissue.
  - Data matrix: Gene expression measurements.
  - Grouping structure: Genes were assigned to pathways from all nine gene sets on the Molecular Signature Database.

Table A39: Dataset information for the six datasets used in the real data analysis.

Dataset	p	n	m	Group sizes	Type	Source
brca1	17322	536	243	$[1, 6505]$	Linear	[16][1]
scheetz	18975	120	85	$[1, 6274]$	Linear	[35][1]
trust-experts	101	9759	7	$[4, 51]$	Linear	[34][2]
adenoma	18559	64	313	$[1, 741]$	Logistic	[33][3]
celiac	14657	132	276	$[1, 617]$	Logistic	[14][3]
tumour	18559	52	313	$[1, 741]$	Logistic	[8, 21, 30][3]

1. downloaded from https://iowabiostat.github.io/data-sets/
2. downloaded from https://github.com/dajmcdon/sparsegl
3. downloaded from https://www.ncbi.nlm.nih.gov/

E.2. Additional commentary

Three of the datasets, *brca1*, *scheetz*, and *trust-experts*, have continuous responses, so are fit using an SGL linear model. The former two were also analyzed with regards to screening rules in Larsson and Wallin [18], and the later in Liang et al. [23]. The other three datasets, *adenoma*, *celiac*, and *tumour*, have binary responses, so an SGL logistic model is used. The *trust-experts* dataset is low-dimensional, and the other five are high-dimensional.

Even in the case of low-dimensional data (*trust-experts*), DFR provides a clear benefit. DFR-aSGL performs very well for *scheetz* and *adenoma*, improving the computational cost by over 600 times. For the *scheetz* dataset, the aSGL model had more difficulty converging without screening compared to SGL, so DFR-aSGL offered a greater advantage over DFR-SGL. For *adenoma*, the active set for aSGL was smaller (Table A41), due to the increased penalization that comes with the adaptivity. However, despite the advantage of a smaller active set, we do still observe that DFR-aSGL was more efficient at reducing the optimization set, *with respect to the active set*.

DFR is observed to aid in mitigating convergence issues for both SGL and aSGL (Table A42). Across all datasets, DFR encountered no failed convergences. In contrast, sparsegl did not converge at several path points for both *adenoma* and *scheetz*. As sparsegl only screens groups, when a group enters the optimization set, sparsegl is forced to fit with the full group, which can contain noise variables. Applying no screening led to SGL not converging for *adenoma*, *scheetz*, and *tumour*. By drastically reducing the input space, convergence issues arising from large datasets are resolved, which not only improves computational cost, but also solution optimality.

E.3. Additional results for the real data

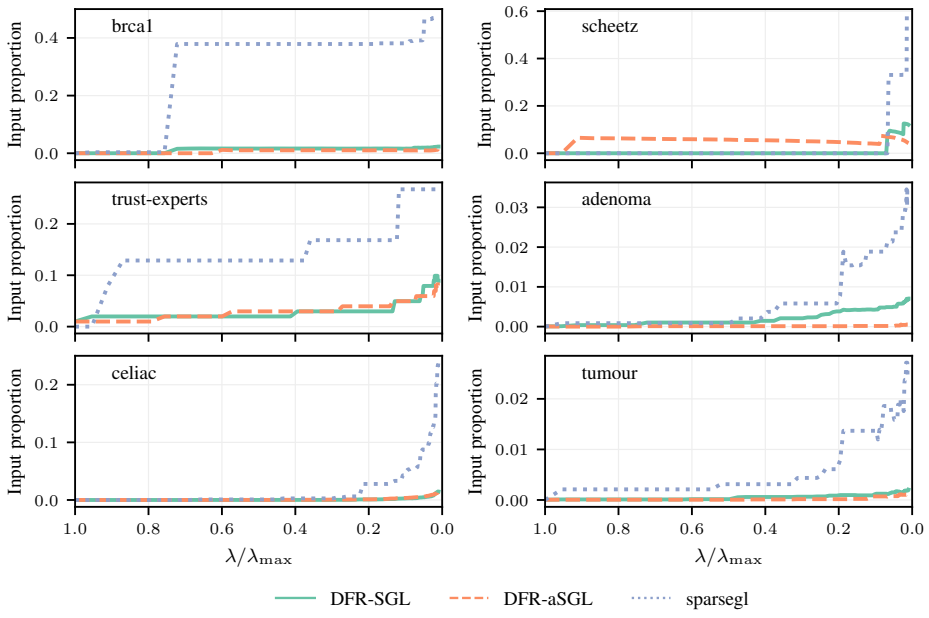

Figure A13: The input proportion as a function of the shrinkage path for the screening methods applied to the real datasets.

Table A40: Group screening metrics corresponding to Figure 3 averaged over all path points, shown with standard errors.

		Cardinality				Input proportion	
Method	Dataset	$\mathcal{A}_g$	$\mathcal{C}_g$	$\mathcal{O}_g$	$\mathcal{K}_g$	$\mathcal{O}_g$ / $\mathcal{A}_g$	$\mathcal{O}_g$ / m
DFR-aSGL	adenoma	1.41 ± 0.08	1.38 ± 0.08	1.42 ± 0.08	-	1.0053 ± 0.0053	$0.0046 \pm 3 \times 10^{-4}$
DFR-SGL	adenoma	3.30 ± 0.21	13.46 ± 0.68	13.46 ± 0.68	-	4.6121 ± 0.2003	0.0430 ± 0.0022
sparsegl	adenoma	3.32 ± 0.21	8.59 ± 0.53	8.59 ± 0.53	0 ± 0	2.6141 ± 0.0942	0.0274 ± 0.0017
DFR-aSGL	celiac	19.41 ± 1.65	29.94 ± 2.45	28.82 ± 2.35	-	1.4412 ± 0.0273	0.1044 ± 0.0085
DFR-SGL	celiac	15.35 ± 1.45	22.04 ± 2.03	22.04 ± 2.03	-	1.4367 ± 0.0276	0.0799 ± 0.0074
sparsegl	celiac	15.36 ± 1.46	19.32 ± 1.84	19.32 ± 1.84	0 ± 0	1.2415 ± 0.0213	0.0700 ± 0.0067
DFR-aSGL	bc-tgca	3.84 ± 0.26	4.17 ± 0.33	4.16 ± 0.33	-	1.0439 ± 0.0121	0.0171 ± 0.0013
DFR-SGL	bc-tgca	5.60 ± 0.48	6.90 ± 0.60	6.90 ± 0.60	-	1.2023 ± 0.0218	0.0284 ± 0.0025
sparsegl	bc-tgca	5.59 ± 0.48	6.25 ± 0.55	6.25 ± 0.55	0 ± 0	1.1062 ± 0.0176	0.0257 ± 0.0022
DFR-aSGL	scheetz	2.19 ± 0.12	2.37 ± 0.16	2.39 ± 0.16	-	1.0515 ± 0.0121	0.0282 ± 0.0019
DFR-SGL	scheetz	0.61 ± 0.08	0.86 ± 0.12	0.86 ± 0.12	-	1.3537 ± 0.0531	0.0101 ± 0.0014
sparsegl	scheetz	0.61 ± 0.08	0.73 ± 0.10	0.73 ± 0.10	0 ± 0	1.1829 ± 0.0486	0.0086 ± 0.0012
DFR-aSGL	trust-experts	3.41 ± 0.08	3.37 ± 0.09	3.41 ± 0.08	-	1 ± 0	0.4877 ± 0.0118
DFR-SGL	trust-experts	3.34 ± 0.08	3.37 ± 0.08	3.37 ± 0.08	-	1.0185 ± 0.0117	0.4820 ± 0.0113
sparsegl	trust-experts	3.30 ± 0.09	0.04 ± 0.02	3.30 ± 0.09	0 ± 0	1 ± 0	0.4719 ± 0.0127
DFR-aSGL	tumour	3.94 ± 0.22	4.96 ± 0.30	4.98 ± 0.30	-	1.2228 ± 0.0240	$0.0159 \pm 1 \times 10^{-3}$
DFR-SGL	tumour	5.02 ± 0.24	8.80 ± 0.38	8.80 ± 0.38	-	1.8253 ± 0.0357	0.0281 ± 0.0012
sparsegl	tumour	5.02 ± 0.24	6.77 ± 0.29	6.77 ± 0.29	0 ± 0	1.4276 ± 0.0351	$0.0216 \pm 9 \times 10^{-4}$

Table A41: Variable screening metrics corresponding to Figure 3 averaged over all path points, shown with standard errors.

		CARDINALITY				INPUT PROPORTION	
METHOD	DATASET	$\mathcal{A}_v$	$\mathcal{C}_v$	$\mathcal{O}_v$	$\mathcal{K}_v$	$\mathcal{O}_v$ / $\mathcal{A}_v$	$\mathcal{O}_v$ / p
DFR-ASGL	ADENOMA	3.38 ± 0.21	0.81 ± 0.11	4.15 ± 0.30	0.0404 ± 0.0199	1.1828 ± 0.0259	$2 \times 10^{-4} \pm 2 \times 10^{-5}$
DFR-SGL	ADENOMA	14.38 ± 1.11	61.47 ± 2.90	75.51 ± 3.85	0 ± 0	8.9655 ± 0.7776	$0.0041 \pm 2 \times 10^{-4}$
SPARSEGL	ADENOMA	14.41 ± 1.12	308.64 ± 20.81	308.64 ± 20.81	-	26.0099 ± 1.6361	0.0166 ± 0.0011
DFR-ASGL	CELIAC	40.13 ± 3.85	29.63 ± 2.79	68.61 ± 6.51	0.1010 ± 0.0337	1.6545 ± 0.0442	$0.0047 \pm 4 \times 10^{-4}$
DFR-SGL	CELIAC	37.13 ± 3.92	26.11 ± 3.05	61.94 ± 6.77	0 ± 0	1.6187 ± 0.0455	$0.0042 \pm 5 \times 10^{-4}$
SPARSEGL	CELIAC	37.26 ± 3.94	1019.13 ± 106.78	1019.13 ± 106.78	-	27.6753 ± 1.1176	0.0695 ± 0.0073
DFR-ASGL	BC-TGCA	135.75 ± 4.69	31.93 ± 2.77	165.83 ± 6.19	0.0505 ± 0.0221	1.1998 ± 0.0116	$0.0096 \pm 4 \times 10^{-4}$
DFR-SGL	BC-TGCA	241.69 ± 7.34	63.42 ± 4.26	301.8 ± 8.81	0 ± 0	3.9386 ± 2.6945	$0.0174 \pm 5 \times 10^{-4}$
SPARSEGL	BC-TGCA	241.59 ± 7.34	6762.31 ± 186.59	6762.31 ± 186.59	-	95.8171 ± 66.017	0.3904 ± 0.0108
DFR-ASGL	SCHEETZ	743.43 ± 13.50	281.91 ± 13.50	1019.11 ± 20.19	0.0202 ± 0.0142	1.3678 ± 0.0050	0.0537 ± 0.0011
DFR-SGL	SCHEETZ	501.78 ± 60.68	344.92 ± 50.24	836.23 ± 101.02	0 ± 0	1.6688 ± 0.0651	0.0441 ± 0.0053
SPARSEGL	SCHEETZ	501.49 ± 60.66	3030.24 ± 385.60	3030.24 ± 385.60	-	6.1978 ± 0.3472	0.1597 ± 0.0203
DFR-ASGL	TRUST-EXPERTS	4.83 ± 0.20	0.17 ± 0.04	4.96 ± 0.21	0.0404 ± 0.0199	1.0231 ± 0.0067	0.0491 ± 0.0021
DFR-SGL	TRUST-EXPERTS	5.10 ± 0.28	0.35 ± 0.07	5.36 ± 0.29	0 ± 0	1.0608 ± 0.0155	0.0531 ± 0.0029
SPARSEGL	TRUST-EXPERTS	5.07 ± 0.28	0.27 ± 0.14	21.42 ± 0.65	-	4.9628 ± 0.1448	0.2121 ± 0.0065
DFR-ASGL	TUMOUR	7.40 ± 0.58	3.31 ± 0.31	10.57 ± 0.84	0.0202 ± 0.0142	1.3363 ± 0.0394	$6 \times 10^{-4} \pm 5 \times 10^{-5}$
DFR-SGL	TUMOUR	10.70 ± 0.72	9.87 ± 0.50	20.34 ± 1.09	0 ± 0	2.3432 ± 0.1189	$0.0011 \pm 6 \times 10^{-5}$
SPARSEGL	TUMOUR	10.70 ± 0.72	246.8 ± 15.39	246.8 ± 15.39	-	25.9945 ± 0.9193	$0.0133 \pm 8 \times 10^{-4}$

Table A42: Model fitting metrics corresponding to Figure 3 averaged over all path points, shown with standard errors. There are no standard errors for the timing results as the time to calculate the whole path was evaluated.

		TIMINGS			ITERATIONS		ℓ_2 DISTANCE	FAILED CONVERGENCE	
METHOD	DATASET	NO SCREEN (S)	SCREEN (S)	I.F.	NO SCREEN	SCREEN	TO NO SCREEN	NO SCREEN	SCREEN
DFR-ASGL	ADENOMA	8476.04	7.91	1072.10	8903.76 ± 270.64	119.18 ± 14.69	$7 \times 10^{-5} \pm 6 \times 10^{-6}$	0.8384 ± 0.0372	0 ± 0
DFR-SGL	ADENOMA	9017.70	149.10	60.48	9272.90 ± 205.31	4374.64 ± 286.62	$3 \times 10^{-5} \pm 2 \times 10^{-6}$	0.8687 ± 0.0341	0 ± 0
SPARSEGL	ADENOMA	9017.70	198.36	45.46	9272.90 ± 205.31	5140.16 ± 390.91	$3 \times 10^{-5} \pm 2 \times 10^{-6}$	0.8687 ± 0.0341	0.0404 ± 0.0199
DFR-ASGL	CELIAC	1188.14	15.86	74.89	1042.07 ± 90.79	120.96 ± 14.51	$1 \times 10^{-6} \pm 2 \times 10^{-7}$	0 ± 0	0 ± 0
DFR-SGL	CELIAC	1391.78	10.31	134.95	1195.34 ± 98.13	75.37 ± 9.60	$2 \times 10^{-7} \pm 1 \times 10^{-8}$	0 ± 0	0 ± 0
SPARSEGL	CELIAC	1391.78	16.49	84.40	1195.34 ± 98.13	93.29 ± 8.86	$8 \times 10^{-8} \pm 3 \times 10^{-9}$	0 ± 0	0 ± 0
DFR-ASGL	BC-TGCA	21889.33	119.16	183.69	1653.45 ± 33.90	243.27 ± 11.70	$2 \times 10^{-9} \pm 7 \times 10^{-10}$	0 ± 0	0 ± 0
DFR-SGL	BC-TGCA	22227.01	103.78	214.17	1674.07 ± 19.77	334.16 ± 13.44	$2 \times 10^{-12} \pm 3 \times 10^{-13}$	0 ± 0	0 ± 0
SPARSEGL	BC-TGCA	22227.01	4132.04	5.38	1674.07 ± 19.77	1580.73 ± 44.23	$4 \times 10^{-14} \pm 1 \times 10^{-14}$	0 ± 0	0 ± 0
DFR-ASGL	SCHEETZ	90569.05	132.20	685.08	6040.81 ± 457.60	1030.48 ± 90.27	$1 \times 10^{-6} \pm 3 \times 10^{-7}$	0.5657 ± 0.0501	0 ± 0
DFR-SGL	SCHEETZ	68084.77	2246.13	30.31	1891.21 ± 386.40	642.38 ± 136.12	$3 \times 10^{-9} \pm 8 \times 10^{-10}$	0.1818 ± 0.0390	0 ± 0
SPARSEGL	SCHEETZ	68084.77	6666.65	10.21	1891.21 ± 386.40	1890.45 ± 386.44	$5 \times 10^{-22} \pm 2 \times 10^{-22}$	0.1818 ± 0.0390	0.1818 ± 0.0390
DFR-ASGL	TRUST-EXPERTS	5.96	3.10	1.92	76.74 ± 1.80	85.11 ± 3.96	$1 \times 10^{-11} \pm 3 \times 10^{-12}$	0 ± 0	0 ± 0
DFR-SGL	TRUST-EXPERTS	7.29	3.20	2.28	98.36 ± 4.17	92.55 ± 5.93	$4 \times 10^{-11} \pm 9 \times 10^{-12}$	0 ± 0	0 ± 0
SPARSEGL	TRUST-EXPERTS	7.29	4.52	1.61	98.36 ± 4.17	104.24 ± 5.15	$2 \times 10^{-6} \pm 2 \times 10^{-6}$	0 ± 0	0 ± 0
DFR-ASGL	TUMOUR	7027.69	86.25	81.48	8783.1 ± 265.36	82.75 ± 5.61	$3 \times 10^{-8} \pm 9 \times 10^{-9}$	0.6768 ± 0.0472	0 ± 0
DFR-SGL	TUMOUR	7466.83	89.43	83.49	9272.02 ± 224.60	186.27 ± 7.64	$3 \times 10^{-9} \pm 2 \times 10^{-10}$	0.8586 ± 0.0352	0 ± 0
SPARSEGL	TUMOUR	7466.83	90.03	82.93	9272.02 ± 224.60	197.40 ± 9.38	$2 \times 10^{-9} \pm 2 \times 10^{-10}$	0.8586 ± 0.0352	0 ± 0