# Methodological Advancements in High-dimensional Sparse-group Models

A thesis submitted for the degree of
*Doctor of Philosophy*
by

Fabio Feser
Department of Mathematics
Imperial College London

March 22, 2026

## Statement of Originality

All of the work in this thesis is my own unless otherwise referenced. The work in this thesis was supervised by Dr Marina Evangelou.
The contents of Chapters 1 and 2 were written by me for this thesis and include an introduction and relevant background material.
The content of Chapter 3 is largely based on our manuscript Feser and Evangelou [66], which is currently under review.
The content of Chapter 4 is largely based on our paper Feser and Evangelou [69], published at *AISTATS '25*.
The content of Chapter 5 is largely based on our workshop paper published at *NeurIPS OPT Workshop '24* [67] and full conference paper published at *ICML '25* [68].
The content of Chapter 6 was written by me especially for this thesis and contains relevant background materials for Chapter 7.
The content of Chapter 7 is currently being prepared for submission for publication.
The contents of Chapter 8 were written by me for this thesis and contain a discussion.

## Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

# Abstract

This thesis develops novel sparse-group methodologies for high-dimensional data analysis, with a particular emphasis on genetic applications. Genes naturally form in groups, and utilising this inherent grouping structure can substantially enhance model performance. Traditional group-only models often perform poorly due to the inclusion of noise variables within active groups. Sparse-group models address this limitation by introducing additional sparsity within groups. Moreover, effective control of the false discovery rate (FDR) is crucial, as it supports cost-efficient validation of findings.

The main methodological contribution of this thesis is the Sparse-group SLOPE (SGS) model. The SGS framework achieves FDR control at both the group and variable levels, allowing it to outperform competitive methods in disease prediction. To improve the computational efficiency of SGS, a new sparse-group strong screening framework is introduced, which performs two layers of screening. It is applied to both SGS and group SLOPE (gSLOPE), substantially reducing runtime by excluding inactive features before optimisation, without compromising solution optimality. This framework is also extended to the Sparse-group Lasso (SGL) and adaptive SGL models using the Dual Feature Reduction (DFR) method, with our empirical results establishing it as the state-of-the-art screening approach for SGL. Further, this thesis tackles the problem of model selection for SLOPE models under general settings by developing Bayesian formulations of gSLOPE and SGS. The models employ spike-and-slab priors, providing simultaneous estimation of noise and model parameters, uncertainty quantification, and adaptivity to underlying sparsity structures. A two-step procedure and a scaled regression approach are also proposed for SLOPE models. Empirical studies demonstrate that the proposed Bayesian methods provide the optimal implementation of SLOPE under general settings, achieving reliable FDR control across a wide range of scenarios and demonstrating superior power and predictive performance compared to existing approaches.

## Acknowledgements

First and foremost, I must thank my wonderful supervisor, Dr Marina Evangelou. Your help and guidance throughout the PhD have been absolutely invaluable, and I'm grateful for the constant support you have given me. You have enabled many opportunities for me, and I am proud to be presenting our work together in this thesis. I would also like to thank the many academics I had the opportunity of working with throughout my PhD, including Dr Heather Battey, Professor Nancy Reid, Professor Chiara Sabatti, and Dr Johan Larsson. I am grateful for the insights and knowledge I gained through these collaborations.

I would also like to thank Imperial College London and the StatML programme for providing me with the opportunity to pursue a PhD. I have learned so much and had many exciting experiences through the StatML programme, including academic visits to the University of Toronto and Stanford University, as well as enjoyable schools in Lake Como and Windsor, and the numerous trips to Oxford. Despite the District line trying its utmost to stop me getting to campus, I was always happy to arrive there and spend time with the other fantastic PhD students at Imperial, especially Josh and Luke, who brightened every day for me. I would also like to thank my friends who have supported me throughout my PhD journey, always providing social occasions to look forward to. In particular, I am thankful to Rich and Lorenz, as well as Henry for his help with the internship.

My deepest thanks go to my family and partner. My brothers, Luca and Paolo, have been there for me my whole life, and the PhD was no exception. I can never fully express my gratitude to my greatest supporter, my Mum, who has always provided an endless supply of love and encouragement throughout my life. Finally, thank you to Kate for the wonderful years spent in Putney. I am so glad we were able to go on this adventure together, and I am excited for the next ones coming up.

# Contents

**Abstract** iii

**Acknowledgements** v

**1 Introduction** 1

1.1 Outline of the Thesis . . . 3

1.2 Code Availability . . . 4

**2 Background** 5

2.1 Penalised Regression . . . 5

2.2 Sorted $\ell_1$ Penalised Estimation (SLOPE) . . . 10

2.3 Group-based Models . . . 14

**3 Sparse-group SLOPE: Adaptive Bi-level Selection with FDR Control** 17

3.1 Introduction . . . 18

3.2 Sparse-group SLOPE . . . 20

3.3 Proximal Fitting Algorithm . . . 23

3.4 FDR Control . . . 27

3.5 Synthetic Data Experiments . . . 37

3.6 Real Data Experiments . . . 43

3.7 Discussion . . . 48

**4 Strong Screening Rules for Group-based SLOPE Models** 51

4.1 Screening Rules Background . . . 52

4.2 Introduction . . . 56

4.3 Screening Rule Framework . . . 57

4.4 Group SLOPE Screening . . . 58

4.5 Sparse-group SLOPE Screening . . . 64
4.6 Synthetic Data Experiments . . . 68
4.7 Real Data Experiments . . . 74
4.8 Discussion . . . 78

**5 Dual Feature Reduction for the Sparse-group Lasso and its Adaptive Variant 81**
5.1 Introduction . . . 82
5.2 Dual Feature Reduction for SGL . . . 86
5.3 Dual Feature Reduction for aSGL . . . 93
5.4 Synthetic Data Experiments . . . 97
5.5 Real Data Experiments . . . 105
5.6 Discussion . . . 107

**6 Background Materials for "Model Selection for SLOPE Models: A Bayesian Perspective" 109**
6.1 Model Selection . . . 109
6.2 Relevant Approaches . . . 112
6.3 Bayesian Models . . . 117

**7 Model Selection for SLOPE Models: A Bayesian Perspective 121**
7.1 Introduction . . . 122
7.2 SLOPE in a Bayesian Context . . . 124
7.3 Group-based Bayesian SLOPE Models . . . 125
7.4 Stochastic Approximation EM (SAEM) Algorithm . . . 130
7.5 Initialisations and Sensitivity Analysis . . . 141
7.6 Other Key Proposals . . . 143
7.7 Synthetic Data Experiments . . . 147
7.8 Real Data Experiments . . . 160
7.9 Discussion . . . 163

**8 Discussion 165**
8.1 Summary of Contributions . . . 165
8.2 Methodological Limitations . . . 167
8.3 Future Work . . . 170
8.4 Concluding Remarks . . . 174

**Appendices** **199**

**A General** **199**

A.1 Notation . . . 199
A.2 Definitions . . . 202
A.3 Computational Information . . . 202

**B Additional Material for Chapter 3** **211**

B.1 Sparse-group SLOPE . . . 211
B.2 FDR Control . . . 214
B.3 Synthetic Data Experiments . . . 241
B.4 Real Data Experiments . . . 244

**C Additional Material for Chapter 4** **249**

C.1 Group SLOPE . . . 249
C.2 Sparse-group SLOPE . . . 250
C.3 Screening Rule Framework . . . 251
C.4 Group-based OSCAR . . . 252
C.5 Data Experiments . . . 255

**D Additional Material for Chapter 5** **263**

D.1 Sparse-group Lasso . . . 263
D.2 Adaptive Sparse-group Lasso . . . 265
D.3 Data Experiments . . . 266

**E Additional Material for Chapter 7** **277**

E.1 Group-based Bayesian SLOPE Models . . . 277
E.2 SAEM for BGSLOPE . . . 284
E.3 SAEM for BSGS . . . 288
E.4 Illustrative Example . . . 300
E.5 Initialisations and Stability Analysis . . . 301
E.6 Other Approaches . . . 304
E.7 Synthetic Data Experiments . . . 306

# List of Figures


2.1 SLOPE BH and gSLOPE sequences for varying $q_v$ and $q_g$ . . . . . 11

3.1 Penalty unit balls in $\mathbb{R}^3$ for SLOPE models . . . . . 21
3.2 Sensitivity and FDR trade-off diagram as function of $(\lambda, \alpha)$ . . . . . 21
3.3 Pathwise fitted coefficients for SGS . . . . . 22
3.4 Variable and group sequences for SGS . . . . . 32
3.5 vFDR and gFDR under different $a_j$ estimators . . . . . 33
3.6 vSensitivity and gSensitivity under different $a_j$ estimators . . . . . 34
3.7 vFDR and gFDR under orthogonal design with even groups . . . . . 36
3.8 vFDR and gFDR under orthogonal design with uneven groups . . . 36
3.9 Global metrics for SLOPE- and lasso-based models . . . . . 39
3.10 Global metrics for SGS variants . . . . . 39
3.11 Metrics for SLOPE-based models under decreasing sparsity . . . . . 40
3.12 Metrics for SLOPE-based models under varying $p/n$ ratio . . . . . 43
3.13 Classification rate for SLOPE-based models for real data . . . . . 46
3.14 Additional metrics for SGS for Colitis data . . . . . 47
3.15 Hyperparameter choice for SGS for Colitis data . . . . . 48

4.1 Screened proportion for group-only vs bi-level screening . . . . . 70
4.2 Proportion of variables/groups in screened sets along a path . . . . . 71
4.3 Proportion of variable/groups relative to input along a path . . . . . 71
4.4 Proportion of variable/groups in screened sets as function of correlation 71
4.5 Runtime for screen vs no screen as a function of dimensionality . . . 72
4.6 Proportion of KKT violations for synthetic study . . . . . 73
4.7 Improvement factor on real data . . . . . 75
4.8 Screened proportion on a path for real data . . . . . 76
4.9 Input proportion on real data . . . . . 76

4.10 KKT violations for real data . . . . . 78

5.1 aSGL weights . . . . . 94
5.2 Screening metrics for GAP safe comparison under varying $\alpha$ . . . . . 99
5.3 Screening metrics for GAP safe comparison under varying signal strength 99
5.4 Screening metrics under varying dimensionality . . . . . 101
5.5 Improvement factor under varying data parameters . . . . . 102
5.6 Input proportion under varying data parameters . . . . . 102
5.7 Screening metrics for interaction data . . . . . 103
5.8 Improvement factor for aSGL under varying weights . . . . . 104
5.9 Screening metrics for real data . . . . . 106
5.10 Input proportion as a function of a path for real data . . . . . 107

7.1 FDR for SLOPE and SGS along a path for CV . . . . . 122
7.2 BGSLOPE dependency graph . . . . . 128
7.3 BSGS dependency graph . . . . . 128
7.4 Bayesian parameters for illustrative example . . . . . 140
7.5 Bias of BSGS under different $\boldsymbol{\beta}$ initialisation models . . . . . 142
7.6 Comparison between AS-SGS and Scaled SGS for different $\alpha$ . . . . 145
7.7 BH and Gaussian sequences for SLOPE . . . . . 147
7.8 FDR and power of best approaches under varying signal . . . . . 151
7.9 OOS of best approaches under varying signal . . . . . 152
7.10 Bias of all approaches under varying signal . . . . . 152
7.11 FDR and power of best approaches under varying data parameters . 153
7.12 MAE($\sigma$) of all approaches under varying data parameters . . . . . 154
7.13 Power and FDR of BSGS under different schemes under varying noise 154
7.14 FDR and power of best approaches under varying correlation . . . . 156
7.15 Impact of $\alpha$ on SGS models . . . . . 159

B.1 FDR metrics of SGS Original and SGS Double . . . . . 237
B.2 Sensitivity metrics of SGS Original and SGS Double . . . . . 237
B.3 FDR metrics of SGS max under different $a_j$ estimators . . . . . 238
B.4 Sensitivity metrics of SGS max under different $a_j$ estimators . . . . 238
B.5 Sensitivity metrics of SGS under orthogonal design with even groups 239
B.6 Metrics of SGS max under orthogonal design with even groups . . . 239
B.7 Metrics of SGS mean under orthogonal design with even groups . . . 240

B.8 Sensitivity metrics under orthogonal designs with uneven groups . . 240
B.9 Number of features selected the models under synthetic data . . . . 241
B.10 Metrics for SLOPE- and lasso-based models under varying correlation 241
B.11 Sensitivity for SLOPE-based models under decreasing sparsity . . . 242
B.12 Metrics for lasso-based models and SGS under decreasing sparsity . 242
B.13 Sensitivity for SLOPE-based models under varying $p/n$ ratio . . . . 243
B.14 Metrics for lasso-based models and SGS under varying $p/n$ ratio . . 243
B.15 Classification rate for lasso-based models for real data . . . . . 245
B.16 Classification rate for OSCAR-based models for real data . . . . . 245

C.1 Screened proportion for group-only vs bi-level screening for SGO . . 253
C.2 Proportion of variables/groups in screened sets along a path for SGO 253
C.3 Proportion of variables/groups in screened sets as a function of correlation for SGO . . . . . 253
C.4 Runtime for screen vs no screen as a function of dimensionality for OSCAR models . . . . . 253
C.5 Proportion of KKT violations for synthetic study for OSCAR models 254
C.6 Improvement factor on real data for OSCAR models . . . . . 254
C.7 Input proportion on real data for OSCAR models . . . . . 255
C.8 Proportion of variables/groups in screened sets along a path under different parameters . . . . . 257
C.9 Proportion of variables/groups in screened sets along a path under logistic model . . . . . 258
C.10 Equivalent of Figure 4.2 for the logistic model . . . . . 259
C.11 Runtime for screened vs no screen as a function of dimensionality for logistic model . . . . . 259
C.12 Improvement factor of all SLOPE models on real data . . . . . 259
C.13 Ratio of fitting set to active of all SLOPE models on real data . . . 260

D.1 Improvement factor under varying signal strength and $\alpha$ . . . . . 267
D.2 Input proportion under varying signal strength and $\alpha$ . . . . . 267
D.3 Input proportion of aSGL under varying weights . . . . . 268
D.4 Screening metrics for interaction data under the logistic model . . . 271
D.5 Improvement factor under varying data parameters for logistic model 271
D.6 Input proportion under varying data parameters for logistic model . 272

D.7 Improvement factor under varying signal strength and $\alpha$ for logistic model . . . 272
D.8 Input proportion under varying signal strength and $\alpha$ for logistic model 272

E.1 Bayesian parameters for illustrative example for BSGS-$\alpha$ . . . 300
E.2 The log-likelihood for the models in the illustrative example . . . 300
E.3 MAE($\sigma$) of BSGS under different $\boldsymbol{\beta}$ initialisations . . . 301
E.4 Bias of SLOBE models under different $\boldsymbol{\beta}$ initialisations . . . 301
E.5 Metrics for BGSLOPE under different $\boldsymbol{\beta}$ initialisations . . . 302
E.6 Distribution of learned $\alpha$ values for BSGS-$\alpha$ . . . 303
E.7 Coin flip property test for SGL and SGS . . . 305
E.8 $F_1$ scores for all model selection approaches . . . 306
E.9 MAE($\sigma$) for all model selection approaches . . . 307
E.10 FDR and power of other approaches under varying signal . . . 308
E.11 FDR and power of other approaches under varying data parameters 308
E.12 Bias of all approaches under varying data parameters . . . 309
E.13 OOS of all approaches under varying data parameters . . . 309
E.14 Runtime for all model selection approaches as function of dimensionality 310
E.15 FDR and power of other approaches under varying correlation . . . 310
E.16 Bias of all approaches under varying correlation . . . 311
E.17 OOS of all approaches under varying correlation . . . 311
E.18 Variable and group FDR for CV and BSGS . . . 312
E.19 Metrics under equal groups . . . 313

# List of Tables


3.1 Metrics for models under random signal . . . . . 41
3.2 Metrics for SLOPE-based models under large groups . . . . . 42
3.3 Number of false positives for all models . . . . . 43
3.4 Peak classification rate for all models for real data . . . . . 45
3.5 Active genes found by SGS for Colitis data . . . . . 47

4.1 Averaged cardinality of screened sets for synthetic and real data . . 72
4.2 Average runtime for screen vs no screen for synthetic study . . . . . 73
4.3 Dataset information for real data study . . . . . 74

5.1 Figure 5.2 runtime . . . . . 93
5.2 Scheetz runtime (Figure 5.9) . . . . . 93
5.3 Summary of SGL screening rules . . . . . 97
5.4 Improvement factor for CV . . . . . 103
5.5 Dataset information for real data study . . . . . 105

7.1 Summary of approaches used in synthetic study . . . . . 150
7.2 Key variable metrics averaged over all simulation cases . . . . . 158
7.3 Dataset information for real data study . . . . . 160
7.4 NMSE for approaches on the real data . . . . . 161
7.5 NMSE for BSGS variants on the real data . . . . . 162

A.1 General notation used in the thesis . . . . . 199
A.2 Chapter-specific notation . . . . . 200
A.3 Common acronyms used in the thesis . . . . . 201
A.4 Algorithmic hyperparameters . . . . . 203
A.5 Base case data hyperparameters . . . . . 203
A.6 Model hyperparameters . . . . . 204

B.1 Metrics averaged over all cases for all models in synthetic study . . . 244
B.2 Metrics for lasso-based models under large groups . . . . . 244
B.3 Top ten active genes found by SGS for Cancer . . . . . 246
B.4 Number of selected features for Cancer and Colitis data . . . . . 246
B.5 Peak classification rate for SGS applied to all genesets . . . . . 247

C.1 Variable screening metrics for SGS for synthetic study . . . . . 256
C.2 General and group screening metrics for synthetic study . . . . . 257
C.3 Variable metrics for SGS on real data . . . . . 260
C.4 General and group metrics for real data . . . . . 261

D.1 Group screening metrics corresponding to the GAP safe simulations 268
D.2 Variable screening metrics corresponding to the GAP safe simulations 269
D.3 Model fitting metrics corresponding to the GAP safe simulations . . 269
D.4 Group screening metrics for other linear model simulations . . . . . 270
D.5 Variable screening metrics for other linear model simulations . . . . 270
D.6 Model fitting metrics for other linear model simulations . . . . . 270
D.7 Group screening metrics corresponding to the logistic model . . . . . 273
D.8 Variable screening metrics corresponding to the logistic model . . . . 273
D.9 Model fitting metrics corresponding to the logistic model . . . . . 273
D.10 Group screening metrics corresponding to the real data . . . . . 274
D.11 Variable screening metrics corresponding to the real data . . . . . 275
D.12 Model fitting metrics corresponding to the real data . . . . . 276

E.1 Mean number of iterations for the $\boldsymbol{\beta}$ initialisations . . . . . 302
E.2 Mean MSE($\boldsymbol{\beta}$) and MAE($\sigma$) values for the initialisation schemes . . 303
E.3 Mean MSE($\boldsymbol{\beta}$) and MAE($\sigma$) values for the Beta prior schemes . . . . 303
E.4 Metrics for short synthetic study . . . . . 305
E.5 Key group metrics averaged over all simulation cases . . . . . 312
E.6 NMSE for SLOBE variants on real data . . . . . 313

# CHAPTER 1

## Introduction

In statistics, the focus is broadly on discovering signals and discarding noise by learning from the data available. By discovering signals, hidden relationships are uncovered that govern real-world processes. Gaining an insight into the hidden forces that shape real-world outcomes allows us to infer the underlying patterns and relationships that drive them, as well as predict these outcomes with accuracy. This is especially useful in the medical domain, where discoveries about the relationships between *genotypes* (genes) and *phenotypes* (observable characteristics, such as disease traits) enable the development of targeted treatments and drug therapies. Furthermore, these relationships can be used to predict a patient's risk of specific diseases, allowing preventative measures to be taken to reduce risk. However, finding signal is never trivial, and noise often finds ways to mask itself as signal. If a model is poorly designed, its inferences and predictions may be dominated by noise, resulting in inaccurate outcomes. It is extremely rare for there to be perfect separation between signal and noise. This is especially true in high-dimensional settings, where the curse of dimensionality rears its ugly head.

Extracting as much signal as possible from a dataset while eliminating noise is referred to as variable selection in statistics, a task that is challenging. This problem is traditionally addressed using penalised regression, in which variables are shrunk towards zero, effectively setting some to be inactive. Generally, it is accepted that no matter how much information is available, it is unlikely that penalised regression models only capture signals without noise. Instead, the focus shifts to reducing noise (or the error rate) in the model, providing stronger guarantees on the accuracy of discovered associations and ensuring that predictions are driven by true signal rather than by noise.

Reducing the error rate is particularly important in genetics. Statistical associations do not necessarily correspond to true biological effects, so costly verification studies are often required. By controlling the error rate of a list of statistical associations, follow-up studies can be conducted more cost-effectively. To this end, the Sorted $\ell_1$ Penalised Estimation (SLOPE) model was proposed, which provides guarantees on the error rate of its selected variables.

However, like many penalised models, SLOPE does not fully utilise all available information, which is crucial for accurately identifying true associations. In many real-world applications, predictors exhibit a grouping structure; in genetics, genes are members of pathways, which are functional groups of genes working together to achieve a biological task. Pathway analysis has attracted considerable attention in the scientific community, as examining sets of genes is often regarded as more powerful and biologically informative than traditional univariate analysis [188]. When conducting pathway analysis of genetics data, the interest is in identifying both genes and pathways associated with a change in the risk profile of a disease. Evangelou et al. [58] illustrated how genes discovered through pathway analysis have often been missed from conventional analyses, even though they can have an important biological role in the development of a disease.

This grouping information is often used for group selection or, perhaps more importantly in genetics, to enhance variable selection by making use of the information contained within the grouping structure. In this thesis, we present new methodologies designed to exploit grouping or pathway information to detect signals in genetics data, while providing theoretical guarantees on the accepted amount of noise (also called *false discoveries*). These methods are developed in a general statistical framework to ensure broad applicability, with each chapter concluding by demonstrating their use on real genetic data. While grounded in rigorous theory, the research is driven by practical considerations, with a strong emphasis on accessibility, aiming to make these tools readily usable for uncovering biological insights.

To allow grouping information to be used with SLOPE, we introduce the sparse-group SLOPE (SGS) model, which combines the benefits of SLOPE and sparse-group norms to provide bi-level selection with false discovery rate (FDR) control. Bi-level FDR control is achieved by developing new penalty sequences that guarantee concurrent FDR control at both the variable and group levels. SGS is applied to predicting disease cases and is shown to outperform existing methods.

For high-dimensional data, only a small proportion of variables are typically rel-

evant. Penalised regression models, particularly sparse-group models such as SGS, can be computationally demanding when applied to large datasets. To make SGS practical for large-scale genetic studies, we develop strong screening rules that remove inactive predictors before model fitting, resulting in substantial computational savings. Screening rules are also developed for the group SLOPE (gSLOPE) model. This work introduces the first sparse-group strong screening framework, which is further extended to the sparse-group lasso (SGL) through the Dual Feature Reduction (DFR) approach.

Another challenge with penalised regression models is model selection, as several models are typically fit along a regularisation path. The model is often chosen using cross-validation, but this approach prioritises predictive performance, whereas for SLOPE models, controlling the FDR is of greater importance. To address this, we consider the general problem of model selection for SLOPE models and develop new Bayesian methods for group-based SLOPE variants using spike-and-slab priors. These Bayesian models provide uncertainty quantification, adapt to unknown sparsity structures, and allow flexibility in incorporating prior beliefs. Extensive simulation and real data studies show that the Bayesian approach maintains FDR control under general settings while achieving higher power and predictive accuracy compared to existing methods. Together, these results demonstrate that the Bayesian approach provides the optimal implementation for SLOPE models under general designs.

## 1.1 Outline of the Thesis

The thesis is structured as described below. Each chapter includes both synthetic and real data studies, with the majority of the real datasets coming from genetics. The SLOPE-based developments in each chapter also apply to the wider Ordered Weighted $\ell_1$ (OWL) class of models. The methodologies are primarily developed for SGS but are often first formulated for group SLOPE as a stepping stone.

- **Background (Chapter 2):** Provides background and literature review materials on penalised regression models, including the lasso, SLOPE, and group-based methods.
- **Chapter 3:** Introduces the sparse-group SLOPE model, proven to have bi-level FDR control. The model is fit using a proximal algorithm and applied

to disease prediction tasks.

- **Chapter 4:** Develops the first sparse-group strong screening framework in the literature, deriving gSLOPE and SGS screening rules from their subdifferentials and achieving substantial computational savings.
- **Chapter 5:** Presents the Dual Feature Reduction (DFR) screening method for SGL and aSGL, based on the sparse-group strong screening framework. DFR is shown to be the state-of-the-art SGL screening approach.
- **Model selection background (Chapter 6):** Provides additional background relevant for Chapter 7, including spike-and-slab models, Knockoffs, and noise estimation procedures.
- **Chapter 7:** Tackles the problem of model selection under general designs for SLOPE models. Introduces new spike-and-slab Bayesian methodology for group-based SLOPE models.
- **Discussion (Chapter 8):** Provides an overview of the main contributions and limitations of this thesis, and outlines potential directions for future research to address these challenges.

## 1.2 Code Availability

Two R packages were developed to implement the methods in this thesis (both are available on CRAN):

- `sgs` R package [64]: Implementation of SGS models. Linear and logistic regression models are supported, both of which can be fit using k-fold cross-validation. In addition, a general ATOS implementation is provided. Group SLOPE and group-based OSCAR models are also implemented. All models are available with strong screening rules for computational speed-up. The package was used to fit group-based SLOPE and OSCAR models in Chapters 3, 4, and 7.
- `dfr` R package [65]: Implementation of DFR for SGL and aSGL. DFR is implemented using the ATOS algorithm, with linear and logistic SGL models supported, both of which can be fit using k-fold cross-validation. The package was used to fit SGL and aSGL models using DFR in Chapter 5.

# CHAPTER 2

# Background

This section provides key background material relevant to the entire thesis, reviewing penalised regression methods such as the lasso, SLOPE, and group-based models.

**A note on notation.** We introduce notation as we go, but for ease of reference, we have summarised the key notations and acronyms in Tables A.1 and A.3. For chapter-specific notation, see Table A.2. We adopt the following notational convention: all vectors and matrices in this thesis are bold. Generally, variable indices are represented by $i$ and group ones by $j$. Similarly, any operator has a subscript $v$ when applied to variables and a subscript $g$ for groups.

## 2.1 Penalised Regression

We begin by formally defining the regression task and the penalised formulation of the problem. Suppose we have $n$ observations and $p$ variables, then we are interested in exploring the relationships between a continuous response, $\mathbf{y} \in \mathbb{R}^n$, and a design matrix, $\mathbf{X} \in \mathbb{R}^{n \times p}$. A common approach is to fit a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ are the regression coefficients and $\boldsymbol{\epsilon} \sim \boldsymbol{\mathcal{N}}_n(\mathbf{0}, \mathbf{I}_n\sigma^2)$ the residuals, such that $\sigma^2 > 0$ denotes the noise level of the model. The relationship between $\mathbf{X}$ and $\mathbf{y}$ is encoded via the model choice (in this case linear) and the entries of $\boldsymbol{\beta}$. Without loss of generality, we assume the coefficients and response are mean-centred, so the intercept can be omitted.

Finding relationships between $\mathbf{X}$ and $\mathbf{y}$ corresponds to discovering variable associations, *i.e.*, the entries of $\boldsymbol{\beta}$ that are non-zero, given by $S_v := \operatorname{supp}(\boldsymbol{\beta}) = \{i \in [p] :$

$\beta_i \neq 0\}$, where $[p] := \{1, \ldots, p\}$. This is more widely known in statistics as *variable selection*.

As an example, a researcher in genetics may wish to determine which genes, represented by the columns of $\mathbf{X}$, are associated with a measure of disease severity captured by $\mathbf{y}$. In this case, the variables selected correspond to the genes that have an association with the severity.

In an ideal world, we would discover the *Markov blanket* (Definition 2.1). However, as alluded to earlier with our discussion on noise infiltrating the signal, in most situations, there is little hope in recovering the blanket without some error. Instead, as is common in variable selection literature, we do not attempt to recover the blanket but instead aim to find an approximation of $S_v$ with guarantees on the error rate.

**Definition 2.1** (Markov blanket)**.** A *Markov blanket* of a random variable $\mathbf{y}$ in a set of random variables $\mathbb{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_p\}$ is any subset $\mathbb{X}_1 \subseteq \mathbb{X}$ such that $\mathbf{y} \perp\!\!\!\perp \mathbb{X} \setminus \mathbb{X}_1 \mid \mathbb{X}_1$. In other words, $\mathbb{X}_1$ contains all information needed to predict $\mathbf{y}$.

In this context, two fundamental types of statistical errors are relevant: a *type I* error occurs when a null variable ($\beta_i = 0$) is incorrectly identified as active, while a *type II* error occurs when a truly active variable ($\beta_i \neq 0$) is missed. Often, one type of error is avoided at the cost of another, so balancing these errors is central to variable selection procedures.

To perform variable selection, $p$ hypotheses are tested: $H_i : \beta_i = 0$ for $i \in [p]$. For these tests, we are interested in controlling the type I error. More formally, we wish to control the *False Discovery Rate* (FDR) (Definition 2.2), by keeping it below a specified limit. The FDR metric has widespread use in the multiple testing literature, as spurious discoveries increase in probability as a function of the number of tests being performed. Other useful metrics in multiple testing include *Power* and the $F_1$ *score*; see Definitions A.1 and A.2.

**Definition 2.2** (False discovery rate)**.** Suppose a variable selection procedure returns an approximated support set $\hat{S}_v \subseteq [p]$, then the *False Discovery Rate* (FDR) is defined as

$$\text{FDR} = \mathbb{E}\left[\frac{|\hat{S}_v \setminus S_v|}{\max(1, |\hat{S}_v|)}\right].$$

The numerator counts the number of false discoveries (type I errors), so that the FDR metric quantifies the expected rate of false discoveries.

Controlling the FDR is of particular importance in genetic applications, where statistical associations must often be validated through subsequent biological experiments. Such experiments are typically costly and time-consuming; therefore, ensuring that the pool of selected associations contains no more than a specified proportion of false discoveries is crucial for enabling efficient follow-up studies.

In a low-dimensional setting ($n > p$), inference on $\boldsymbol{\beta}$ often occurs through the use of the (unbiased) ordinary least squares (OLS) estimator

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} \sim \boldsymbol{\mathcal{N}}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}).$$

Using the OLS estimator, the fitted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}$ can be considered as an orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$. This perspective highlights that OLS decomposes $\mathbf{y}$ into two orthogonal components: the fitted component $\hat{\mathbf{y}}$ lying in the column space of $\mathbf{X}$ and the residual component $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$ lying in the orthogonal complement. Historically, low-dimensional settings were the most common and much of classical statistical theory assumes that $p$ is fixed as $n \to \infty$. Under this regime, many useful results hold, such as the Law of Large Numbers and the Central Limit Theorem, which provide the foundation for deriving asymptotic properties of parameter estimators [102].

However, as data collection rates have increased dramatically, it has become very common to encounter $p \gg n$, also known as the *high-dimensional setting*. This is particularly prevalent in genetics, which provides the primary motivation for our work. In a genetics study, the number of measured genes often far exceeds the number of individuals. More generally, high-dimensional datasets are increasingly common across many fields due to advances in data collection technologies. Under this setting, the matrix $\mathbf{X}$ is no longer of full rank, and as a result, the OLS estimator is not unique. This leads to the model being unidentifiable [85]. Consequently, alterations to the traditional regression approach are required.

A widely used resolution is to impose *sparsity* on the feature space, which shrinks coefficients towards zero, effectively performing variable selection while fitting the model. Sparsity can be interpreted geometrically as projecting the response $\mathbf{y}$ onto a lower-dimensional subspace spanned by a subset of the columns of $\mathbf{X}$. This reduces the effective dimensionality of the model, resolving the identifiability issue. Additionally, it also improves interpretability and reduces variance, but in turn introduces bias. However, procedures exist to debias the coefficients [130].

This leads to the penalised regression framework, given in Lagrangian form by

$$\hat{\boldsymbol{\beta}}(\lambda) \in \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min} \left\{ f(\mathbf{b}) + \lambda J(\mathbf{b}; \mathbf{v}) \right\}, \tag{2.1}$$

where $f$ is a differentiable and convex loss function, $J$ is a sparsity-inducing convex penalty norm with weights $\mathbf{v} \succeq 0$, and $\lambda > 0$ is the regularisation parameter. In the linear case, which is adopted throughout most of this thesis, the loss function is $f(\mathbf{b}) = \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2$, where the $1/2n$ factor is a reparameterisation of $\lambda$ that makes it comparable across sample sizes. This formulation proves particularly useful when tuning $\lambda$ using *cross-validation* (CV) [85]; a data-splitting technique that helps in selecting optimal hyperparameter values while assessing model performance. The parameter $\lambda$ determines the level of sparsity: larger values encourage more variables to be shrunk toward zero. In practice, $\lambda$ is often selected along a path $\lambda_1, \ldots, \lambda_l$, producing a sequence of models that trade-off sparsity and goodness-of-fit.

### 2.1.1 Ridge and the Lasso

Two commonly used convex penalty norms in Equation 2.1 are *Ridge* [88] and the *least absolute shrinkage and selection operator* (lasso) [192]. Ridge regression uses the $\ell_2$ norm, $J_{\text{ridge}}(\mathbf{b}) = \|\mathbf{b}\|_2^2$, which penalises large coefficients but does not induce exact sparsity; all variables typically remain in the model with shrunken values, so no variable selection is performed. The ridge solution can be interpreted in terms of OLS by $\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}$ and can be seen as projecting $\mathbf{y}$ onto a shrunk column space, where high-variance directions dominate and low-variance directions are dampened. This reduces variance but introduces a small bias toward zero.

The lasso is defined by the $\ell_1$ norm, $J_{\text{lasso}}(\mathbf{b}) = \|\mathbf{b}\|_1$. It is among the most widely used penalised regression methods, as the non-differentiability of its $\ell_1$ norm enables coefficients to be shrunk exactly to zero, thereby performing variable selection [192]. At zero, the subdifferential spans $[-1, 1]$, so that $\beta_i = 0$ when the gradient is small (*i.e.*, the feature is weakly correlated with the residual). In contrast, ridge does not allow exact zeros because its gradient is smooth and continuous.

The lasso embodies the *betting on sparsity principle* [84], which states that $\ell_1$ methods work well when the true model is sparse, but if it is dense, no method can recover it without extensive data and overparameterisation. Another way to view this is that we must assume the true number of signals is substantially smaller than $p$, so that some form of projection into a lower-dimensional subspace is feasible (*i.e.*,

a sparse representation can be found). In genetic studies, this principle is justified by the common belief that only few variants are disease-associated [83, 218].

Further properties of the lasso include that it can select at most $n$ variables, it reduces variance at the cost of bias, and it is computationally cheap and simple to implement. The lasso estimator is unique with probability one when the predictor variables are drawn from a continuous distribution, irrespective of the dimensionality of the problem [194].

There are many approaches to solving the lasso optimisation problem, including coordinate-wise descent [71], Fast Iterative Shrinkage-Thresholding Algorithm (FISTA; a proximal algorithm) [21], and least-angle regression (LARS) [54], which fits the full solution path.

Despite the success of applying the lasso across many fields, it is inconsistent under certain scenarios [239]. Over the years, several extensions of the lasso have been proposed to address its limitations. One such method is the adaptive lasso, which assigns adaptive weights to each coefficient's penalty, improving variable selection consistency [239]. Another is the elastic net, which combines ridge and lasso penalties in a convex combination to exploit the strengths of both norms [240].

However, a key limitation of interest in this thesis is the lack of guarantees for FDR control in the lasso. Not only does the lasso provide no formal assurances of error-rate control, but Su et al. [187] also demonstrate that false discoveries tend to occur early along the lasso path. This indicates that signal and noise variables are consistently interspersed throughout the path, regardless of signal strength. The authors further establish a trade-off between type I and type II errors, revealing that regions of simultaneously high power and low false discoveries are unattainable. Further, Fan and Li [59] noted that the lasso's $\ell_1$ penalty overshrinks large coefficients, causing it to include unimportant variables to compensate, which increases false positives.

Another perspective on false discoveries comes from examining the residuals. In a lasso model, the residuals can be seen as projecting $\mathbf{y}$ onto the lasso null polyhedron [171]. The residuals retain effects from selected variables, known as shrinkage noise, which amplifies as more strong predictors are added, projecting onto null directions and potentially overpowering true signals [187]. This can lead to the selection of null variables, causing false discoveries.

As a result, the lasso is unsuitable for our problem of interest. This thesis focuses on an extension of the lasso that addresses FDR control: SLOPE, introduced next.

## 2.2 Sorted $\ell_1$ Penalised Estimation (SLOPE)

An important extension of the lasso that addresses FDR control in variable selection is the SLOPE model, which is a key method of this thesis. Before introducing SLOPE, we first describe the broader family of OWL models, of which SLOPE is a special case.

### 2.2.1 OWL Models

The *Ordered Weighted* $\ell_1$ (OWL) family of models includes those with penalty functions given by

$$J_{\text{owl}}(\mathbf{b}; \mathbf{v}) = \sum_{i=1}^{p} v_i |b|_{(i)}, \tag{2.2}$$

where $|b|_{(1)} \geq \cdots \geq |b|_{(p)}$ are the order statistics of the absolute coefficients and $v_1 \geq \cdots \geq v_p \geq 0$ are non-increasing weights. OWL models offer valuable adaptivity by assigning larger penalty weights to larger coefficients, which enhances their selection properties, especially in scenarios with highly correlated predictors [230].

An example of an OWL model is given by the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) model [25]. For OSCAR, the weights are defined as $v_i = a_1 + a_2(p-i)$, where $a_1, a_2$ are to be specified by the user. They are generally set to $a_1 = b_i \|\mathbf{X}^\top \mathbf{y}\|_\infty$ and $a_2 = a_1/p$, where $b_i = i \cdot \exp\{-2\}$ [12, 143, 238]. The OSCAR weights are linearly decaying values. As implied by its name, OSCAR is particularly effective at clustering highly correlated variables [230]; this improves upon a key weakness of the lasso, which tends to arbitrarily select one variable from a set of highly correlated ones, while setting the others to zero [25].

The methods developed in this thesis are primarily for SLOPE models, but are applicable in all cases to the wider class of OWL models (simply by changing the weights).

### 2.2.2 SLOPE

The *Sorted* $\ell_1$ *Penalised Estimation* (SLOPE) model is an important special case of OWL. SLOPE bridges the gap between the lasso and FDR control in multiple testing by applying the same sorted $\ell_1$ penalty as OWL (Equation 2.2) with the weights designed around FDR control.

The weights are defined through a connection to a widely used approach in multiple testing: the Benjamini-Hochberg (BH) procedure [22], which ranks the $p$-values and rejects those below a threshold that controls the expected proportion of false discoveries at a specified level. The connection between SLOPE and the BH procedure arises through the choice of the penalty sequence in SLOPE, given by

$$v_i = \Phi^{-1}\left(1 - \frac{q_v i}{2p}\right), \quad i \in [p], \tag{2.3}$$

where $q_v \in (0, 1)$ is the target FDR level, $\Phi$ is the cumulative distribution function of the standard normal distribution, and $\Phi^{-1}$ denotes its quantile function. The sequence is termed the *BH sequence* and is shown in Figure 2.1. Under orthogonal input designs, *i.e.*, $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$, this sequence enables SLOPE to control the FDR at level $q_v$ [24]. SLOPE, when applied with this sequence, was asymptotically shown to have the FDR converging to zero and power converging to one under Gaussian design matrices, given certain conditions [106]. Any use of SLOPE in this thesis will use the BH sequence, unless otherwise stated.

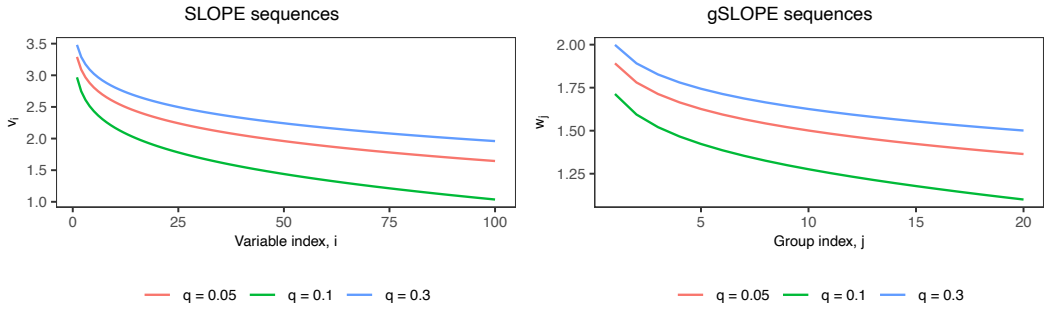



Figure 2.1: SLOPE BH sequence for different values of $q_v$ (left) and gSLOPE max sequence for different values of $q_g$ (right) under $m = 20$ unequal groups of sizes $[3, 7]$ with $p = 100$.

*Remark* 2.1 (SLOPE–BH connection). Under a linear model with an orthogonal design, we have $\tilde{\mathbf{y}} = \mathbf{X}^\top\mathbf{y} = \boldsymbol{\beta} + \mathbf{X}^\top\boldsymbol{\epsilon} \sim \boldsymbol{\mathcal{N}}_n(\boldsymbol{\beta}, \mathbf{I}_n\sigma^2)$. In a traditional regression setting, we would use the z-score to perform inference [84]:

$$z_i = \frac{\beta_j}{\sigma\sqrt{v_j}} \sim \mathcal{N}(0, 1), \quad \text{where} \quad v_i = (\mathbf{X}^\top\mathbf{X})^{-1}_i, \; i \in [p].$$

In the orthogonal SLOPE setting, inference is directly on $\tilde{\mathbf{y}}$, which has a mean of

$\boldsymbol{\beta}$. Additionally, as $\mathbf{X}$ is orthogonal, we have that $v_j = 1$. Therefore, the null distribution is given by $\tilde{\mathbf{y}}/\sigma \sim \boldsymbol{\mathcal{N}}_n(\mathbf{0}, \mathbf{I}_n)$. Under a two-sided test with significance level $q_v/2$ in each tail, the null is rejected if $\tilde{y}_i/\sigma \geq \Phi^{-1}(1 - q_v/2)$. Applying the BH adjustment with $\sigma = 1$ to these multiple tests gives rise to the SLOPE penalty sequence.

SLOPE reduces to the lasso under constant weights $v_1 = \cdots = v_p$. Unlike the adaptive lasso, which tends to apply smaller penalties to larger coefficients, SLOPE penalises larger coefficients more, which, in principle, makes it more difficult for noise to enter the model. Another key difference between the adaptive lasso and SLOPE lies in the form of their penalties. The (adaptive) lasso uses a separable penalty, whereas SLOPE employs a non-separable one, and as a result, they are fitted using different optimisation algorithms (Remark 2.2).

*Remark* 2.2 (Non-separability of SLOPE). Due to their sorting procedures, which induce dependence among the variables, OWL models (including SLOPE) are non-separable. That is, the penalty cannot be expressed as the sum of individual penalty terms: $J(\mathbf{b}) \neq \sum_{i=1}^{p} J_i(b_i)$, where $J_i$ acts only on the individual feature. An example of a separable penalty is the lasso. Most descent algorithms are not guaranteed to find the global optimum for non-separable penalties. Therefore, proximal algorithms are used instead, as they update all coordinates simultaneously. This motivated the modification of a proximal algorithm for SGS in Chapter 3; see Section 3.3.1.

Beyond FDR control, SLOPE possesses other useful statistical properties: it achieves the minimum total squared error over a range of sparsity classes without requiring prior knowledge of the true sparsity level, is asymptotically minimax, and clusters similar features [70, 171, 186]. SLOPE induces clustering by assigning identical absolute values to certain features, making it useful for highly correlated designs; a setting in which the lasso struggles [98]. The clustering behaviour stems from the geometry of its dual norm, whose unit ball is a signed permutahedron that partitions the input space in an ordered manner, thereby naturally forming feature clusters [171]. This geometry is used to describe conditions for the uniqueness of the SLOPE solution in Schneider and Tardivel [171].

As with the lasso, SLOPE cannot simultaneously achieve small type I and type II errors. However, Bu et al. [32] demonstrated that SLOPE is capable of attaining arbitrarily high power, thereby breaking the Donoho–Tanner (DT) power limit that constrains the lasso. Moreover, there exist settings in which the SLOPE trade-off curve is strictly better than that of the lasso.

SLOPE has found extensive use with genetics data due to its FDR control property [24, 73, 80], as well as in fields such as machine learning and finance [108, 158, 197].

As part of the investigations in this thesis, it was discovered that SLOPE is also connected to e-values, which are a different approach to FDR control. An *e-value* is a random variable $E \in [0, \infty]$ satisfying $\mathbb{E}[E] \leq 1$, such that a null hypothesis is rejected if it exceeds a threshold [198]. In contrast, a *p-value* is a random variable $P \in [0, 1]$ such that $\mathbb{P}(P \leq t) \leq t$ for any $t \in [0, 1]$. For both, the probability of making a type I error at level $\alpha$ is less than or equal to $\alpha$.

The average of multiple e-values is itself an e-value, making them useful alternatives to p-values, as they can be easily combined to aggregate results. This property is particularly valuable for derandomising procedures such as Knockoffs and for high-dimensional settings where valid p-values may be difficult or impossible to construct [155]. p-values can be converted to e-values via a p-to-e calibrator function $f : [0, 1] \to [0, \infty]$. We discovered that the SLOPE BH penalty sequence is in fact a p-to-e calibrator that takes p-values $1 - q_v i/2p$ and outputs a decreasing sequence in the range $[0, \infty]$. This discovery is surprising, given that the sequences were proposed prior to the development of e-values.

The BH multiple testing adjustment has been extended to operate on e-values rather than p-values, resulting in the eBH procedure, which controls the FDR with no correction for any dependence structure among the e-values [209]. Ignatiadis et al. [96] demonstrated that the eBH procedure is universal for FDR control; specifically, any FDR-controlling method can be expressed as the eBH procedure applied to an appropriate set of compound e-values.

### 2.2.3 Other FDR Control Approaches

By assuming an orthogonal design matrix, FDR control was proved for SLOPE in Bogdan et al. [24] and group SLOPE (Section 2.3) in Brzyski et al. [30]. SLOPE models are motivated by the BH procedure (see Remark 2.1), which provides FDR control through p-values and assumes that the tests generating these p-values are independent. The BH procedure is more conservative than SLOPE, leading to lower power, as well as requiring the computation of p-values [24]. The ideas of BH have been adapted to group settings using weighted p-values in Hu et al. [91] and Koner et al. [105].

SLOPE was adapted from the lasso to control the FDR. Another adaptation,

the debiased lasso, has been shown to control the (directional) FDR and achieve asymptotic power of one under conditions where the design matrix has sub-Gaussian rows, along with certain assumptions about the true number of signals [97]. The debiased lasso adjusts the lasso solution using a debiasing matrix to control the FDR, although it can be overly conservative.

FDR control under the most general setting is provided by the Knockoffs procedure [14] (described in detail in Chapter 6), which requires only that the distribution of the predictors is known. The procedure has been extended to the group setting in Dai and Barber [46], but there are no results specific to sparse-group settings. Knockoffs have been implemented to provide multilayer FDR control in Katsevich and Sabatti [103] and Gablenz and Sabatti [74], enabling FDR control at different resolutions. However, joint FDR control is not guaranteed to hold simultaneously, and the procedures reuse data, which can lead to a loss of power.

## 2.3 Group-based Models

In many real-world settings, such as genetics with pathways, variables form groups. To incorporate the additional information from these groupings into penalised regression models, group models were developed. Group-only norms enable group selection, while sparse-group models use the grouping information to also improve variable selection.

To formally define group models, consider an $m$-partition of the input space, $\mathcal{G} = \{G_1, \dots, G_m\}$, containing $m$ non-overlapping groups of variable indices of sizes $p_1, \dots, p_m$. Then, the *Group Lasso* (gLasso) norm is given by $J_{\text{glasso}}(\mathbf{b}; \mathbf{w}) = \sum_{j=1}^{m} \sqrt{p_j} \|\mathbf{b}^{(j)}\|_2$, where $\mathbf{b}^{(j)} \in \mathbb{R}^{p_j}$ is a vector of the coefficients for group $j \in [m]$ [227]. The gLasso performs *group selection* with the aim of discovering $S_g := \{j \in [m] : \|\beta^{(j)}\|_2 \neq 0\}$. It is group selection consistent in a random design model for fixed $p$ under some conditions [7]. Further, the gLasso was proven to outperform the lasso under strong group sparsity (and other) conditions [92]. However, these conditions are generally difficult to satisfy under $p \gg n$, limiting the usefulness of gLasso.

SLOPE was also extended to group regression in the *Group SLOPE* (gSLOPE) model [30]

$$J_{\text{gslope}}(\mathbf{b}; \mathbf{w}) = \sum_{j=1}^{m} \sqrt{p_j} w_j \|\mathbf{b}^{(j)}\|_2. \tag{2.4}$$

The norm has ordered penalty weights $w_1 \geq \cdots \geq w_m \geq 0$ which are matched to $\sqrt{p_1}\|\mathbf{b}^{(1)}\|_2 \geq \cdots \geq \sqrt{p_m}\|\mathbf{b}^{(m)}\|_2$. gSLOPE was shown to control the group FDR under orthogonal designs, using the gSLOPE max sequence

$$w_i^{\max} = \max_{j \in [m]} \left\{ \frac{1}{\sqrt{p_j}} F^{-1}_{\chi_{p_j}} \left( 1 - \frac{q_g i}{m} \right) \right\}, \text{ for } i \in [m], \tag{2.5}$$

where $F_{\chi_{p_j}}$ is the cumulative distribution function of a $\chi$ distribution with $p_j$ degrees of freedom and $q_g \in (0, 1)$ is the group FDR control parameter. gLasso and gSLOPE reduce to the lasso and SLOPE, respectively, under singleton groups. The gSLOPE max sequence is shown in Figure 2.1.

*Remark* 2.3. In comparison to the SLOPE penalty sequence (Equation 2.3), the penalty sequence of gSLOPE is taken as the maximum over all groups, accounting for the uncertainty in grouping structure (when we do not have evenly sized groups). To understand this, note that in the FDR proof for SLOPE [24], a key step for deriving the sequence requires that, for $i \in [p]$ and where $r$ is the number of hypothesis rejections,

$$1 - F_{|y_i|}(v_r) \leq \frac{q_v r}{p} \implies v_r \geq \Phi^{-1} \left( 1 - \frac{q_v r}{2p} \right),$$

by the fact that $y_i \sim \mathcal{N}(0, 1)$. As all variables have the same distribution, we can allow the inequality to be an equality, obtaining Equation 2.3. The equivalent step in the gSLOPE FDR proof is, for $j \in [m]$,

$$1 - F_{\chi_{p_j}}(w_r p_j) \leq \frac{q_g r}{m} \implies w_r \geq F^{-1}_{\chi_{p_j}} \left( 1 - \frac{q_g j}{m} \right).$$

As the distribution of the groups is not the same, the maximum over all groups is taken, to ensure the inequality holds for all groups, obtaining Equation 2.5. This leads to the gSLOPE FDR control being overly conservative, whereas the SLOPE FDR control is exact. This was verified in synthetic results in Brzyski et al. [30], where the FDR control was not exact under uneven groups.

### 2.3.1 Sparse-group Models

A key limitation of group models is that activating a group forces all its variables to be active, implicitly assuming every variable in a relevant group carries signal. In practice, this assumption is rarely true and often introduces substantial noise. This is especially pronounced in genetics, where the proportion of relevant genes

in a pathway tends to be low [83]. In such cases, the aim is often to identify associated individual genes while using group information to boost detection power. As a result of using all variables in a group, group-only norms can exhibit poor predictive performance, as shown in Simon et al. [178] and in Chapters 3 and 7.

This limitation led to the development of *sparse-group models*, which apply both variable and group penalisation to generate *bi-level selection*, and are the core topic of this thesis. Bi-level models come in various forms, such as concave one-norm group and composite penalties; in this thesis, we focus on additive penalties.

Sparse-group models utilise the grouping information without the restriction of keeping all variables in a group active. The lasso model was extended to a sparse-group model through the *Sparse-group Lasso* (SGL) [178], given by

$$\hat{\boldsymbol{\beta}}_{\text{sgl}}(\lambda) \in \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min} \left\{ f(\mathbf{b}) + \lambda J_{\text{sgl}}(\mathbf{b}) \right\}, \tag{2.6}$$

$$\text{where} \quad J_{\text{sgl}}(\mathbf{b}) = \alpha\|\mathbf{b}\|_1 + (1-\alpha)\sum_{j=1}^{m}\sqrt{p_j}\|\mathbf{b}^{(j)}\|_2, \tag{2.7}$$

where $\alpha \in [0, 1]$ defines the balance between variable and group shrinkage. Since models that apply only group penalisation typically perform poorly in practice, it is often more informative to compare sparse-group models with variable penalisation models. This allows assessment of whether the grouping structure provides any meaningful inferential or predictive value. SGL was found to outperform both the lasso and gLasso when applied to predicting breast cancer cases using genomics data [178]

SGL has been extended to have adaptive penalisation using the adaptive SGL (aSGL) [132, 151], which applies adaptive penalties to achieve the oracle property in a double-asymptotic framework. Other sparse-group models include sparse-group versions of Smoothly Clipped Absolute Deviation (SCAD), Minimax Concave Penalty (MCP), and Exponential Penalty (EP) [33].

In the next chapter, we propose a new sparse-group model: *Sparse-group SLOPE* (SGS), which performs adaptive bi-level selection with FDR control.

# CHAPTER 3

## Sparse-group SLOPE: Adaptive Bi-level Selection with FDR Control

*The content of this chapter is largely based on our manuscript Feser and Evangelou [66]. The SGS model developed in this chapter is implemented in the* `sgs` *R package [64], available on CRAN.*

In high-dimensional settings where variables naturally form groups, it is often important to induce sparsity both at the group level and within groups. This chapter introduces the key method of this thesis: Sparse-group SLOPE (SGS), a new penalisation methodology that achieves bi-level selection, identifying both relevant variables and groups. Bi-level selection is widely applicable across scientific domains, particularly in genetics, where genes are organised into pathways; understanding which pathways and genes are associated with an outcome of interest is critical for disease understanding and prediction.

SGS extends the SLOPE and Group SLOPE norms to provide adaptive variable- and group-level selection, while controlling the FDR at both levels. A proximal splitting algorithm, Adaptive Three Operator Splitting (ATOS), is implemented to efficiently fit SGS models for linear and logistic regression.

Achieving bi-level FDR control for SGS required the derivation of new interdependent penalty sequences and associated norm results, as the standard SLOPE and gSLOPE penalties are not inherently compatible with sparse-group structures. Through extensive simulations with orthogonal designs, we demonstrate that SGS successfully maintains bi-level FDR control under a variety of scenarios.

Application to both synthetic and real gene expression datasets shows that SGS outperforms existing lasso- and SLOPE-based approaches in terms of selection ac-

curacy and predictive performance. In particular, SGS identifies both known and novel genes associated with colitis, producing parsimonious models with minimal noise and improved disease prediction accuracy.

## 3.1 Introduction

In genetic studies, we often face high-dimensional settings where the goal is to identify associations between a large number of genetic predictors and a continuous response. Controlling the FDR of such discoveries is crucial, as false discoveries can lead to misleading biological conclusions and wasted experimental resources on follow-up studies. Furthermore, genes naturally form biological groups or pathways, and incorporating this structure can improve model performance. However, none of the existing methodologies satisfy all key features required for such applications.

A common approach to variable selection in linear regression is through the lasso, which encourages sparsity by shrinking some regression coefficients to zero. While effective, the lasso treats each predictor independently and ignores any group structure among variables. To incorporate grouping information, gLasso is used instead, which selects whole groups. However, gLasso retains all variables within an active group, which can be problematic in genetics, as many genes within a pathway may be irrelevant. SGL addresses this by combining the lasso and gLasso penalties to achieve bi-level selection.

SGL was shown to outperform the lasso and gLasso for disease prediction in Simon et al. [178], but has no FDR guarantees, making its use for discovering associations in genetics limited. Aside from SGL, examples of sparse-group models in literature are limited, but include sparse-group versions of SCAD, MCP, and EP [33].

Genetics settings, via the existence of pathways, provide the perfect natural setting for applying sparse-group models. However, existing sparse-group models emphasise predictive performance without offering any theoretical guarantees on error control, on either the variable or group level.

Aside from the lack of error control, lasso-type models suffer from other limitations. They can select at most $n$ variables and exhibit inconsistent selection under certain conditions. Extensions such as the adaptive lasso and elastic net improve selection consistency and performance under correlated designs but still lack explicit error-rate control.

To directly link penalised regression models with FDR control, the SLOPE model was introduced. By assigning decreasing penalties to ordered coefficients, SLOPE achieves FDR control under orthogonal designs while adapting automatically to unknown sparsity levels. Building on this, the gSLOPE model extends SLOPE to grouped predictors, enabling control of the group-level FDR. These advances make SLOPE-based approaches particularly attractive for high-dimensional genetic applications, where both sparsity and group structure are key and FDR control is essential for reliable inference. However, gSLOPE suffers from the same issue as gLasso in that it must select all variables in an active group as active, leading to no guarantees on the variable FDR, as well as poor predictive performance.

Motivated by these shortcomings of current sparse-group and SLOPE models, in this chapter, we incorporate SLOPE into a sparse-group framework, achieving bi-level FDR control.

### 3.1.1 Contributions and Organisation of this Chapter

In this chapter, we introduce the Sparse-Group SLOPE (SGS) model (Section 3.2), which integrates the SLOPE penalty into a sparse-group framework, thereby unifying the strengths of both approaches. SGS combines the SLOPE and gSLOPE penalties by incorporating both sorting procedures into a convex combination, analogous to how SGL merges the lasso and gLasso, thereby achieving bi-level selection. This bi-level structure allows the SLOPE framework to overcome the limitation of selecting all variables in active groups.

Combining SLOPE and gSLOPE to form the SGS penalty is relatively trivial, but the challenges arise from developing a suitable fitting algorithm and providing bi-level FDR control. Due to the sparse-group structure of the SGS norm, which involves two sorting operations, the respective fitting algorithms for SLOPE and gSLOPE cannot be applied directly. Similarly, the block coordinate descent algorithm used for SGL is unsuitable, as SGS employs a non-separable penalty unlike SGL. Therefore, to fit SGS, a proximal splitting algorithm, Adaptive Three Operator Splitting (ATOS) (Section 3.3), is adapted to efficiently incorporate the existing SLOPE and gSLOPE proximal operators, using a derived transformation that enables the gSLOPE operator's direct application within ATOS.

Since the SLOPE and gSLOPE penalties are not inherently suited to sparse-group structures, applying them directly with SGS does not provably ensure FDR control at either the variable or group level; as confirmed by our experimental results

in Section 3.4.4. To address this limitation, Section 3.4 introduces new, interdependent penalty sequences that achieve bi-level FDR control for SGS, supported by both extensive theoretical results (Section 3.4.2) and numerical studies (Section 3.4.4). In particular, our main result, Theorem 3.1, is the first in the literature to concurrently prove bi-level FDR control for a penalised regression method. The sequences are designed to jointly account for the two concurrent types of penalisation. Novel norm results are developed for deriving new sequences that provide FDR control in a sparse-group setting.

Simulated data analysis further demonstrates that SGS outperforms existing competitive lasso- and SLOPE-based approaches in terms of selection performance (Section 3.5). SGS was found to achieve more accurate variable and group selection than these methods under various scenarios, including random signals, large groups, and under the null model.

Finally, the analysis of real genetic datasets evaluates SGS as a disease prediction tool for colitis and breast cancer using gene expression data (Section 3.6). SGS identified both known and new genes associated with colitis. By applying adaptive sparse-group penalisation, SGS produced models with minimum levels of noise, leading to improved disease prediction performance over the other tested approaches.

Overall, SGS offers a new method for identifying associations in high-dimensional data with grouped structures, providing bi-level FDR control and demonstrating strong predictive capability.

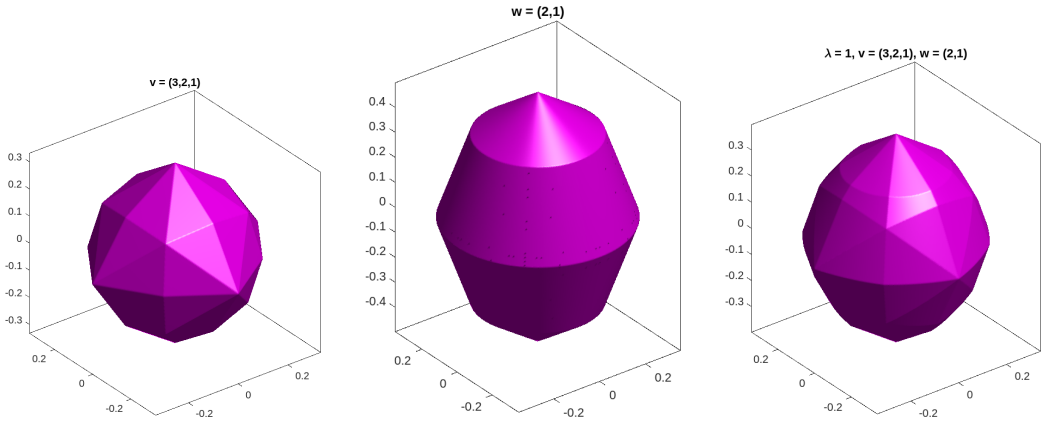
## 3.2 Sparse-group SLOPE

To incorporate the SLOPE concept into a sparse-group framework, SGS is defined as the solution to the convex optimisation problem given by

$$\hat{\boldsymbol{\beta}}_{\text{sgs}}(\lambda) \in \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda\alpha \sum_{i=1}^{p} v_i |b|_{(i)} + \lambda(1-\alpha) \sum_{j=1}^{m} w_j \sqrt{p_j} \|\mathbf{b}^{(j)}\|_2 \right\}, \tag{3.1}$$

which is defined for a linear model through the least squares loss function, but is also applicable for logistic regression (described in Appendix B.1.2). SGS uses adaptive penalty weights, with variable weights $\mathbf{v} = [v_1 \ \dots \ v_p]^\top$, where $v_1 \geq \cdots \geq v_p$ are matched to $|b|_{(1)} \geq \cdots \geq |b|_{(p)}$, and group weights $\mathbf{w} = [w_1 \ \dots \ w_m]^\top$, where $w_1 \geq \cdots \geq w_m$ are matched to $\sqrt{p_1}\|\mathbf{b}^{(1)}\|_2 \geq \cdots \geq \sqrt{p_m}\|\mathbf{b}^{(m)}\|_2$; the choice of

these weights is discussed in Section 3.4.3. SGS is a generalisation of existing high-dimensional penalisation approaches, including the lasso, gLasso, SGL, SLOPE, and gSLOPE, under certain hyperparameter choices.




(a) SLOPE (b) gSLOPE (c) SGS with $\alpha = 0.5$

Figure 3.1: Penalty unit balls in $\mathbb{R}^3$ for SLOPE (a), gSLOPE (b), and SGS (c).

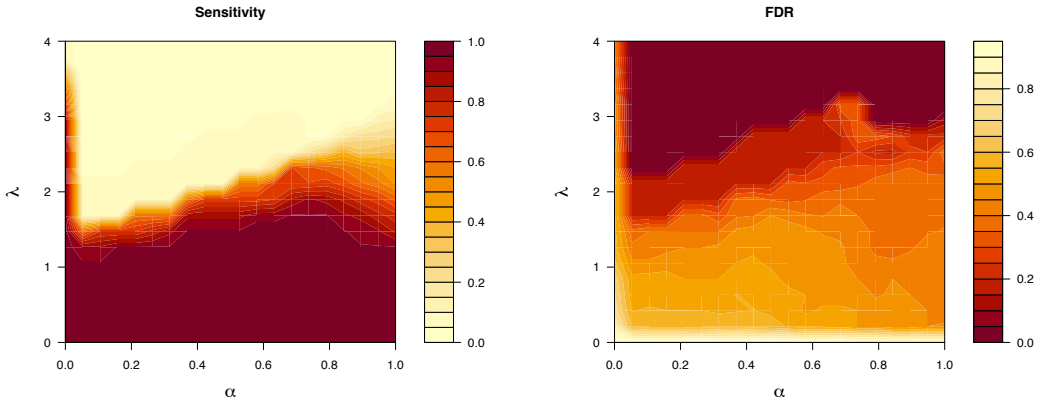



Figure 3.2: The trade-off between sensitivity and FDR as a function of the grid $(\lambda, \alpha)$; dark values are better. The data was generated as described in Remark 3.1.

SGS can be seen to be a convex combination of SLOPE and gSLOPE (Figure 3.1), balanced through $\alpha \in [0, 1]$, such that it reduces to SLOPE for $\alpha = 1$ and to gSLOPE for $\alpha = 0$. Simon et al. [178] recommends using $\alpha = 0.95$ for SGL, and this is the value we will use for most of the thesis when applying SGS. Figure 3.2 explores the tradeoff between FDR and sensitivity for SGS under different values of $\lambda$ and $\alpha$; an optimal region for $\alpha$ in $[0.9, 1]$ is observed in the trade-off diagram,

where the FDR does not increase but the sensitivity does. Further, the experiments in this chapter confirm that values in this region are optimal (Sections 3.5 and 3.6; in particular, Figure 3.10).

The regularisation parameter $\lambda > 0$ can be used to define a pathwise solution, as shown in Figure 3.3. This is discussed in Section 2.1 and explored in Chapter 7. The start of the path is defined as per Proposition 3.1.

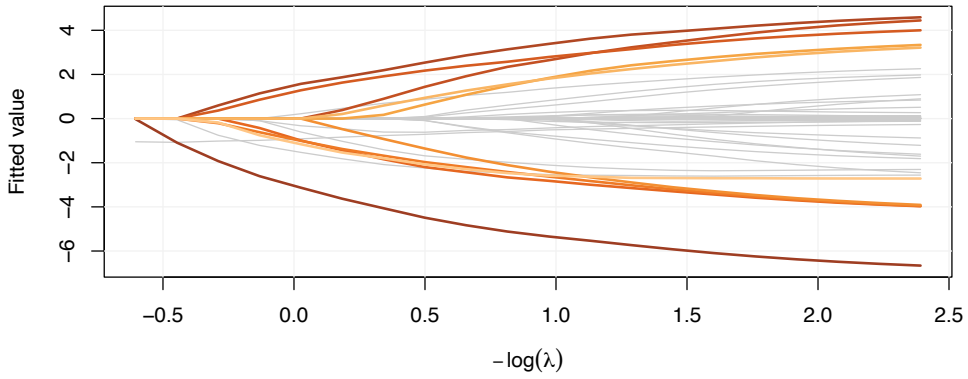



Figure 3.3: Pathwise fitted coefficient values for an SGS model for $\alpha = 0.95, q_v = 0.1, q_g = 0.1$, shown for the top ten largest coefficients (with the others shown in grey). The darkness of the lines represents the absolute magnitude of the coefficients. The data was generated as described in Remark 3.1.

**Proposition 3.1** (SGS path start)**.** *For SGS, the path value at which the first variable enters the model is given by*

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

*where the operator $(\cdot)_{\downarrow}$ sorts a vector into decreasing form, and $\boldsymbol{\tau}$ and $\boldsymbol{\omega}$ are expanded vectors of the group sizes ($\sqrt{p_j}$) and penalty weights ($w_j$) to $p$ dimensions.*

*Remark* 3.1 (Computational setup of the plots). For Figures 3.2 and 3.3, the computational setup was as in Section 3.5. A design matrix $\mathbf{X} \sim \boldsymbol{\mathcal{N}}_p(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^{200\times 800}$ was generated with correlation matrix $\boldsymbol{\Sigma}$, where $\Sigma_{i,j} = 0.3$, where $i \neq j$ and $i$ and $j$ belong to the same group. The response was generated using the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\sigma$ chosen adaptively so that the signal-to-noise ratio (SNR) was set at 6, defined as $\text{SNR} = \text{Var}(\mathbf{X}\hat{\boldsymbol{\beta}})/\hat{\sigma}^2$. The

signal of the active variables was set at $\beta \sim \mathcal{N}(0, 5^2)$. The variables were split into 160 non-overlapping groups of sizes $\{3, \dots, 7\}$. The group sparsity proportion, defined as the proportion of inactive groups, was set at 0.8. The variable sparsity proportion within an active group was set at 0.4. The SGS model used $\alpha = 0.95$ and $q_v = q_g = 0.1$ with the vMean and gSLOPE mean sequences (see Section 3.4.3). The data was $\ell_2$-standardised and an intercept fit.

### 3.2.1 Extensions

The SGS framework can be extended to the wider class of OWL models (Section 2.2.1) to define any Sparse-group OWL model by changing the weights. In particular, the framework can be used to define the *Sparse-group OSCAR* (SGO) model, with variable and group weights as, for $i \in [p]$ and $j \in [m]$,

$$v_i = a_1 + a_2(p - i), \quad w_j = a_1 + a_3(m - j),$$

where $a_3 = a_1/j$. The hyperparameters $a_1$ and $a_2$ are chosen as for OSCAR (Section 2.2.1).

## 3.3 Proximal Fitting Algorithm

The SGS penalty proposed in Equation 3.1 is convex (Theorem B.1). Both the SLOPE and gSLOPE penalties are non-separable (Remark 2.2) and data-dependent due to the sorting procedures [31]. As a result, the SGS penalty is non-separable at both the variable and group level. An advantage of non-separable penalties is their ability to share information across variables, enabling the detection of grouping structures, which makes them preferable for group regression settings [161].

Penalised regression models are typically optimised using either coordinate descent or proximal gradient algorithms. Descent methods iteratively reduce the objective cost by updating one variable at a time, exploiting separability of the penalty for efficiency. Examples of descent algorithms include coordinate descent [72, 90, 177] and Block Coordinate Descent (BCD) [152, 178] for group-based models.

In contrast, proximal algorithms are designed for composite objectives, which may involve non-smooth terms, combining a gradient step on the loss function with an explicit proximal mapping for the penalty. Examples of proximal approaches are given by Iterative Shrinkage-thresholding Algorithm (ISTA) [48], its acceler-

ated variant Fast ISTA (FISTA) [21], and operator-splitting approaches such as the Alternating Direction Method of Multipliers (ADMM) [27] and forward-backward splitting [146].

Coordinate descent approaches are not guaranteed to converge to the global optimum for a non-separable penalty, which means that BCD, which is used to fit SGL, is not suitable for use with SGS. Instead, to fit SLOPE-based models, proximal algorithms are used, which have no separability assumptions; although, recently, a coordinate descent algorithm was proposed for SLOPE, which uses its clustering property [112]. In proximal algorithms, the coordinates are updated simultaneously, in contrast to the cyclic updates used in gradient descent.

Both proximal and gradient descent algorithms have convergence rates of $\mathcal{O}(1/t)$ and time complexities of $\mathcal{O}(tp^2)$ [141, 147, 237], and benefit from efficient implementation in a pathwise setting using warm starts (which we utilise for SGS).

Proximal gradient algorithms solve optimisation problems of the form $\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$, where $f$ and $g$ are convex functions and $f$ is differentiable. Using a proximal algorithm requires access to the *proximal mapping*, given by [146]

$$\operatorname{prox}_g(\mathbf{x}) := \arg\min_{\mathbf{z}} \left\{ \frac{1}{2} \|\mathbf{z} - \mathbf{x}\|_2^2 + g(\mathbf{z}) \right\}.$$

The proximal mapping is sometimes written as $\operatorname{prox}_g(\mathbf{x}; \mathbf{v})$ to emphasise the dependence of the function $g$ on some weights $\mathbf{v}$. The mapping often has a closed form and depends only on $g$, so even if $f$ is complex, only its gradient is needed, making the computational cost primarily determined by $g$.

To use a proximal algorithm for SGS, the proximal mapping is required, which would be challenging to derive due to the two layers of non-separability. Instead, we exploit the fact that the mappings of SLOPE and gSLOPE are already known; given by Algorithms 7 and 8 and denoted by $\operatorname{prox}_{\text{slope}}$ and $\operatorname{prox}_{\text{gslope}}$, respectively. To do this, SGS is reformulated as a problem of the form

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{x}), \tag{3.2}$$

where $f$ is convex and $L_f$-smooth (differentiable with Lipschitz gradient), and both $g$ and $h$ are convex and proximal (*i.e.*, the proximal mapping is available). The function $f$ corresponds to the loss function, and the smoothness required is satisfied by the linear and logistic regression loss functions. The functions $g$ and $h$ are given

by the SLOPE and gSLOPE penalty norms.

### 3.3.1 Adaptive Three Operator Splitting (ATOS)

To solve Equation 3.2, a proximal splitting algorithm, *Adaptive Three Operator Splitting* (ATOS) [147], is used. ATOS requires only the evaluation of the gradient of $f$ and the proximal mappings of $g$ and $h$; all of which are available for SGS.

We first describe the non-adaptive version of ATOS, three operator splitting (TOS). TOS introduces two auxiliary variables, $\mathbf{y}$ and $\mathbf{z}$, to Equation 3.2 to reparameterise the optimisation problem as

$$\min_{\mathbf{x},\mathbf{y},\mathbf{z}} f(\mathbf{x}) + g(\mathbf{y}) + h(\mathbf{z}) \quad \text{s.t.} \quad \mathbf{x} \equiv \mathbf{y} \equiv \mathbf{z}.$$

This allows the problem to be broken down into three smaller (and simpler) subproblems, keeping the solutions to them as close as possible to each other. For a step size $\gamma > 0$, the update for iteration $t$ is

$$\begin{aligned}
\mathbf{z}_{[t]} &= \operatorname{prox}_h(\mathbf{y}_{[t]}), \\
\mathbf{x}_{[t+1]} &= \operatorname{prox}_g(2\mathbf{z}_{[t]} - \mathbf{y}_{[t]} - \gamma \nabla f(\mathbf{z}_{[t]})), \\
\mathbf{y}_{[t+1]} &= \mathbf{y}_{[t]} + \mathbf{x}_{[t+1]} - \mathbf{z}_{[t]},
\end{aligned}$$

where the subscript $[t]$ indicates the value of a variable at iteration $t$ [49]. This is a generalisation of two popular splitting approaches: when $h = 0$, we obtain ADMM, and for $g = 0$, the forward-backward proximal splitting algorithm [146].

ATOS modifies TOS in two key ways. First, it applies an adaptive search to the step size (described in Section 3.3.2). Second, it reformulates the optimisation task as a saddle point problem. Denoting $h^*$ as the convex conjugate of $h$, the problem becomes

$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{x}) = \min_{\mathbf{x}} \max_{\mathbf{u}} \left\{ f(\mathbf{x}) + g(\mathbf{x}) + \langle \mathbf{x}, \mathbf{u} \rangle - h^*(\mathbf{u}) \right\}.$$

The problem reduces to finding a saddle point, $(\mathbf{x}^*, \mathbf{u}^*)$, of the objective function, where $\mathbf{x}^*$ is the global minimum of the original optimisation problem [147]. ATOS is given in full detail, applied to SGS, in Algorithm 1 (the algorithm can also be used for SGO by using the OSCAR weights). From this, it is clear that ATOS recovers TOS by applying the transformation $\mathbf{y}_{[t+1]} = \mathbf{b}_{[t+1]} + \gamma_{[t]}\mathbf{u}_{[t]}$ and using a constant

step size.

The sorting of the variables and groups for proximal operators in Algorithm 1 is set according to the coefficients at the previous iteration or path point. The operators perform this sorting as part of their evaluation.

**Algorithm 1** ATOS for SGS

**Input:** $\mathbf{b}_{[0]}, \mathbf{z}_{[0]}, \mathbf{u}_{[0]}, \mathbf{v} \in \mathbb{R}^p$, $\mathbf{w} \in \mathbb{R}^m$, $\eta \in (0,1)$, $\alpha \in [0,1]$, $\gamma_{[0]}, t_{\max}, \epsilon, \lambda > 0$
**for** $t = 0, 1, 2, \ldots$ **until** $\|\mathbf{b}_{[t+1]} - \mathbf{z}_{[t]}\|_2 \leq \epsilon$ **or** $t > t_{\max}$ **do**
  $\mathbf{b}_{[t+1]} \leftarrow \operatorname{prox}_{\text{slope}}(\mathbf{z}_{[t]} - \gamma_{[t]}\mathbf{u}_{[t]} - \gamma_{[t]}\nabla f(\mathbf{z}_{[t]}); \gamma_{[t]}\lambda\alpha\mathbf{v})$ ▶ SLOPE
  **while** $f(\mathbf{b}_{[t+1]}) > Q(\mathbf{b}_{[t+1]}; \mathbf{z}_{[t]}, \gamma_{[t]})$ **do**
    $\mathbf{b}_{[t+1]} \leftarrow \operatorname{prox}_{\text{slope}}(\mathbf{z}_{[t]} - \gamma_{[t]}\mathbf{u}_{[t]} - \gamma_{[t]}\nabla f(\mathbf{z}_{[t]}); \gamma_{[t]}\lambda\alpha\mathbf{v})$
    $\gamma_{[t]} \leftarrow \eta\gamma_{[t]}$ ▶ Backtracking
  **end while**
  $\mathbf{z}_{[t+1]} \leftarrow \operatorname{prox}_{\text{gslope}}(\mathbf{D}\mathbf{b}_{[t+1]} + \mathbf{D}^{-1}\gamma_{[t]}\mathbf{u}_{[t]}; \gamma_{[t]}\lambda(1-\alpha)\mathbf{w})$ ▶ gSLOPE
  $\mathbf{u}_{[t+1]} \leftarrow \mathbf{u}_{[t]} + (\mathbf{b}_{[t+1]} - \mathbf{D}^{-1}\mathbf{z}_{[t+1]})/\gamma_{[t]}$
**end for**
**Output:** $(\mathbf{b}_{[t+1]}, \mathbf{u}_{[t+1]})$, where $\mathbf{b}_{[t+1]}$ solves Equation 3.1

### 3.3.2 Adaptive Step Size Search

Using a constant step size in the update may cause the algorithm to converge to a non-stationary point [85]. ATOS uses an adaptive line search to calculate the step size. It works in a similar way to a backtracking line search, which is guaranteed to converge to the global optimum for convex functions [85]. To perform the search, an initial step size $\gamma_{[0]}$ is set (see Appendix B.1.3), and is decreased using $\gamma_{[t]} = \eta\gamma_{[t]}$, such that $\eta \in (0,1)$ defines the level of backtracking, until $f(\mathbf{b}_{[t+1]}) \leq Q(\mathbf{b}_{[t+1]}; \mathbf{z}_{[t]}, \gamma_{[t]})$, where [147]

$$Q(\mathbf{b}; \mathbf{z}, \gamma) := f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{b} - \mathbf{z} \rangle + \frac{1}{2\gamma} \|\mathbf{b} - \mathbf{z}\|_2^2 . \tag{3.3}$$

### 3.3.3 Group SLOPE Proximal Weight Adjustment

In the derivation of the proximal operator for gSLOPE, a transformation is applied, so that it cannot be used directly in the ATOS algorithm [80]. The gSLOPE method is defined by the solution to the convex minimisation problem

$$\min_{\mathbf{b} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda \sum_{j=1}^{m} w_j \sqrt{p_j} \|\mathbf{b}^{(j)}\|_2 \right\}. \tag{3.4}$$

Gossmann et al. [80] reformulate this using the transformation $c_j = \sqrt{p_j} b_j$. Let $\mathbf{D}$ be the diagonal matrix with entries $\sqrt{p_i}$, so that $\mathbf{c} = \mathbf{D}\mathbf{b}$. Then, Equation 3.4 can be written as

$$\min_{\mathbf{c} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{D}^{-1}\mathbf{c}\|_2^2 + \lambda \sum_{j=1}^{m} w_j \|\mathbf{c}^{(j)}\|_2 \right\}.$$

Defining $g(\mathbf{c}) := \lambda \sum_{j=1}^{m} w_j \|\mathbf{c}^{(j)}\|_2$, the update step for $\mathbf{c}$ and $\mathbf{b}$ are given by

$$\mathbf{c}_{[t+1]} = \operatorname{prox}_g \left( \mathbf{c}_{[t]} - \gamma_{[t]} \nabla f(\mathbf{b}_{[t]}) \right), \quad \mathbf{b}_{[t+1]} = \mathbf{D}^{-1}\mathbf{c}_{[t+1]},$$

where $\nabla f(\mathbf{b}_{[t]}) = (\mathbf{X}\mathbf{D}^{-1})^\top (\mathbf{X}\mathbf{b}_{[t]} - \mathbf{y})/n$ is the gradient of the linear loss function. The proximal mapping returns the vector $\mathbf{c}$, instead of the desired vector $\mathbf{b}$, and takes $\mathbf{c}$ as input to the proximal mapping. As such, a transformation to the input is required, which is then reversed after applying the proximal mapping. The gSLOPE update steps are altered as follows:

$$\begin{aligned}
& \mathbf{z}_{[t+1]} = \operatorname{prox}_{\text{gslope}}(\mathbf{b}_{[t+1]} + \gamma_{[t]} \mathbf{u}_{[t]}; \gamma_{[t]} \mathbf{w}) \\
\rightarrow \quad & \mathbf{z}_{[t+1]} = \operatorname{prox}_{\text{gslope}}(\mathbf{D}\mathbf{b}_{[t+1]} + \mathbf{D}^{-1}\gamma_{[t]} \mathbf{u}_{[t]}; \gamma_{[t]} \mathbf{w}), \\
& \mathbf{u}_{[t+1]} = \mathbf{u}_{[t]} + (\mathbf{b}_{[t+1]} - \mathbf{z}_{[t+1]})/\gamma_{[t]} \\
\rightarrow \quad & \mathbf{u}_{[t+1]} = \mathbf{u}_{[t]} + (\mathbf{b}_{[t+1]} - \mathbf{D}^{-1}\mathbf{z}_{[t+1]})/\gamma_{[t]}.
\end{aligned}$$

The transformation $\mathbf{D}\mathbf{b}_{[t+1]}$ ensures $\mathbf{c}_{[t+1]}$ is the input into the operator and the transformation $\mathbf{D}^{-1}\mathbf{z}_{[t]}$ recovers $\mathbf{b}_{[t+1]}$. Additionally, the gSLOPE transformation alters the gradient, $\nabla f$, to include an additional $\mathbf{D}^{-1}$ term, so the transformation $\mathbf{D}^{-1}\gamma_{[t]}$ accounts for this difference.

## 3.4 FDR Control

Applying SGS to linear regression under orthogonal designs using the standard SLOPE and gSLOPE penalty sequences (Equations 2.3 and 2.4) does not provide FDR control (see Section 3.4.3). These sequences were not designed for the sparse-group setting, particularly when scaled by $\alpha$ and $1 - \alpha$. Since the goal is to control the FDR at both the variable and group levels, the penalty sequences must interact to account for the two layers of penalisation and respective scaling. Our result introduces sequences that achieve this goal and is the first to theoretically establish concurrent bi-level FDR control.

### 3.4.1 SLOPE FDR Proofs

FDR control at a single level was shown for both SLOPE [24] and gSLOPE [30] under orthogonal designs. To achieve bi-level FDR control, we derive new, inter-dependent penalty sequences. Theorem 3.1 introduces variable and group penalties that concurrently control both FDRs under the Normal Means model. Corollary 3.1 links the case of an orthogonal design matrix to the Normal Means model to demonstrate FDR control under orthogonality. The sequences depend on each other, facilitating concurrent FDR control. Lemma 3.1 guarantees that such sequences can always be found and provides a condition for their uniqueness. The proofs of the results are given in Appendix B.2.1. The sequences are derived in Appendix B.2.4 and are applied to orthogonal settings in Section 3.4.4.

The leave-one-out arguments used to prove FDR control for SLOPE and gSLOPE do not readily extend to SGS. This is due to the presence of two non-separable, sorted norms, as well as the simultaneous testing of hypotheses at both the variable and group levels. To extend the leave-one-out argument to SGS, it was necessary to derive new norm results (see Appendix B.2.5).

For SLOPE models, FDR control has been established under the linear model, where $\text{Var}(\boldsymbol{\beta})$ depends solely on the common noise level $\sigma$. In contrast, for logistic regression, even with orthogonal designs, each $\text{Var}(\boldsymbol{\beta})$ depends on the values of the other variances, making controlling the FDR significantly more complex.

### 3.4.2 Bi-level FDR Control

**Theorem 3.1** (Bi-level FDR control for SGS under the Normal Means model)**.** *Suppose we have a linear model* $\mathbf{y} = \boldsymbol{\beta} + \boldsymbol{\epsilon}$*, where* $n \geq p$*,* $\mathbf{y} \sim \boldsymbol{\mathcal{N}}_p(\boldsymbol{\beta}, \mathbf{I}_p)$*, and* $\boldsymbol{\epsilon} \sim \boldsymbol{\mathcal{N}}_p(\mathbf{0}, \mathbf{I}_p)$*. Further, suppose there exists some* $m$*-partition of the input space* $\mathcal{G} = \{G_1, \dots, G_m\}$*, such that* $G_i \cap G_j = \varnothing$ *for* $i \neq j$ *and* $\bigcup_{i=1}^m G_i = [p]$*. Suppose we apply SGS (Equation 3.1) to obtain* $\hat{\boldsymbol{\beta}}$*, using the variable weights*

$$v_i = \max_{j \in [m]} \left\{ \frac{1}{\alpha} \Phi^{-1} \left( 1 - \frac{q_v i}{2p} \right) - \frac{1}{3\alpha} (1 - \alpha) a_j w_j \right\}, \; i \in [p],$$

*and group weights*

$$w_i = \max_{j \in [m]} \left\{ \frac{F_{FN}^{-1} \left( 1 - \frac{q_g i}{m} \right) - \alpha \sum_{k \in G_j} v_k}{(1 - \alpha) p_j} \right\}, \; i \in [m],$$

*where* $a_j = |\{i \in G_j : \beta_i \neq 0\}|$, $q_v, q_g \in (0,1)$ *are the desired variable and group FDR levels, and* $\Phi$ *and* $F_{FN}$ *are the cumulative distribution functions of a standard Gaussian distribution and a folded Gaussian distribution, respectively.*

*Then, concurrently testing the multiple variable and group hypotheses*

$$H_i^v : \beta_i = 0,\ i \in [p], \quad H_i^g : \|\boldsymbol{\beta}^{(i)}\|_2 = 0,\ i \in [m],$$

*and defining* $V^v$ *and* $R^v$ *as the number of false and total variable rejections, and* $V^g$ *and* $R^g$ *as the number of false and total group rejections, given by*

$$\begin{aligned} V^v &= |\{i \in [p] : \hat{\beta}_i \neq 0, \beta_i = 0\}|, & R^v &= |\{i \in [p] : \hat{\beta}_i \neq 0\}|, \\ V^g &= |\{i \in [m] : \|\hat{\boldsymbol{\beta}}^{(i)}\|_2 \neq 0, \|\boldsymbol{\beta}^{(i)}\|_2 = 0\}|, & R^g &= |\{i \in [m] : \|\hat{\boldsymbol{\beta}}^{(i)}\|_2 \neq 0\}|, \end{aligned}$$

*SGS provides bi-level FDR control by bounding the variable and group FDR by*

$$\begin{aligned} \mathit{vFDR} &:= \mathbb{E}\left[\frac{V^v}{\max(R^v, 1)}\right] \leq q_v \frac{p_0}{p},\ \textit{where } p_0 = |\{i \in [p] : \beta_i = 0\}|, \\ \mathit{gFDR} &:= \mathbb{E}\left[\frac{V^g}{\max(R^g, 1)}\right] \leq q_g \frac{m_0}{m},\ \textit{where } m_0 = |\{i \in [m] : \|\boldsymbol{\beta}^{(i)}\|_2 = 0\}|. \end{aligned}$$

**Corollary 3.1** (Bi-level FDR control for SGS under orthogonality)**.** *Theorem 3.1 applied to the linear model* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ *when* $\mathbf{X}$ *is orthogonal leads to bi-level FDR control.*

**Lemma 3.1** (Existence and uniqueness of penalty sequences)**.** *For the SGS weight sequences defined in Theorem 3.1, the following hold*

*(i) Existence: there exists a pair* $(\mathbf{v}^*, \mathbf{w}^*) \in \mathbb{R}^p_{\geq 0} \times \mathbb{R}^m_{\geq 0}$ *satisfying both weight equations simultaneously.*

*(ii) Uniqueness: if* $\max_{j \in [m]} a_j < 3$*, then the penalty pair* $(\mathbf{v}^*, \mathbf{w}^*) \in \mathbb{R}^p_{\geq 0} \times \mathbb{R}^m_{\geq 0}$ *satisfying both weight equations simultaneously is unique, and can be computed iteratively from a starting point* $\mathbf{v}^{(0)} \in \mathbb{R}^p_{\geq 0}$*, with geometric convergence*

$$\|\mathbf{v}^{(n)} - \mathbf{v}^*\|_\infty \leq \left(\frac{\max_{j \in [m]} a_j}{3}\right)^n \|\mathbf{v}^{(0)} - \mathbf{v}^*\|_\infty.$$

*Remark* 3.2. The uniqueness statement in Lemma 3.1 requires $\max_{j \in [m]} a_j < 3$ for sequence uniqueness, limiting groups to at most two active variables. In practice, this bound is often exceeded, yet FDR control still holds. The restriction arises from

a norm bound in Lemma B.7 used in Theorem 3.1; any tighter bound would reduce the dependency of $\mathbf{v}$ on $\mathbf{w}$, suggesting that the current bound is likely already the sharpest necessary to ensure FDR control. Importantly, Theorem 3.1 only requires existence, not uniqueness, for bi-level FDR control.

*Remark* 3.3. In the proof of Theorem 3.1, we assume without loss of generality that $\lambda = 1$, corresponding to the noise level under the orthogonal model used for FDR control. In practice, the SGS implementation employs a loss function scaled by a factor of $1/2n$, whereas the theoretical analysis uses a factor of $1/2$ for simplicity. This practical scaling allows for comparability across different sample sizes, which is useful when applying CV. The two formulations are easily reconciled by scaling $\lambda$ by $1/n$.

### 3.4.3 Penalty Sequences

Theorem 3.1 shows that the penalty sequences which guarantee bi-level FDR control for SGS are given by sequences, which we term the vMax and gMax sequences,

$$v_i^{\max} = \max_{j \in [m]} \left\{ \frac{1}{\alpha} \Phi^{-1}\left(1 - \frac{q_v i}{2p}\right) - \frac{1}{3\alpha}(1-\alpha) a_j w_j \right\}, \; i \in [p], \qquad (vMax) \qquad (3.5)$$

$$w_i^{\max} = \max_{j \in [m]} \left\{ \frac{F_{\mathrm{FN}}^{-1}\left(1 - \frac{q_g i}{m}\right) - \alpha \sum_{k \in G_j} v_k}{(1-\alpha) p_j} \right\}, \; i \in [m], \qquad (gMax) \qquad (3.6)$$

where $a_j$ represents the number of active variables within an active group (discussed in Section 3.4.4). In contrast to the SLOPE sequences, the SGS sequences are defined by taking the maximum across groups, reflecting the fact that the distribution of each group is different (see Remark 2.3 for further details). The sequences are shown in Figure 3.4.

The penalty sequences that guarantee FDR control (under orthogonality) for SLOPE and gSLOPE are given in Equations 2.3 and 2.4. As the SLOPE BH sequence only controls the FDR under orthogonal settings, a modified sequence for non-orthogonal designs, termed the *Gaussian sequence*, was derived for SLOPE by Bogdan et al. [24], but it reduces to the lasso under high-dimensional settings [111]. These settings are the primary focus of this thesis, so the sequence is not considered further in this chapter, although it finds an interesting application in Chapter 7.

One could use Equations 2.3 and 2.4 for SGS (which we term *SGS Original*), but this would be a rather naive approach, as these penalty sequences were not designed

for bi-level selection. Indeed, applying SGS Original to orthogonal data (with the setup of Section 3.4.4) does not achieve bi-level FDR control (Figures B.1 and B.2). Additionally, our synthetic data results show that SGS Original does worse than using the derived sequences from Theorem 3.1 in practice (Table 3.1).

An alternative approach is to set $\alpha = 0.5$ and $\lambda = 2$ in SGS Original, which applies both penalties at the same magnitude as in the single-norm variants (termed *SGS Double*). However, this was also found to be unsatisfactory, as too much penalisation is applied, so the FDR-sensitivity trade-off is not optimised (Figures B.1 and B.2).

### Relaxed Sequences

Brzyski et al. [30] observed that under orthogonal data, the gSLOPE max sequence (Equation 2.5) penalises excessively, leading to an unnecessary loss of power. On the other hand, SLOPE achieves exact FDR control. Remark 2.3 demonstrates that this is because the gSLOPE sequence is taken over the maximum of all groups, which causes the FDR control to be overly conservative.

This suggests that the magnitude of the penalties can be decreased, achieving higher power while still maintaining FDR control. Indeed, Brzyski et al. [30] apply a relaxation to the gSLOPE sequence to form the gSLOPE mean sequence, given by, for $i \in [m]$,

$$w_i^{\text{mean}} = \overline{F}^{-1}_{\chi_{p_j}}\left(1 - \frac{q_g i}{m}\right), \ \overline{F}_{\chi_{p_j}}(x) = \frac{1}{m}\sum_{j=1}^{m} F_{\chi_{p_j}}(\sqrt{p_j}x). \quad (gSLOPE\ mean) \quad (3.7)$$

The relaxed sequences were found to still control the FDR at the specified level and recover much of the lost power [30]. Figure 3.4 compares the relaxed gSLOPE sequence to its more stringent counterpart.

A relaxation of our derived penalty sequences is also possible. Observe that in the proof for Theorem 3.1 (see Appendix B.2.1), Equation B.5 can be recast as, for $i \in [p]$,

$$\frac{1}{m}\sum_{j=1}^{m} \Phi\left(\alpha v_i + \frac{1}{3}(1-\alpha)a_j w_j\right) \geq 1 - \frac{q_v i}{2p}.$$

So, the variable sequence can be relaxed to be

$$v_i^{\text{mean}} = \overline{\Phi}^{-1}\left(1 - \frac{q_v i}{2p}\right), \quad (vMean) \tag{3.8}$$
$$\text{where} \quad \overline{\Phi}(x) = \frac{1}{m}\sum_{j=1}^{m} \Phi\left(\alpha x + \frac{1}{3}(1-\alpha)a_j w_j\right).$$

Applying a similar relaxation to the gMax sequence (Equation 3.6) gives, for $i \in [m]$,

$$w_i^{\text{mean}} = \overline{F}_{\text{FN}}^{-1}\left(1 - \frac{q_g i}{m}\right), \quad (gMean) \tag{3.9}$$
$$\text{where} \quad \overline{F}_{\text{FN}}(x) = \frac{1}{m}\sum_{j=1}^{m} F_{\text{FN}}\left((1-\alpha)p_j x + \alpha \sum_{k \in G_j} v_k\right).$$

The relaxed penalty sequences are visualised in Figure 3.4 and are experimentally compared with the maximum sequences. The relaxation at the variable level has minimal impact, whereas the group-level relaxation produces a noticeable effect, as expected, since the relaxation is applied across groups.

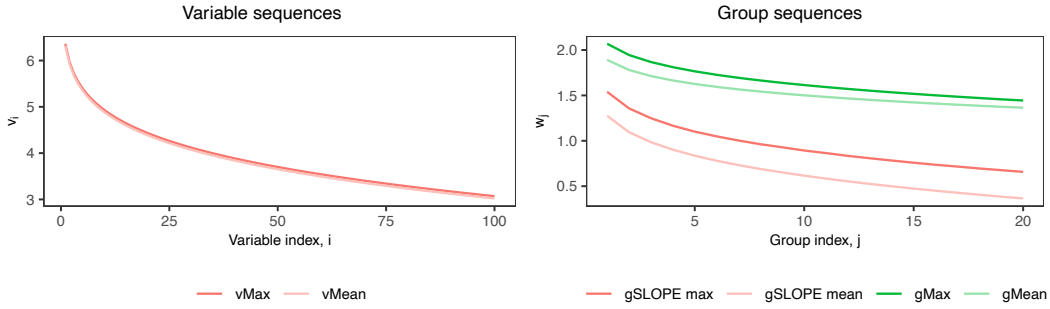



Figure 3.4: Variable and group sequences under $m = 20$ uneven groups of sizes $[3, 7]$ with $p = 100$ and $\alpha = 0.5$.

### Practical Considerations

A few practical considerations are required when implementing the SGS sequences.

For the variable sequences, the quantity $a_j$, which represents the number of active variables within an active group such that $0 \leq a_j \leq p_j$, is generally unknown and requires estimation. We estimate it using $\hat{a}_j = \lfloor \alpha p_j \rfloor$. Figures 3.5 and 3.6 show that this estimator ($\hat{a}_j = 3$ in the case considered) achieves the highest sensitivity,

while maintaining bi-level FDR control. FDR control was achieved for all values of $a_j$ under the max sequences (Figures B.3 and B.4); showing that the theoretical properties hold for any estimator.

To generate the group sequences, the quantity $\sum_{k \in G_j} v_k$ is required. While the sequence of variables is known, the exact assignment of penalties to individual variables, and consequently to the groups they belong to, is not fully specified a priori. Instead, partial information of the group sizes is used to assign the highest ranking groups (those with the largest $\sqrt{p_j}\|\boldsymbol{\beta}^{(j)}\|_2$ values) as those with the largest group sizes, so that they are assigned the largest variable penalties, $\{v_1, \dots, v_{p_j}\}$.

Finally, as the variable and group sequences depend on each other, they are initialised. This is done via the SLOPE BH sequence: plug the BH sequence (Equation 2.3) into the SGS group sequence (Equation 3.6), then plug this resulting group sequence into the SGS variable sequence (Equation 3.5) to generate $\mathbf{v}^{\max}$ and $\mathbf{w}^{\max}$.

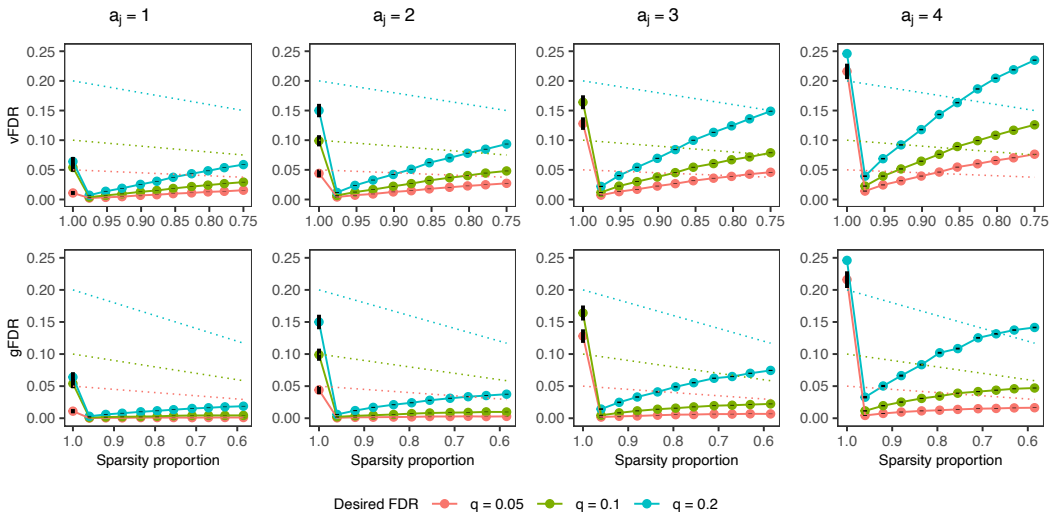



Figure 3.5: vFDR and gFDR of SGS with vMean and gSLOPE mean sequences under orthogonal matrix $\mathbf{X} = \mathbf{I}_{1000}$ with 200 equal groups of size 5. This is shown for different values of $a_j$, as a function of decreasing sparsity proportion. The signal effect was set at $\beta_i = 5\delta\sqrt{2\log p}, i \in S_v$, $\delta \sim \mathcal{N}(0,1)$, with 1000 simulation repetitions performed per sparsity proportion. The dotted lines represent the FDR theoretical upper bounds.

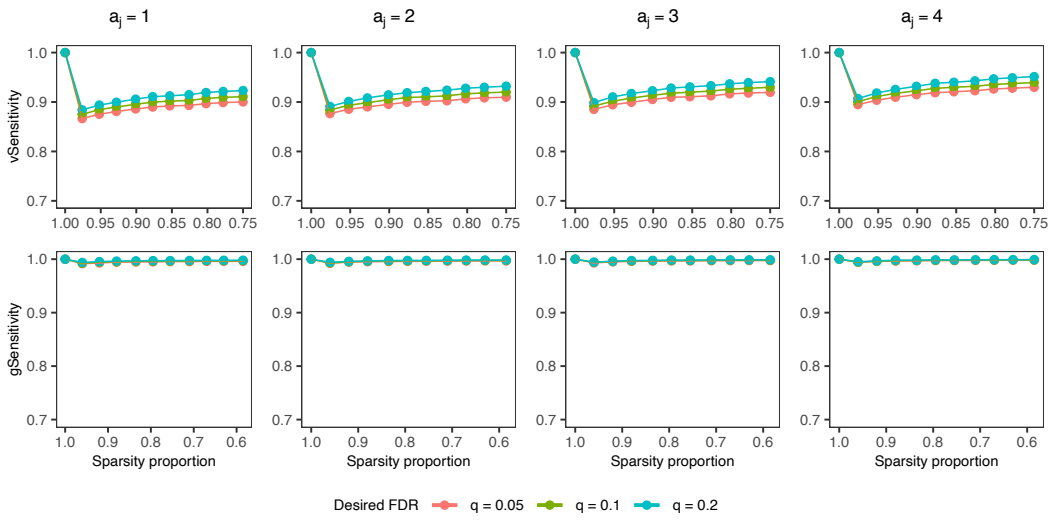



Figure 3.6: vSensitivity and gSensitivity of SGS with vMean and gSLOPE mean sequences under orthogonal matrix $\mathbf{X} = \mathbf{I}_{1000}$ with 200 equal groups of size 5. This is shown for different values of $a_j$, as a function of decreasing sparsity proportion. The signal effect was set at $\beta_i = 5\delta\sqrt{2\log p}, i \in S_v$, $\delta \sim \mathcal{N}(0,1)$, with 1000 simulation repetitions performed per sparsity proportion.

### 3.4.4 Computational Experiments

In this section, we confirm computationally the bi-level FDR control of the sequences that guarantee FDR control in Theorem 3.1 and explore whether the relaxed sequences can maintain FDR control while recovering lost power.

To satisfy the conditions required for FDR control, an orthogonal design matrix is used $\mathbf{X} = \mathbf{I}_{1000}$. For the grouping structure, 200 groups of 5 variables each were generated. The group sparsity proportion of the true signal varied from 1 to 0.59, where *sparsity proportion* refers to the proportion of inactive groups in the true model (a sparsity proportion of 1 corresponds to the null model, whereas a value of 0 corresponds to the fully saturated model). Within an active group, the variable sparsity proportion was set to 0.4, resulting in the global variable sparsity proportion varying from 1 to 0.75.

The true effects were set to $\beta_i = 5\delta\sqrt{2\log p}$, for $i \in S_v$, and $\delta \sim \mathcal{N}(0,1)$, as the expected value of the maximum of $p$ independent standard Gaussian variables is approximately $\sqrt{2\log p}$ [37]. The response was generated using the linear model $\mathbf{y} = \boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \boldsymbol{\mathcal{N}}_p(\mathbf{0}, \mathbf{I}_p)$, with 1000 simulation repetitions performed per sparsity

proportion. SGS was applied with varying $q_v, q_g \in \{0.05, 0.1, 0.2\}$, and $\alpha = 0.6$ and $\lambda = 1/n$ fixed; see Remark 3.3 for the justification behind the choice of $\lambda$.

SGS achieves bi-level FDR control using the vMax and gMax sequences (Figure 3.7, with sensitivity levels around 0.9; see Figure B.5). SGS also achieves this under $\alpha = 0.95$, which is used in the synthetic data study in Section 3.5 (Figure B.6). Using the SGS relaxed sequences, the bi-level FDR is kept close to the desired level, but group FDR control is not achieved (Figure B.7). Combining the vMean sequence with the gSLOPE mean sequence achieves bi-level FDR control while maintaining the highest sensitivity among the SGS sequences considered (Figures 3.7 and B.5).

Groups of uneven sizes were also considered: for the grouping structure, 40 groups of sizes between 3 and 7 were generated, with the simulation parameters remaining otherwise identical. Under both the max sequences and the combination of vMean and gSLOPE mean, bi-level FDR control is achieved (Figure 3.8 with sensitivity in Figure B.8).

In this section, we have observed overly conservative FDR control for SGS, which is also observed for gSLOPE in Brzyski et al. [30]. On the other hand, SLOPE was shown to achieve exact FDR control in experimental studies [24]. For gSLOPE, this arises solely from the requirement that the sequence be defined by the maximum over all groups (Remark 2.3). In the case of SGS, this is also present, but additional conservativeness is introduced through the extra assumptions needed to derive the sequences under the sparse-group norm.

*Remark* 3.4 (Weight choices for SLOPE models). For the remainder of this thesis (unless stated otherwise), SGS will employ the vMean and gSLOPE mean sequences (Equations 3.8 and 3.7). Consequently, the assumption required to compute the quantity $\sum_{k \in G_j} v_k$ in the gMean and gMax sequences is no longer necessary. Furthermore, gSLOPE will be implemented using the gSLOPE mean sequence, and SLOPE will be implemented using the BH sequence.

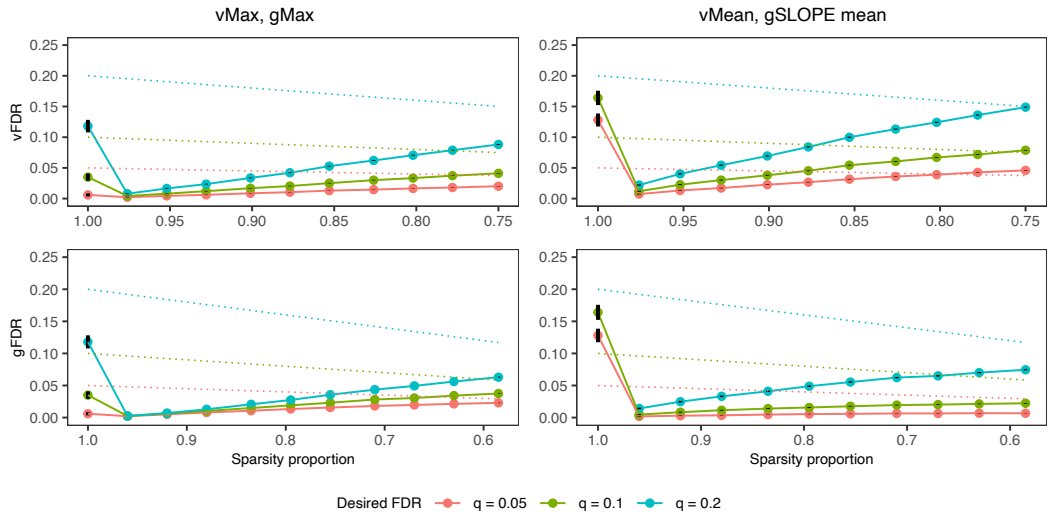



Figure 3.7: vFDR and gFDR of SGS with the vMax, gMax (left), and the vMean, gSLOPE mean (right) sequences under orthogonal design with 200 equal groups of size 5. Shown as a function of decreasing sparsity proportion, with 1000 simulation repetitions performed per sparsity proportion. The dotted lines represent the FDR theoretical upper bounds.

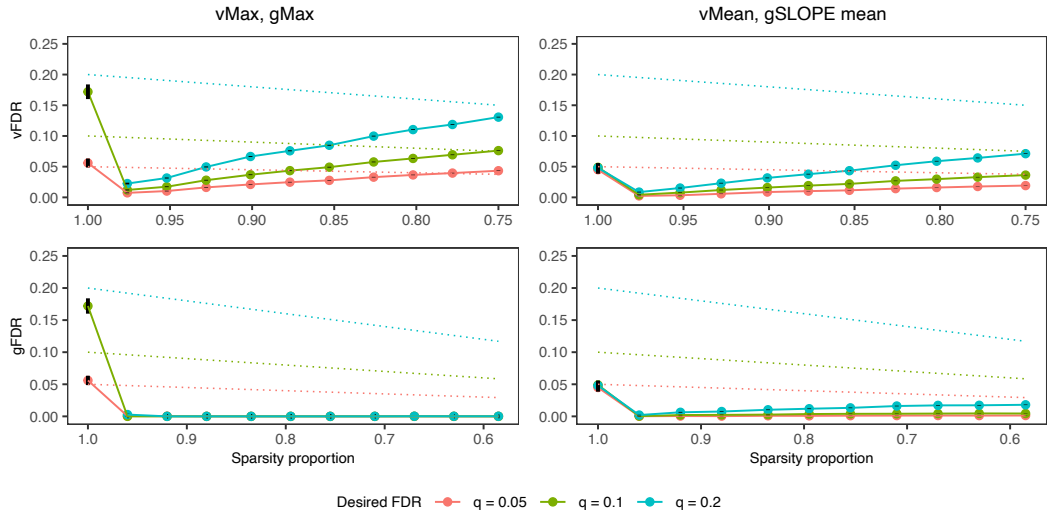



Figure 3.8: vFDR and gFDR of SGS with the vMax, gMax (left) and the vMean, gSLOPE (right) mean sequences under orthogonal design with 40 uneven groups of sizes $\{3, \ldots, 7\}$. Shown as a function of decreasing sparsity proportion, with 1000 simulation repetitions performed per sparsity proportion. The dotted lines represent the FDR theoretical upper bounds.

## 3.5 Synthetic Data Experiments

An extensive simulation study was conducted to investigate the variable and group selection performance of SGS on non-orthogonal data. SGS is compared to the lasso, gLasso, SGL, SLOPE, and gSLOPE under various scenarios, including varying the signal sparsity, the group sizes, the $p/n$ ratio, and comparing fixed versus random signals.

### 3.5.1 Simulation Design

The design matrix $\mathbf{X} \sim \boldsymbol{\mathcal{N}}_p(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^{200\times 800}$ was used with correlation matrix $\boldsymbol{\Sigma}$. Three cases of within-group correlation were considered: no, medium, and high correlation, corresponding to $\rho = 0, 0.3, 0.9$, for $\Sigma_{i,j} = \rho$, where $i \neq j$ and $i$ and $j$ belong to the same group. The response was generated using the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2), i \in [p]$, and $\sigma$ chosen adaptively so that the signal-to-noise ratio was set at 6.

The following parameters were set up for the base scenario of data generation. The signal of the active variables was fixed at $\beta_i = 5, i \in S_v$. The variables were split into 160 non-overlapping groups of sizes $\{3, \dots, 7\}$. Six different cases of sparsity proportions were considered, varying the group sparsity proportion from 1 to 0.84. The variable sparsity proportion in an active group is set at 0.4, resulting in the global variable sparsity proportion varying from 1 to 0.82.

The simulations were performed over six sparsity and three correlation cases, giving 18 unique simulation setups, each repeated 100 times.

### 3.5.2 Model Hyperparameters

SGS and SGL were applied using $\alpha = 0.95$ and $q_v = q_g = 0.1$ for SGS, corresponding to a target FDR of 0.1. The SLOPE models used their default penalty sequences (see Remark 3.4). For each model, the data was $\ell_2$-standardised and an intercept fit. Models were fit using 10-fold CV along a 20-$\lambda$ log-linear path, starting at a value of $\lambda_{\max}$ which generates the null model (Proposition 3.1 for SGS) and terminating at $0.15\lambda_{\max}$, with the optimum model chosen such that the test error is within one standard error of the minimum (known as the 1se model).

### 3.5.3 Metrics

The $F_1$ score is used as a primary selection comparison metric (Definition A.2), providing a balance between sensitivity and FDR, with a high score being preferable. The mean squared error (MSE) and mean absolute error (MAE) of the fitted values to the true solution are also reported.

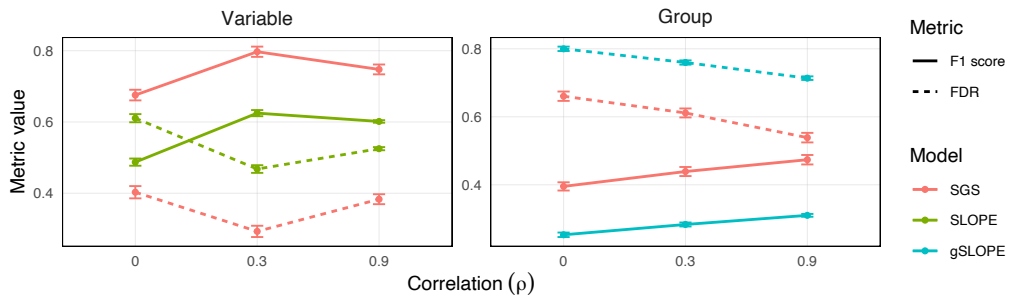
Further simulation design and model implementation information can be found in Section A.3.

### 3.5.4 Results

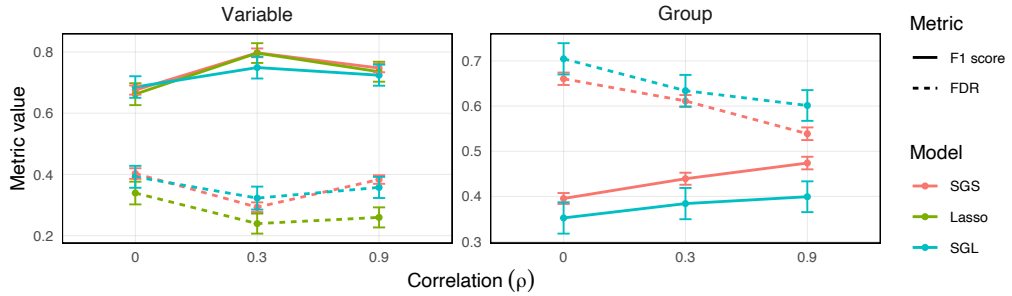
Averaging the metrics across the six sparsity cases, SGS achieves substantially higher $F_1$ scores and lower FDR levels than both SLOPE and gSLOPE, across all correlation values (Figure 3.9a). The grouping information is found to be useful in selecting the relevant variables and groups. Improved performance is observed as the correlation increases, especially for group selection, as the correlation acts within a group, making the grouping information more important.

SGS has an $F_1$ score almost identical to the lasso, although surprisingly higher FDR (Figure 3.9b). This is caused by using CV for model selection, as CV has no aim of FDR control, so the FDR properties of SLOPE models are lost (an issue tackled for SLOPE models in Chapter 7). However, SLOPE models do still perform adaptive penalisation, the benefit of which is demonstrated by comparing the two bi-level selection approaches, where SGS outperforms SGL for both selection types. The difference in performance increases as the correlation increases for group selection, providing evidence that SLOPE-based models perform stronger under correlated designs, which is in agreement with findings presented in Zeng and Figueiredo [230]. Averaging across all correlation cases, SGS obtains the highest variable $F_1$ score ($0.74 \pm 10^{-3}$) and group $F_1$ score ($0.44 \pm 10^{-3}$) of all the models considered (Table B.1). To achieve these metrics, SGS used sparser models than SLOPE and gSLOPE, with sparsity levels more in line with those of the lasso and SGL (Figure B.9).

SGS clearly performs best when using $\alpha = 0.95$, compared to $\alpha = 0.6, 0.05$ (Figure 3.10), showing that emphasis on variable penalisation, while enforcing some grouping, yields the strongest performance. Introducing across-group correlation ($\Sigma_{ij} = 0.05$ for $i \neq j$ in different groups), while maintaining within-group correlation, led to the lasso-based models outperforming the SLOPE-based ones (Figure B.10). This contrasts with Figure 3.9b, where SGL performed worse than SGS.




(a) SLOPE-based models




(b) Lasso-based models

Figure 3.9: $F_1$ score and FDR of the SLOPE-based (a) and lasso-based (b) models under $\beta_i = 5, i \in S_v$, for $\rho = 0, 0.3, 0.9$, with standard errors shown. The results were averaged over the six sparsity cases, with the average variable and group sparsity proportions at 0.91 and 0.92. 160 groups of sizes $\{3, \dots, 7\}$ were used and 600 simulation repetitions performed per correlation case.

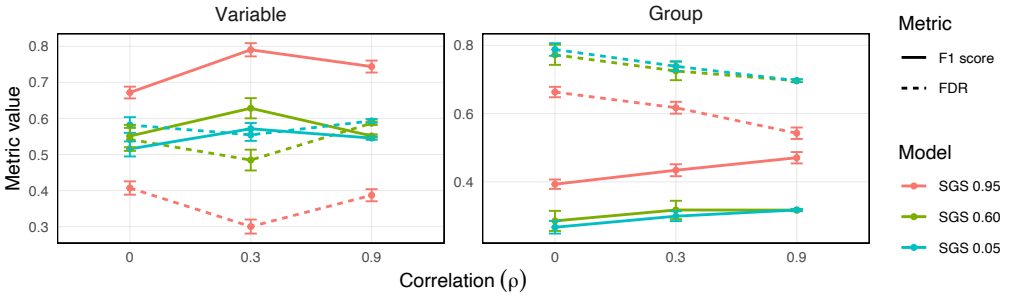



Figure 3.10: $F_1$ score and FDR of SGS for different values of $\alpha$ under $\beta_i = 5, i \in S_v$, for $\rho = 0, 0.3, 0.9$, with standard errors shown. The number next to SGS in the model legend indicates the value of $\alpha$ used. The results were averaged over the six sparsity cases, with the average variable and group sparsity proportions at 0.91 and 0.92. 160 groups of sizes $\{3, \dots, 7\}$ were used and 600 simulation repetitions performed per correlation case.

### Decreasing Sparsity

By not averaging across the six sparsity cases, we observe a drop to the $F_1$ score of SGS, with the FDR increasing, as a function of decreasing sparsity (Figure 3.11 with sensitivity in Figure B.11, and the lasso-based models shown in Figure B.12, with SGS found to have similar performance to SGL). This drop is slowed as the correlation increases and is present for all of the models. Under very sparse models, the FDR is found to be below the 0.1 desired level for SGS, as well as obtaining very high $F_1$ scores, but both metrics weaken as more predictors enter the model. The results are unsurprising, as decreasing the sparsity in the underlying model means there are more true signals for the models to detect, which makes obtaining a higher $F_1$ score and a lower FDR challenging.

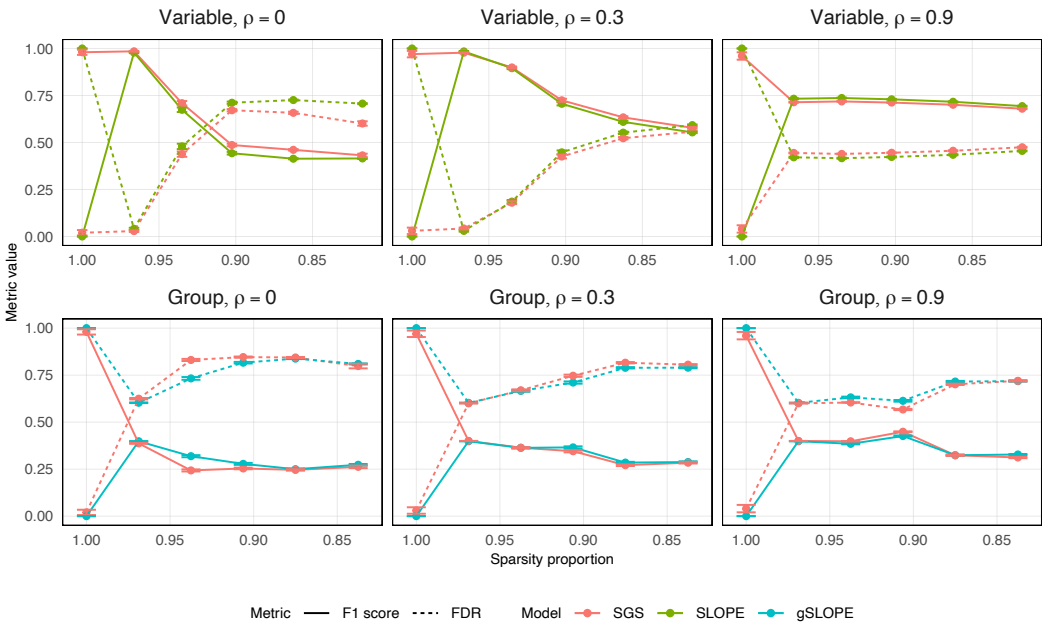



Figure 3.11: $F_1$ score and FDR as functions of decreasing sparsity proportion, for the SLOPE-based models, under $\beta_i = 5, i \in S_v$, with 160 groups of sizes $\{3, \dots, 7\}$. Standard errors are shown and 100 simulation repetitions were performed per sparsity proportion and correlation case.

### Random Signal

A clear drop-off in selection performance, in terms of the $F_1$ score, is observed for all models when the signal follows a Gaussian distribution and lower sparsity pro-

portions are considered (Table 3.1). Amongst the tested methods, SGS is found to have the highest $F_1$ score for both types of selection. SGS, with the derived sequences of Section 3.4.3, obtains far superior performance than the naive application of SGS Original, highlighting the importance of using theory to extract the full performance of SGS. Here, the SLOPE-based models have lower MSEs than the lasso-based ones, in particular, compared to SGL, which suffers from inflated estimates. As the SLOPE-based models apply stronger penalisation, which introduces additional bias, one would generally expect this trend to go the other way.

| **Model** | ***β* error** | | **Variable mean** | | | **Group mean** | | |
|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | $F_1$ | FDR | Sens. | $F_1$ | FDR | Sens. |
| SGS | 1.93 $_{(0.01)}$ | 0.43 $_{(10^{-4})}$ | **0.55** $_{(10^{-3})}$ | 0.46 $_{(10^{-3})}$ | 0.58 $_{(10^{-3})}$ | **0.46** $_{(10^{-3})}$ | **0.57** $_{(10^{-3})}$ | **0.56** $_{(10^{-3})}$ |
| SLOPE | **1.77** $_{(0.01)}$ | 0.43 $_{(10^{-3})}$ | 0.39 $_{(10^{-3})}$ | 0.64 $_{(10^{-3})}$ | **0.61** $_{(10^{-3})}$ | – | – | – |
| gSLOPE | 2.03 $_{(0.01)}$ | 0.48 $_{(10^{-3})}$ | – | – | – | 0.30 $_{(10^{-3})}$ | 0.74 $_{(10^{-4})}$ | 0.55 $_{(10^{-3})}$ |
| Lasso | 2.03 $_{(0.01)}$ | **0.41** $_{(10^{-3})}$ | 0.45 $_{(10^{-3})}$ | **0.35** $_{(10^{-3})}$ | 0.43 $_{(10^{-3})}$ | – | – | – |
| SGL | 144.62 $_{(0.99)}$ | 2.60 $_{(0.01)}$ | 0.43 $_{(0.01)}$ | 0.51 $_{(0.01)}$ | 0.47 $_{(10^{-3})}$ | 0.38 $_{(10^{-3})}$ | 0.63 $_{(10^{-3})}$ | 0.49 $_{(10^{-3})}$ |
| SGS Original | 3.28 $_{(0.01)}$ | 0.54 $_{(10^{-4})}$ | 0.25 $_{(10^{-3})}$ | 0.70 $_{(10^{-3})}$ | 0.23 $_{(10^{-4})}$ | 0.36 $_{(10^{-3})}$ | 0.58 $_{(10^{-3})}$ | 0.35 $_{(10^{-3})}$ |

Table 3.1: MSE, MAE, $F_1$ score, FDR, and sensitivity under $\beta_i \sim \mathcal{N}(0, 5^2), i \in S_v$, averaged over all correlation and sparsity cases, for SLOPE- and lasso-based models, with standard errors shown in grey. The average variable and group sparsity proportions were 0.78 and 0.80, with 160 groups of sizes $\{3, \dots, 7\}$, and 1800 simulation repetitions were performed in total. **Bold** represents the best value obtained for each metric.

### Large Groups

Considering larger groups of sizes $\{5, \dots, 75\}$, a drop in the variable $F_1$ score for SGS is observed (Table 3.2) in comparison to Figure 3.9a. As there were fewer groups present, a false group discovery is less likely, leading to a lower group FDR. Here, SGS obtains a higher $F_1$ score than the other two models, as well as lower FDR levels. The lasso-based models are shown in Table B.2, where on this occasion the non-adaptive lasso-based models perform stronger. This highlights a challenge of applying adaptive group penalisation under large groups.

| Model | $\beta$ error | | Variable mean | | | Group mean | | |
|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | $F_1$ | FDR | Sens. | $F_1$ | FDR | Sens. |
| SGS | 0.27 $_{(0.01)}$ | 0.11 $_{(10^{-3})}$ | **0.59** $_{(0.01)}$ | **0.51** $_{(0.01)}$ | **0.99** $_{(10^{-3})}$ | **0.71** $_{(0.01)}$ | **0.30** $_{(0.01)}$ | 0.98 $_{(10^{-3})}$ |
| SLOPE | 0.26 $_{(0.01)}$ | 0.10 $_{(10^{-3})}$ | 0.56 $_{(10^{-3})}$ | 0.54 $_{(10^{-3})}$ | **0.99** $_{(10^{-4})}$ | - | - | - |
| gSLOPE | **0.19** $_{(10^{-3})}$ | **0.10** $_{(10^{-3})}$ | - | - | - | 0.69 $_{(10^{-3})}$ | 0.35 $_{(10^{-3})}$ | **1.00** $_{(10^{-4})}$ |

Table 3.2: MSE, MAE, $F_1$ score, FDR, and sensitivity under $\beta_i = 5, i \in S_v$, averaged over all correlation and sparsity cases, for SLOPE-based models, with standard errors shown in grey. The average variable and group sparsity proportions were 0.97 and 0.92, with 25 groups of sizes $\{5, \dots, 75\}$. In total, 1800 simulation repetitions were performed. **Bold** represents the best value obtained for each metric.

### Impact of the $p/n$ Ratio

By reducing the $p/n$ ratio through increasing $n$ from 200 to 800, we gain insight into how SGS performs depending on the number of available observations (Figure 3.12 with sensitivity in Figure B.13, and the lasso-based models shown in Figure B.14). Under no and medium correlations, the $F_1$ score increases as the ratio decreases, while the FDR quickly decreases, with variable FDR control achieved as $p/n$ tends towards 1 (the interpolation threshold). Under high correlation, the $F_1$ score stagnates as the ratio decreases. The stagnation of the score under high correlation is similar to the trend seen in Figure 3.11, where the $F_1$ score stays the same under decreasing sparsity proportion. In terms of model performance, SGS tends to have stronger performance at a higher $p/n$ ratio, in comparison to SLOPE and gSLOPE, but the gap decreases with the ratio, showing that SGS provides a clear advantage when there are fewer observations available.

### False Positives

Finally, the number of false positives selected by each model was computed by generating data under the null model of no associations between the design matrix and the response (Table 3.3, using the setup from Section 3.5.1). SGS and the lasso have the fewest number of false positives. This case illustrates the downside of applying only group sparsity in gSLOPE, as the method has the highest number, selecting all variables in a group as active, leading to many inactive variables and additional noise in the model.

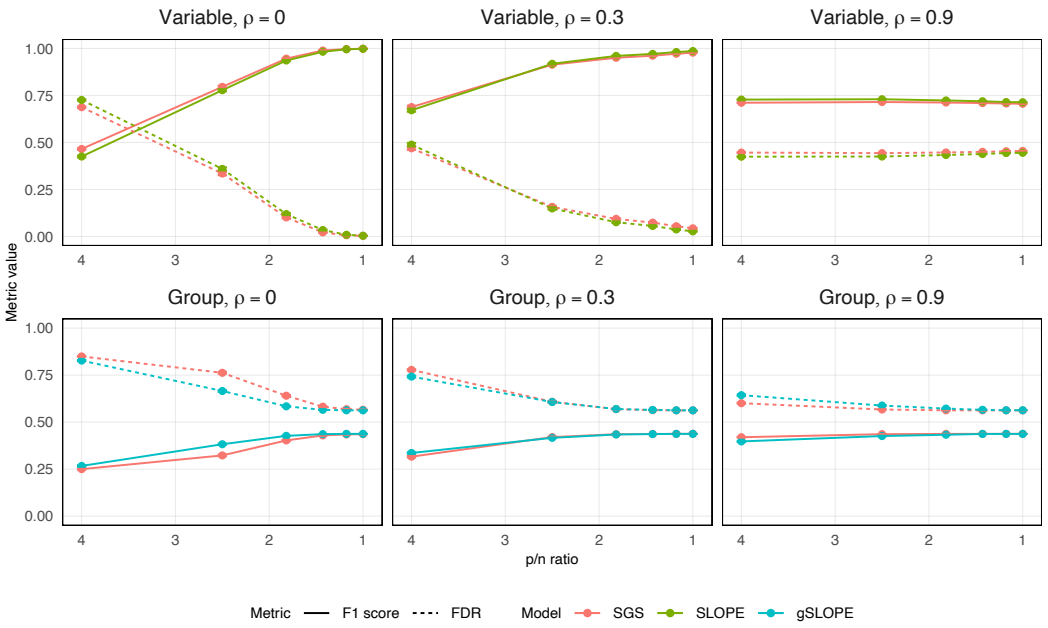



Figure 3.12: Average $F_1$ score and FDR across 100 simulation iterations as functions of decreasing $p/n$ ratio, for the SLOPE-based models under $\beta_i = 5, i \in S_v$. A single sparsity case was considered: variable/group proportions were set at 0.82 and 0.84, with 160 groups of sizes $\{3, \ldots, 7\}$.

| **Metric** | SGS | SLOPE | gSLOPE | Lasso | SGL |
|---|---|---|---|---|---|
| False positives | 0.51 (0.20) | 23.2 (1.54) | 208 (3.60) | **0.40** (0.06) | 8.33 (1.01) |

Table 3.3: The mean number of false variable positives selected for SLOPE- and lasso-based models for $p = 800$, with standard errors shown in grey for the 300 simulation repetitions performed. **Bold** represents the best value.

## 3.6 Real Data Experiments

We can use the associations discovered between genes and diseases to predict disease risk and outcome. In this section, the use of SGS as a disease prediction tool is explored through its application to two real datasets. The classification performance of SGS is compared to both lasso- and SLOPE-based models.

### 3.6.1 Datasets

The first dataset, Colitis, includes 127 individuals of which 85 have an inflammatory bowel disease (ulcerative colitis or Crohn's) and the other 42 are controls [34]. The expression of 22283 genes were microarrayed across the individuals. The second dataset, Cancer, contains data from 60 patients who had suffered from early-stage estrogen receptor-positive breast cancer and had been treated with tamoxifen [128]. The patients were classified based on whether the cancer had recurred. The initial dataset contained over 22575 genes but had high levels of missingness. As a result, genes with over 50% missingness were removed, leaving 12071 remaining genes, and mean imputation on those genes was applied to remove any remaining missingness.

Both datasets were accessed using the `GEOquery` R function.[1] The two datasets have previously been analysed using lasso-based models in Simon et al. [178]. As the dataset sources have been updated since this publication, the analysis is repeated here for those models, using the same cleaning steps as in Simon et al. [178].

The nine major collections of gene sets, C1-C8 and H, of the Human Molecular Signatures Database (MSigDB)[2] were downloaded for grouping the genes of the two datasets into pathways. Table B.5 presents the number of pathways and their allocated genes for each dataset and collection. For each pathway, the overlapping genes were duplicated into the different pathways to which they belong [191].

For both datasets, the samples were split into training and test sets, and the classification rate on the test set was computed using the trained models. All nine collections were analysed for both datasets (Table B.5). The C3 and C8 collections achieved the highest peak classification results for Colitis and Cancer, respectively, and the results of these are presented in this section.

Further details on the datasets are provided in Appendix A.3.3.

### 3.6.2 Model Hyperparameters

SGS and SGL were applied with $\alpha = 0.99$, as this was found to yield better predictive performance on the real data. In the dataset considered here, many groups are large, and choosing an $\alpha$ value close to 1 allows the sparse-group models to more effectively discard noise variables in active groups. Each model was applied to a log-linear path of 100 $\lambda$ values, starting at a value of $\lambda_{\max}$ which generates the null

[1] Accessed on 08/03/2023.
[2] gsea-msigdb.org/gsea/msigdb/human/collections.jsp. Accessed on 08/03/2023.

model and terminating at $\lambda_{\min}$.

The vMean and gSLOPE mean sequences were used for SGS, with $q_g = 10^{-10}$ for both datasets, $q_v = 10^{-5}$ for Cancer, and $q_v = 10^{-4}$ for Colitis. SLOPE and gSLOPE were applied using the BH sequence and gSLOPE mean sequence, respectively, using the same FDR control parameters as SGS. SGO used $\alpha = 0.99, a_1 = b_i\|\mathbf{X}^\top \mathbf{y}\|_\infty, a_2 = a_1/p$, where $b_i = i \cdot \exp\{-2\}$, and gOSCAR used the group weights of SGO.

Detailed model and algorithmic hyperparameter information is provided in Tables A.4 and A.6.

### 3.6.3 Cancer Dataset Results

The 60 patients of the breast cancer dataset were split evenly to train/test sets. The final dataset contains 6375 genes that are grouped into 550 pathways with sizes in the range $[1, 533]$. For this dataset, the path was extended to $\lambda_{\min} = 0.01\lambda_{\max}$ to allow for denser models. The development of breast cancer follows a complex genetic landscape, requiring more genes for better predictive accuracy [180].

SGS was found to outperform both SLOPE and gSLOPE, with a peak accuracy of 66.7%, in comparison to 60.0% and 50.0% for SLOPE and gSLOPE (Figure 3.13). Table 3.4 shows that it also outperformed the lasso-based models (Figure B.15) and OSCAR-based models (Figure B.16). The optimal SGS model was found at the peak index of 67 and selected 32 genes from 20 pathways (Table B.3). From the associated genes found by SGS, COX6A1 and SUSD3 have been previously shown to have an association with breast cancer [94, 225].

Additional fitting information, including the number of selected variables and groups for each method, can be found in Appendix B.4.

| **Dataset** | **SLOPE-based** | | | **Lasso-based** | | | **OSCAR-based** | | |
|---|---|---|---|---|---|---|---|---|---|
| | SGS | SLOPE | gSLOPE | SGL | Lasso | gLasso | SGO | OSCAR | gOSCAR |
| Colitis | **97.4** | 94.8 | 84.4 | 92.2 | 93.5 | 89.6 | 88.3 | 93.5 | 88.3 |
| Cancer | **66.7** | 60.0 | 56.7 | 50.0 | 56.7 | 36.7 | 63.3 | 56.7 | 56.7 |

Table 3.4: Peak classification rate (%) for the SLOPE-, lasso-, and OSCAR-based models applied to the Colitis and Cancer datasets. **Bold** represents the best value for each dataset.

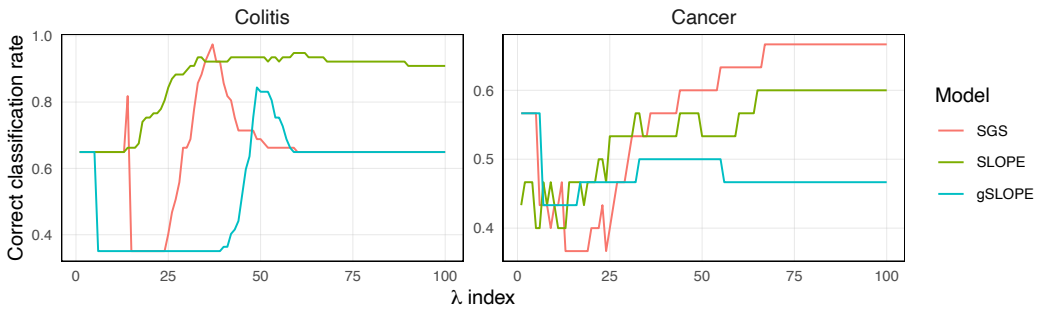



Figure 3.13: Correct classification rate (%) of the SLOPE-based models applied to the Colitis and Cancer datasets, along a log-linear 100-$\lambda$ regularisation path.

### 3.6.4 Colitis Dataset Results

The 127 samples of the Colitis dataset were split into 50/77 train/test set observations. The C3 pathway collection contained 1408 pathways, corresponding to 12031 genes in the dataset. For all models, the path was terminated at $\lambda_{\min} = 0.1\lambda_{\max}$.

SGS achieved the highest peak classification of the nine models considered, at 97.4%, using $\alpha = 0.99$, showing that inducing only a small amount of group sparsity is enough to improve upon the peak of 94.8% for SLOPE (Figure 3.13). The much lower peak of 84.4% for gSLOPE highlights the downside of selecting all variables within a group, as many noisy variables will enter the prediction. Interestingly, the lasso was found to have a higher classification peak than SGL, with 93.5% compared to 92.2%, but with both being lower than SGS (Table 3.4 and Figure B.15). This further highlights the benefit of inducing stringent bi-level penalisation. Figure 3.14 demonstrates that the optimal SGS model had only two misclassifications and that using only eight predictors would also have given the optimal predictive score.

At the peak index of 37, the SGS model selected nine genes from seven pathways, with the MIR3662 pathway having three active genes (Table 3.5). The gene NCK2 was found to be most strongly associated with a change in risk of colitis, which is in agreement with the findings of Burczynski et al. [34]. Among the other genes selected by SGS, TMEM158 and BASP1 were found to be up-regulated for the development of colitis by Xu et al. [216].

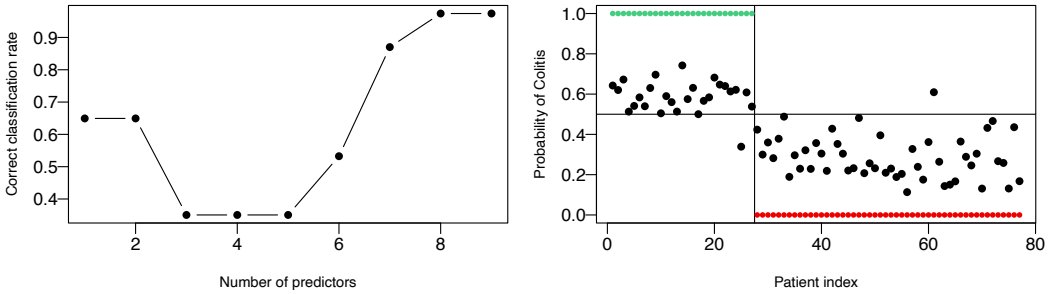



Figure 3.14: Left: correct classification rate (%) as a function of the number of predictors in the model, for the optimal SGS model applied to the Colitis dataset. The genes enter the model in order of their effect size. Right: the probability of a patient having colitis, according to the optimal SGS model. The decision boundaries are shown and the patients are coloured by whether they have the disease. Two misclassifications can be observed.

| **Symbol** | **Gene name** | **Pathway** | $\hat{\beta}$ |
|---|---|---|---|
| NCK2 | NCK adaptor protein 2 | MIR6867_5P | −0.432 |
| SUZ12 | Suppressor of zeste 12 homolog (Drosophila) | MIR607 | 0.357 |
| GOLGA8N | Golgin subfamily A member 8N | MIR3662 | 0.283 |
| ARPC5L | Actin related protein 2/3 complex, subunit 5-like | MIR4659A_3P_MIR4659B_3P | 0.256 |
| BASP1 | Brain abundant, membrane attached signal protein 1 | LET_7A_3P | −0.179 |
| C5AR1 | Complement component 5a receptor 1 | MIR153_5P | −0.158 |
| TMEM158 | Transmembrane protein 158 | MIR5582_3P | −0.107 |
| APP | Amyloid beta (A4) precursor protein | MIR3662 | −0.0692 |
| RAP1A | RAP1A, member of RAS oncogene family | MIR3662 | −0.00356 |

Table 3.5: The nine active genes as found by the optimal SGS solution for the Colitis dataset, given with their estimated coefficient value.

### 3.6.5 Hyperparameter Sensitivity for SGS

For the lasso-based models, we observe that the lasso outperforms SGL for both datasets, showing that for SGL, the grouping information provided no useful information for classification, but instead just increased the model variance. In contrast, by inducing more sparsity, SGS extracts relevant grouping information while discarding noisy variables to improve predictive performance over SLOPE. Across both datasets, SGS selected models with similar sparsity levels to those chosen by SGL and SLOPE (Table B.4), indicating that the superior performance of SGS arises not from increased model complexity, but from its ability to efficiently identify the signal and eliminate noise.

These two data examples highlight the challenges and rewards of applying SGS to real data. In comparison to SGL, SGS has adaptive penalty weights, which require two additional hyperparameters to specify ($q_v$ and $q_g$) that influence the ultimate performance of SGS. Indeed, when optimising the performance of SGS for the Colitis dataset, it obtained the highest peak (97.4%), but beyond the peak, it had a lower classification rate than SLOPE. For SGS, setting both FDR control parameters to 0.01 gave higher classification rates for the second half of the path than for $q_v = 10^{-4}$ and $q_g = 10^{-10}$, despite a lower peak of 96.1% (Figure 3.15). Even though care is required in specifying the hyperparameters for SGS, it has been shown that it consistently achieves strong predictive performance and outperforms other lasso- and SLOPE-based approaches.

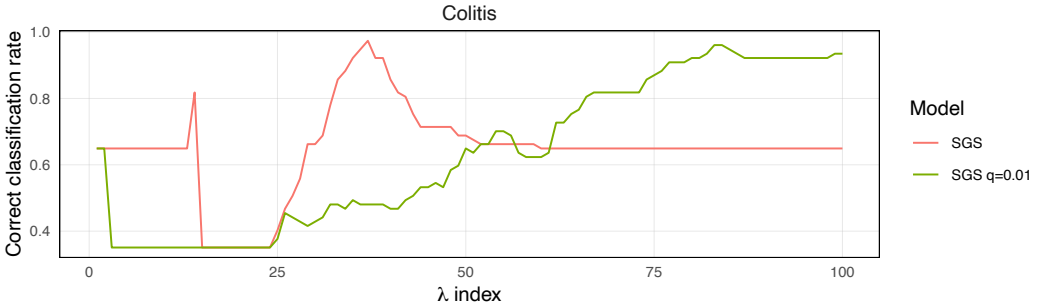



Figure 3.15: Correct classification rate (%) of SGS with $q_v = 10^{-4}, q_g = 10^{-10}$ (red) and $q_v = q_g = 0.1$ (green), applied to the Colitis dataset along a 100-$\lambda$ path.

## 3.7 Discussion

A new approach for both variable and group selection is presented, named SGS, that incorporates SLOPE into a sparse-group framework. SGS was shown theoretically and computationally to control the FDR at both variable and group levels under orthogonal designs, through newly derived variable and group sequences.

Through the analysis of synthetic and real datasets, we showcased that SGS achieves stronger bi-level selection and predictive performance than other high-dimensional models. SGS achieves this by utilising the available grouping information and applying more stringent penalisation to discard noise variables.

SGS was assessed as a disease prediction tool by applying it to two real gene expression datasets, predicting cases of colitis and breast cancer. For both datasets,

SGS obtained the highest peak classification accuracy of all models considered. In particular, SGS achieved a peak accuracy of 97.4% when applied to predicting colitis cases. To form this prediction, SGS required only nine genes, identifying genes previously associated with a change in the risk of colitis, as well as new findings. SGL struggled in comparison to the lasso, showing that, unlike SGS, it was not able to utilise the grouping information.

SGS can be slower to fit than models such as the lasso and SLOPE, due to the two layers of penalisation (both of which are adaptive). As a result, screening rules, which discard inactive features before optimisation, leading to large computational savings, are developed for SGS (and gSLOPE) in Chapter 4. Additionally, as with elastic net and SGL, the hyperparameter $\alpha$ needs to be chosen for SGS, alongside $\lambda$. In most cases, $\alpha$ tends to be set subjectively or found using a grid search approach with CV. A grid search leads to a trade-off between the FDR and sensitivity (Figure 3.2) and is computationally expensive. By reducing computational cost, the screening rules developed in Chapter 4 allow expanded tuning regimes.

Even though there are no theoretical guarantees that SLOPE-based models control the FDR for non-orthogonal data, in our extensive simulation study, SGS was found to maintain strong selection performance under different designs. In particular, it outperforms its non-adaptive counterpart SGL, highlighting the benefit of adaptive penalisation. Additionally, in comparison to SGS, gSLOPE had poor performance, selecting many false variables as a consequence of applying only group-wise sparsity. In Chapter 7, the lack of FDR control for SLOPE models under general settings is resolved using a Bayesian model selection framework. The framework also allows for estimation of $\alpha$, removing subjectivity in its choice.

# CHAPTER 4

## Strong Screening Rules for Group-based SLOPE Models

*The content of this chapter is largely based on our paper Feser and Evangelou [69], published in the Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) (2025). The screening rules developed in this chapter are implemented in the* `sgs` *R package [64], available on CRAN.*

Tuning the regularisation parameter in penalised regression models is an expensive task, requiring multiple models to be fit along a path of parameters. Strong screening rules drastically reduce computational costs by lowering the dimensionality of the input before fitting. In this chapter, we develop the first sparse-group strong screening framework, which performs two layers of screening: first discarding entire groups, and then removing irrelevant variables within the remaining groups. The framework is applied to group-based SLOPE models, Group SLOPE and Sparse-group SLOPE, by deriving model-specific strong rules using subdifferentials. The developed rules apply to the wider family of group-based OWL models. Our experiments on both synthetic and real data show that the screening rules significantly accelerate the fitting process. The screening rules make it accessible for group-based SLOPE models to be applied to high-dimensional datasets, particularly those encountered in genetics. The sparse-group strong screening framework is extended to the sparse-group lasso and its adaptive variant in Chapter 5.

We first present background material on screening rules, useful for this chapter and for Chapter 5. Following this, we proceed with the material of this chapter.

## 4.1 Screening Rules Background

As the amount of data collected increases, the emergence of high-dimensional data is becoming increasingly common in fields ranging from genetics to finance. Performing regression and discovering relevant features on these datasets is a challenging task, as classical statistical methods tend to break down. As seen in this thesis, a popular approach to meeting the challenge is through penalised regression.

A key aspect of fitting a penalised model is to tune the value of $\lambda$ along an $l$-length path $\lambda_1 \geq \cdots \geq \lambda_l \geq 0$. Several approaches exist for tuning this parameter, including CV [43, 89] and exact solution path algorithms, but these can be computationally expensive. An example of a solution path algorithm is the Least Angle Regression (LARS) [54] approach, which calculates solutions for all possible values of $\lambda$ for the lasso. However, LARS is very sensitive to multicollinearity and scales quadratically, rendering its use in high-dimensional settings limited.

*Screening rules* help alleviate the costs of fitting on a path by discarding variables that are inactive at the optimal solution, thus reducing input dimensionality prior to optimisation. The optimisation is then performed using only the variables in the screened set. If the screened set is a small proportion of the total input space, large computational savings are expected.

To formally define screening rules, denote the *active set* of coefficients at path point $\lambda_{k+1}$, for $k \in [l-1]$, by $\mathcal{A}_v(\lambda_{k+1}) = \{i \in [p] : \hat{\beta}_i(\lambda_{k+1}) \neq 0\}$. The goal of a *screening rule* is to use the solution at $\lambda_k$ to recover a *screened set* of features, $\mathcal{S}_v(\lambda_{k+1})$, which is a superset of the (unknown) set of active variables at $\lambda_{k+1}$. Although screening rules are typically defined for fitting the entire path, they can also be applied to a single path model corresponding to a specific $\lambda$ by screening on the path $(\lambda_{\max}, \lambda)$; this is referred to as basic screening.

There are two types of screening rules: *safe/exact* and *heuristic*. Safe rules are guaranteed to only discard inactive variables, but are conservative. They mostly follow the Safe Feature Elimination (SAFE) framework (Definition 4.7) [55]. Notable examples utilising the SAFE framework (for lasso-based models) include the sparse logistic regression rule (Slores) [205], correlation-based screening [215], and Dual Polytope Projections (DPP) [204], including other methods [6, 63, 142].

Heuristic rules tend to follow the strong screening rule framework (Definition 4.2), proposed by Tibshirani et al. [193], and discard considerably more variables than safe rules, at risk of violations. These violations are countered by checking

the Karush–Kuhn–Tucker (KKT) stationarity conditions (Definition 4.1) [109] and adding any offending features back into the optimisation procedure. Other strong rules include BLITZ [100], a variation of the strong rule that prioritises features based on their proximity to violating screening condition, Sure Independence Screening (SIS) [60], a rule that removes features with low correlation to the response for lasso-based models, Extended SIS (ExSIS) [3], an extension of SIS for arbitrary linear models, and the Hessian rule [110], which uses second-order information to discard features in a strong framework.

Hybrid screening rules have been proposed for lasso-based models by Zeng et al. [231] and Wang and Breheny [202], combining safe and heuristic criteria to unify their advantages and improve computational efficiency.

Aside from the safe and heuristic categories, screening rules tend to follow three forms: *static*, where the feature reduction occurs only once at the start [55, 214, 215] (the examples given are all safe rules), *dynamic*, where reduction occurs iteratively [26] (this adapts existing safe rules to make them dynamic), and *sequential*, where information from the previous solution is used [110, 111, 193] (these are all strong rules).

The rules presented in this chapter and Chapter 5 are sequential strong rules. There are approaches for screening observations alongside features [176, 234]. However, the focus of this thesis is high-dimensional settings, in which discarding features is significantly more impactful on computational savings, so these approaches are not considered.

The choice of developing strong screening over safe is motivated by two main reasons. First, strong rules discard significantly more variables than safe rules [193]. Second, safe rules require calculating the dual norm set. For SGS, this is a sum of convex sets. Determining if a point lies in this set requires knowledge of the summation's decomposability, which is a difficult task [206]. This challenge is addressed in Section 4.5.1.

### 4.1.1 Screening Definitions

**Definition 4.1** (KKT stationarity conditions [109])**.** For a general Lagrangian problem

$$\mathcal{L}(\mathbf{b}, \lambda) = f(\mathbf{b}) + \lambda J(\mathbf{b}; \mathbf{v}),$$

the *KKT stationary conditions* are given by

$$\mathbf{0} \in \nabla_{\mathbf{b},\lambda} \mathcal{L}(\mathbf{b}, \lambda). \tag{4.1}$$

**Definition 4.2** (Strong screening rule framework [193])**.** A strong screening rule is formulated through the KKT stationarity conditions (Definition 4.1) for Equation 2.1, given by, at path point $k+1$,

$$\mathbf{0} \in \nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) + \lambda_{k+1} \partial J(\hat{\boldsymbol{\beta}}(\lambda_{k+1}); \mathbf{v}), \tag{4.2}$$

where $\partial J(\cdot)$ is the subdifferential of the norm (Definition 4.5). If the gradient were available, the active set could be identified exactly by checking the subdifferential of the norm at zero:

$$\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) \in \lambda_{k+1} \partial J(\mathbf{0}; \mathbf{v}) = \{\mathbf{x} \in \mathbb{R}^p : J^*(\mathbf{x}; \lambda_{k+1}\mathbf{v}) \leq 1\}, \tag{4.3}$$

where $J^*$ is the dual norm (Definition 4.3) of $J$ and $\partial J(\mathbf{0}; \mathbf{v})$ is the unit ball of the dual norm. So, $J^*(\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})); \lambda_{k+1}\mathbf{v}) \leq 1$ indicates that $\hat{\boldsymbol{\beta}}(\lambda_{k+1}) \equiv \mathbf{0}$. As the gradient at $\lambda_{k+1}$ is not available, a model-specific approximation of the gradient is derived to find a screened subset of the features, $\mathcal{S}_v(\lambda_{k+1})$, such that $\mathcal{A}_v(\lambda_{k+1}) \subset \mathcal{S}_v(\lambda_{k+1})$. As seen in Chapters 4 and 5, an approximation is often found via a Lipschitz assumption (Definition 4.4). The framework is designed for separable norms and requires access to the dual norm.

**Definition 4.3** (Dual norm)**.** Let $X$ be a normed vector space with primal norm $J(\cdot)$, then the *dual norm* is given by, for $\mathbf{x} \in X$,

$$J^*(\mathbf{x}; \mathbf{v}) = \sup \left\{ \mathbf{x}^\top \mathbf{z} : J(\mathbf{z}; \mathbf{v}) \leq 1 \right\}.$$

A key property of the dual norm is that the dual norm of the dual norm is the primal norm, *i.e.*, $J^{**}(\cdot) = J(\cdot)$.

**Definition 4.4** (Lipschitz function)**.** Given two normed spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$, a function $f : X \to Y$ is called a *Lipschitz function* if there exists a real (Lipschitz) constant $K \geq 0$ such that, for all $x_1, x_2 \in X$,

$$\|f(x_1) - f(x_2)\|_Y \leq K \|x_1 - x_2\|_X.$$

The general definition of a Lipschitz function is less restrictive as the normed spaces can be any metric space, but we have given a more relevant version here.

**Definition 4.5** (Subdifferential of a norm)**.** The *subdifferential* of a norm $J(\cdot)$ is given by [138]

$$\partial J(\mathbf{b}; \mathbf{v}) = \begin{cases} \{\mathbf{x} \in \mathbb{R}^p \ : \ J^*(\mathbf{x}; \mathbf{v}) \leq 1\} & \text{at zero,} \\ \left\{\mathbf{x} \in \mathbb{R}^p \ : \ J^*(\mathbf{x}; \mathbf{v}) = 1 \text{ and } \mathbf{x}^\top \mathbf{b} = J(\mathbf{b}; \mathbf{v})\right\} & \text{otherwise.} \end{cases}$$

**Definition 4.6** (Dual problem and Fenchel conjugate)**.** For a general penalised model, there exists a *dual formulation* [8]

$$\hat{\boldsymbol{\theta}} \in \underset{\mathbf{b} \in \mathbb{R}^p, J^*(\mathbf{b}; \mathbf{v}) \leq \lambda}{\arg\min} -f^*(\mathbf{b}),$$

where $f^*$ is the *Fenchel conjugate* function of the loss, given by,

$$f^*(\mathbf{b}) = \max_{\mathbf{x} \in \mathbb{R}^p} \left\{\mathbf{b}^\top \mathbf{x} - f(\mathbf{x})\right\}.$$

**Definition 4.7** (SAFE framework [55])**.** For a dual solution $\hat{\boldsymbol{\theta}}(\lambda)$, Fermat's subdifferential inclusion rule [20, 138] states that, at path point $k+1$,

$$\mathbf{X}_i^\top \hat{\boldsymbol{\theta}}(\lambda_{k+1}) \in \partial J(\hat{\beta}_i(\lambda_{k+1}); v_i), \ \forall i \in [p].$$

Using Definition 4.5, if $J^*(\mathbf{X}_i^\top \hat{\boldsymbol{\theta}}(\lambda_{k+1}); v_i) < 1$, then $\hat{\beta}_i(\lambda_{k+1}) = 0$. This allows the inactive variables to be identified exactly. However, as with the strong framework (Definition 4.2), there is an unknown term. The *SAFE framework* deals with this by constructing a set $\mathcal{R}$, called a *safe region*, that contains the unknown dual solution $\hat{\boldsymbol{\theta}}(\lambda_{k+1})$, for $i \in [p]$,

$$\max_{\boldsymbol{\theta} \in \mathcal{R}} J^*(\mathbf{X}_i^\top \boldsymbol{\theta}; v_i) < 1 \implies J^*(\mathbf{X}_i^\top \hat{\boldsymbol{\theta}}(\lambda_{k+1}); v_i) < 1 \implies \hat{\beta}_i(\lambda_{k+1}) = 0.$$

Constructing a suitable safe region is part of the challenge of developing a safe rule: the region should be as small as possible (and still contain the dual solution), and the computation of the dual norm should be cheap. As such, safe regions are usually chosen to be simple geometric objects, such as balls or domes. Due to the strict requirements on these safe regions, the reduction is generally more conservative in comparison to strong rules.

## 4.2 Introduction

In Chapter 3, SLOPE models, particularly SGS, were shown to be powerful tools for disease prediction. However, their reliance on sorting operations and proximal algorithms can make model fitting computationally intensive. This issue is amplified for SGS, which involves two layers of sorting and penalisation. Consequently, developing efficient screening rules for SLOPE models could substantially reduce computational cost and facilitate their application to a broader range of problems.

Both safe [12, 57] and strong [111] rules have been proposed for SLOPE (Section 2.2.2), as well as exact solution path algorithms [52, 140]. As SLOPE is a non-separable penalty, safe screening requires repeated screening during optimisation, which is expensive due to the repeated dual norm evaluations required for the safe regions [111]. The strong framework cannot be applied to SLOPE, since it is limited to separable penalties. Instead, the subdifferential is used directly to derive the SLOPE strong rule [111]: it presents a condition under which a variable is zero, which can be checked using a gradient approximation.

No screening rules exist for group-based SLOPE models.[1] The extension of strong rules from the lasso to the group lasso was facilitated by the strong framework, which required only knowledge of the dual norm. A similar extension from SLOPE to gSLOPE and SGS is not possible via the framework, as it is restricted to single separable norms. As such, our rules directly use the subdifferential, as is used for the SLOPE strong rule.

A screening rule for SGS should perform bi-level screening to fully exploit the sparse-group nature of its norm. Safe rules that perform bi-level screening exist for SGL [137, 206], but there are no such strong rules (the framework developed here is extended to SGL and the adaptive SGL in Chapter 5).

### 4.2.1 Contributions and Organisation of this Chapter

In this chapter, we propose the first strong screening framework in the literature for sparse-group norms (Section 4.3). The framework applies two layers of screening: first, inactive groups are discarded, and then inactive variables in remaining groups, supplemented by KKT checks to ensure optimality. Strong rules are developed for gSLOPE (Section 4.4) and SGS (Section 4.5) that are incorporated into the

[1] A safe rule was proposed for group SLOPE in Bao et al. [13], published after our rules were introduced in Feser and Evangelou [68].

framework. The screening requires knowledge of the subdifferentials, necessitating a general derivation for gSLOPE (Theorem 4.1).

The analysis of synthetic (Section 4.6) and real (Section 4.7) datasets shows that our proposed screening rules considerably improve runtime. The reduced input dimensionality from the screening also eases convergence issues with large datasets. These improvements are achieved without affecting solution optimality. Our proposed screening rules enable efficient fitting of group-based SLOPE models, achieving computational times comparable to their simpler non-adaptive counterparts and thus promoting broader practical adoption.

As the proposed screening rules only require that the penalty sequences $(\mathbf{v}, \mathbf{w})$ are ordered, they apply to the wider class of group-based OWL models.

## 4.3 Screening Rule Framework

We introduce a new sparse-group framework, based on the strong framework, for applying strong screening to sparse-group norms, allowing for bi-level screening. By applying bi-level screening, a substantially larger proportion of variables are discarded than with group screening alone.

The framework is presented in Algorithm 2. First, a screened set of groups is computed, $\mathcal{S}_g$. An additional layer of screening is then performed to compute $\mathcal{S}_v$ using $\mathcal{S}_g$. This, combined with the previously active variables, forms the reduced input set for fitting, $\mathcal{E}_v$ (the *optimisation set*). Tibshirani et al. [193] also propose using the ever-active set, comprising all variables active at any point along the path. However, because most penalised regression models are not nested (*i.e.*, variables enter and leave the model along the path), this approach can result in overly large optimisation sets. Consistent with the SLOPE strong rule, we therefore use the previously active set. KKT checks are performed on $\mathcal{E}_v$ to ensure no violations have occurred (Sections 4.4.3 and 4.5.3).

The screening for gSLOPE (Section 4.4) does not perform the variable screening step in the framework and instead takes $\mathcal{S}_v$ as all variables in the groups of $\mathcal{S}_g$. The KKT checks are then performed only on the groups.

The main cost of the framework is calculating the fitted values. For $t$ iterations of ATOS (the algorithm used to fit gSLOPE and SGS, see Section 3.3.1), the convergence rate is $\mathcal{O}(1/t)$ [147] with a time complexity of $\mathcal{O}(tp^2)$ (for a general proximal algorithm [237]).

Appendix C.3 describes the implementation of the framework for gSLOPE and SGS.

**Algorithm 2** Sparse-group screening framework

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

## 4.4 Group SLOPE Screening

The screening rule for gSLOPE is formulated using the subdifferential, which is derived next.

### 4.4.1 Subdifferential

The strong rule for gSLOPE is formulated by checking the zero condition of the subdifferential (Equation 4.3), as derived in Theorem 4.1. To derive the subdifferential, we define the operator

$$[\mathbf{b}]_{\mathcal{G},q} := (p_1^q \|\mathbf{b}^{(1)}\|_2, \dots, p_m^q \|\mathbf{b}^{(m)}\|_2)^\top, \tag{4.4}$$

as being applied to all groups in the partition $\mathcal{G}$ with quotient $q$. In particular, $[\mathbf{b}]_{\mathcal{G}_\mathcal{Z},-0.5}$ is the operator applied only to the inactive groups with $q = -0.5$, where $\mathcal{Z} := \{j \in [m] : \|\hat{\beta}^{(j)}\|_2 = 0\}$ such that $\mathcal{G}_\mathcal{Z}$ is the variable indices for the inactive groups.

**Theorem 4.1** (gSLOPE subdifferential)**.** *For a general grouping structure $\mathcal{G}$, the subdifferential for gSLOPE is given by*

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

*Proof of Theorem 4.1.* The proof is similar to that of Theorem 2.7 in Brzyski et al. [30], where the subdifferential of gSLOPE is derived under equal groups. It is derived here under more general terms, *i.e.*, a general grouping structure $\mathcal{G}$. As such, the subdifferential must be derived separately under both inactive and active groups.

*Case 1:* For inactive groups, we consider the subdifferential at zero. The subdifferential of a norm at zero is given by the dual norm of the unit ball [171]

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

The dual norm for gSLOPE is given by [30]

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

Hence, the dual norm unit ball is

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

where $\mathcal{B}_{J^*_{\text{slope}}(\mathbf{0}; \mathbf{w})}[0, 1] = \{\mathbf{x} \in \mathbb{R}^m : \operatorname{cumsum}(|\mathbf{x}|_{\downarrow} - \mathbf{w}) \preceq \mathbf{0}\}$ is the unit ball of the dual norm to $J_{\text{slope}}$ such that $(\cdot)_{\downarrow}$ sorts a vector into decreasing form [111] and $\operatorname{cumsum}(x) = [x_1, x_1 + x_2, \ldots, \sum_{i=1}^{\operatorname{card}(x)} x_i]$. Using this, the subdifferential at zero for the inactive groups is given by

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

*Case 2:* Without loss of generality, denote the group index $s$ such that $\|\boldsymbol{\beta}^{(j)}\|_2 = 0$ for $j > s$ (inactive groups) and $\|\boldsymbol{\beta}^{(j)}\|_2 \neq 0$ for $j \leq s$ (active groups). In other words, $j \in \mathcal{G}_{\mathcal{A}}$ if $j \leq s$. Define a set

$$D = \big\{\mathbf{d} \in \mathbb{R}^p : \|\boldsymbol{\beta}^{(1)} + \mathbf{d}^{(1)}\|_2 > \cdots > \|\boldsymbol{\beta}^{(s)} + \mathbf{d}^{(s)}\|_2, \\ \|\boldsymbol{\beta}^{(s)} + \mathbf{d}^{(s)}\|_2 > \|\mathbf{d}^{(j)}\|_2, j > s\big\}.$$

By definition of a subdifferential, if $\mathbf{x} \in \partial J_{\text{gslope}}(\boldsymbol{\beta}; \mathbf{w})$, then for all $\mathbf{d} \in D$

$$\sum_{j=1}^{m} \sqrt{p_j} w_j \|\boldsymbol{\beta}^{(j)} + \mathbf{d}^{(j)}\|_2 \geq \sum_{j=1}^{m} \sqrt{p_j} w_j \|\boldsymbol{\beta}^{(j)}\|_2 + \mathbf{x}^\top \mathbf{d}.$$

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

$$\begin{aligned} \sum_{j=1}^{s} \sqrt{p_j} w_j \|\boldsymbol{\beta}^{(j)} + \mathbf{d}^{(j)}\|_2 + \sum_{j=s+1}^{m} \sqrt{p_j} w_j \|\mathbf{d}^{(j)}\|_2 \geq & \sum_{j=1}^{s} \sqrt{p_j} w_j \|\boldsymbol{\beta}^{(j)}\|_2 \\ & + \sum_{j=1}^{s} \mathbf{x}^{(j)\top} \mathbf{d}^{(j)} + \sum_{j=s+1}^{m} \mathbf{x}^{(j)\top} \mathbf{d}^{(j)}. \end{aligned} \tag{4.5}$$

Now, for $j \in \mathcal{G}_{\mathcal{A}}$, define a new set $\tilde{D}_j = \{\mathbf{d} \in D : \mathbf{d}^{(k)} \equiv \mathbf{0}, k \neq j\}$. Taking $\mathbf{d} \in \tilde{D}_j$, Equation 4.5 becomes

$$\sqrt{p_j} w_j \|\boldsymbol{\beta}^{(j)} + \mathbf{d}^{(j)}\|_2 \geq \sqrt{p_j} w_j \|\boldsymbol{\beta}^{(j)}\|_2 + \mathbf{x}^{(j)\top} \mathbf{d}^{(j)}.$$

Since the set $\{\mathbf{d}^{(j)} : \mathbf{d} \in \tilde{D}_j\}$ is open in $\mathbb{R}^{p_j}$ and contains zero, by Corollary G.1 in Brzyski et al. [30], it follows that $\mathbf{x}^{(j)} \in \partial f_g(\boldsymbol{\beta}^{(j)})$ where $f_g : \mathbb{R}^{p_j} \to \mathbb{R}, f_g(\mathbf{x}) = w_j \sqrt{p_j} \|\mathbf{x}\|_2$. Now, for $j \leq s$, $f_g$ is differentiable in $\boldsymbol{\beta}^{(j)}$, giving

$$\mathbf{x}^{(j)} = w_j \sqrt{p_j} \frac{\boldsymbol{\beta}^{(j)}}{\|\boldsymbol{\beta}^{(j)}\|_2}.$$

□

### 4.4.2 Screening Rule

The choice of $q = -0.5$ in the operator defined in Equation 4.4 leads to $J^*_{\text{gslope}}(\mathbf{x}; \mathbf{w}) = J^*_{\text{slope}}([\mathbf{x}]_{\mathcal{G},-0.5})$, which allows the gSLOPE subdifferential (Theorem 4.1) to be written using the SLOPE subdifferential [30]. Combining the KKT conditions at zero (Equation 4.3) with the gSLOPE subdifferential (Theorem 4.1) provides a condition that a group is inactive at path point $k+1$ if

$$h(\lambda) := ([\nabla f(\hat{\boldsymbol{\beta}}(\lambda))]_{\mathcal{G},-0.5})_{\downarrow} \in \partial J_{\text{slope}}(\mathbf{0}; \lambda \mathbf{w}).$$

Using the subdifferential of SLOPE (provided in Appendix C.1.1), this is given by

$$\text{cumsum}(h(\lambda_{k+1}) - \lambda_{k+1}\mathbf{w}) \preceq \mathbf{0}. \tag{4.6}$$

This screening condition is checked using the algorithm proposed for the SLOPE strong rule (Algorithm 3), which leads to the strong rule for gSLOPE (Proposition 4.1). The summation in the if step of the algorithm is checking the condition given by Equation 4.6 for each cumsum vector entry until there is a violation, at which point the screened set is updated.

The algorithm exploits the fact that indices corresponding to inactive predictors are ordered last in the input $c$ and in the features $|\hat{\boldsymbol{\beta}}|_\downarrow$, due to the SLOPE subdifferential's invariance to permutations [111]. This property allows the sparsity pattern of $\hat{\boldsymbol{\beta}}$ to be determined directly from $\nabla f(\hat{\boldsymbol{\beta}})$.

**Algorithm 3** SLOPE subdifferential algorithm from Larsson et al. [111]

**Input:** $\mathbf{c} \in \mathbb{R}^p, \boldsymbol{\phi} \in \mathbb{R}^p$, where $\boldsymbol{\phi}_1 \geq \cdots \geq \boldsymbol{\phi}_p \geq 0$
$\mathcal{S}, \mathcal{B} \leftarrow \varnothing$
**for** $i = 1$ **to** $p$ **do**
  $\mathcal{B} \leftarrow \mathcal{B} \cup \{i\}$
  **if** $\sum_{j \in \mathcal{B}}(c_j - \phi_j) \geq 0$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{B}$
    $\mathcal{B} \leftarrow \varnothing$
  **end if**
**end for**
**Output:** $\mathcal{S}$

**Proposition 4.1** (Strong screening rule for gSLOPE)**.** *Taking* $\mathbf{c} = h(\lambda_{k+1})$ *and* $\boldsymbol{\phi} = \lambda_{k+1}\mathbf{w}$ *as inputs for Algorithm 3 returns a superset* $\mathcal{S}_g(\lambda_{k+1})$ *of the active set* $\mathcal{A}_g(\lambda_{k+1})$.

*Proof of Proposition 4.1.* Suppose we have $\mathcal{B} \neq \varnothing$ after running the algorithm. Then, plugging in $h(\lambda_{k+1}) = ([\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))]_{\mathcal{G},-0.5})_\downarrow$ gives

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

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

### Gradient approximation

To use Proposition 4.1, the gradient at path index $k+1$ is needed for the calculation of $h(\lambda_{k+1})$, which is not available in practice. Instead, an approximation is required that does not lead to any violations in Algorithm 3. By the if condition in this algorithm, an approximation for a group $j \in [m]$ is needed such that $h_j(\lambda_{k+1}) \leq h_j(\lambda_k) + R_j$, where $R_j \geq 0$ is a group-specific constant.

An approximation is found by assuming that $h_j(\lambda_{k+1})$ is a Lipschitz function (Definition 4.4) of $\lambda_{k+1}$ with respect to the $\ell_1$ norm, *i.e.*,

$$|h_j(\lambda_{k+1}) - h_j(\lambda_k)| \leq w_j|\lambda_{k+1} - \lambda_k|.$$

By again noting that $J^*_{\text{gslope}}(\mathbf{x}) = J^*_{\text{slope}}([\mathbf{x}]_{\mathcal{G},-0.5})$, it can be seen that the assumption is equivalent to the Lipschitz assumptions used for the lasso [193] and SLOPE [111] strong rules. By the reverse triangle inequality,

$$|h_j(\lambda_{k+1})| \leq |h_j(\lambda_k)| + \lambda_k w_j - \lambda_{k+1} w_j.$$

This leads to the choice of constant $R_j = \lambda_k w_j - \lambda_{k+1} w_j$ and the corresponding gradient approximation strong screening rule (Proposition 4.2). To apply gSLOPE screening on a path in practice, Proposition 4.3 describes the calculation of the first path value.

**Proposition 4.2** (Gradient approximation strong screening rule for gSLOPE)**.** *Denote $r_v(\cdot)$ as the operator that returns the index of a vector sorted into decreasing absolute form. Then, taking $\mathbf{c} = h(\lambda_k) + \lambda_k \mathbf{w} - \lambda_{k+1}\mathbf{w}$ and $\boldsymbol{\phi} = \lambda_{k+1}\mathbf{w}$ as inputs for Algorithm 3, and assuming that for any $k \in [l-1]$,*

$$|h_j(\lambda_{k+1}) - h_j(\lambda_k)| \leq w_j|\lambda_{k+1} - \lambda_k|, \ \forall j \in [m],$$

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

*Proof of Proposition 4.2.* Since $\text{cumsum}(\mathbf{y}) \succeq \text{cumsum}(\mathbf{x}) \iff \mathbf{y} \succeq \mathbf{x}$ [111], we only need to show that for a group $j \in [m]$,

$$|h_j(\lambda_{k+1})| \leq |h_j(\lambda_k)| + \lambda_k w_j - \lambda_{k+1} w_j.$$

Applying the reverse triangle inequality to the Lipschitz assumption gives

$$|h_j(\lambda_{k+1})| - |h_j(\lambda_k)| \leq |h_j(\lambda_{k+1}) - h_j(\lambda_k)| \leq \lambda_k w_j - \lambda_{k+1} w_j$$
$$\implies |h_j(\lambda_{k+1})| \leq |h_j(\lambda_k)| + \lambda_k w_j - \lambda_{k+1} w_j.$$

□

**Proposition 4.3** (gSLOPE path start)**.** *For gSLOPE, the path value at which the first group enters the model is given by*

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

*Proof of Proposition 4.3.* When all groups are zero, the gSLOPE KKT conditions are

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

This is satisfied when

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

Rearranging this gives

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

Picking the maximum possible $\lambda$ such that this holds yields

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

This can be verified by noting that $\lambda_{\max} = J^*_{\text{gslope}}(\nabla f(\mathbf{0}); \mathbf{w})$ [138]. So, as $J^*_{\text{gslope}}(\mathbf{x}; \mathbf{w}) = J^*_{\text{slope}}([\mathbf{x}]_{\mathcal{G},-0.5}; \mathbf{w})$ and the dual norm of SLOPE is given by [139]

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

Therefore, $\lambda_{\max}$ is as before. □

### 4.4.3 KKT Checks

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

$$\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) \in \lambda_{k+1} \partial J_{\text{gslope}}(\hat{\boldsymbol{\beta}}(\lambda_{k+1}); \mathbf{w}).$$

Hence, the conditions check whether the gradient of the loss function lies within the subdifferential of the penalty. Since the goal is only to identify incorrectly discarded groups, it suffices to evaluate the subdifferential condition at zero. Therefore, a violation occurs if a group is discarded but

$$\begin{aligned}
&\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) \notin \lambda_{k+1} \partial J_{\text{gslope}}(\mathbf{0}; \mathbf{w}_{\mathcal{G}_\mathcal{Z}}) \\
\implies &\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) \notin \left\{ \mathbf{x} \in \mathbb{R}^{\operatorname{card} \mathcal{G}_\mathcal{Z}} : [\mathbf{x}]_{\mathcal{G}_\mathcal{Z}, -0.5} \in \partial J_{\text{slope}}(\mathbf{0}; \lambda_{k+1} \mathbf{w}_{\mathcal{G}_\mathcal{Z}}) \right\} \\
\implies &[\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))]_{\mathcal{G}_\mathcal{Z}, -0.5} \notin \partial J_{\text{slope}}(\mathbf{0}; \lambda_{k+1} \mathbf{w}_{\mathcal{G}_\mathcal{Z}}) \\
\implies &\operatorname{cumsum}(([\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))]_{\mathcal{G}_\mathcal{Z}, -0.5})_\downarrow - \lambda_{k+1} \mathbf{w}_{\mathcal{G}_\mathcal{Z}}) \succ \mathbf{0}.
\end{aligned}$$

## 4.5 Sparse-group SLOPE Screening

This section presents the group and variable screening rules for SGS. The proofs to many of the results in this section are similar to those used for gSLOPE (Section 4.4) and are presented in Appendix C.2.1.

The screening rules are derived using the SGS KKT conditions, formulated in terms of SLOPE and gSLOPE (by the sum rule of subdifferentials),

$$\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) \in \lambda_{k+1} \alpha \partial J_{\text{slope}}(\hat{\boldsymbol{\beta}}(\lambda_{k+1}); \mathbf{v}) + \lambda_{k+1}(1-\alpha) \partial J_{\text{gslope}}(\hat{\boldsymbol{\beta}}(\lambda_{k+1}); \mathbf{w}). \quad (4.7)$$

### 4.5.1 Group Screening

Combining the SGS KKT conditions with the condition for a group to be inactive in the gSLOPE subdifferential (Equation 4.6), we have

$$\operatorname{cumsum}\left( \left( [\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) + \lambda_{k+1} \alpha \partial J_{\text{slope}}(\mathbf{0}; \mathbf{v})]_{\mathcal{G}, -0.5} \right)_\downarrow - \lambda_{k+1}(1-\alpha) \mathbf{w} \right) \preceq \mathbf{0}. \quad (4.8)$$

The problem reduces to a form similar to the gSLOPE screening rule (Section 4.4), with inputs for Algorithm 3 given by

$$\mathbf{c} = ([\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}) + \lambda_{k+1}\alpha\partial J_{\text{slope}}(\mathbf{0};\mathbf{v})]_{\mathcal{G},-0.5})_{\downarrow} \quad \text{and} \quad \boldsymbol{\phi} = \lambda_{k+1}(1-\alpha)\mathbf{w}.$$

However, this input contains an unknown quantity, $\partial J_{\text{slope}}(\mathbf{0};\mathbf{v})$. To determine its form for an inactive group, the term inside the $[\cdot]_{\mathcal{G},q}$ operator in Equation 4.8 needs to be as small as possible for the inactive condition to be met. To achieve this, the soft thresholding operator is used, shown by the following lemma.

**Lemma 4.1.** *In Equation 4.8, choosing $\partial J_{slope}(\mathbf{0};\mathbf{v}) = S(\nabla f(\boldsymbol{\beta}), \lambda\alpha\mathbf{v})$, under general $\lambda$, as the soft thresholding operator minimises $[\nabla f(\boldsymbol{\beta}) + \lambda\alpha\partial J_{slope}(\mathbf{0};\mathbf{v})]_{\mathcal{G}}$, where*

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

*Proof.* To determine the form of the quantity $\partial J_{\text{slope}}(\mathbf{0};\mathbf{v})$, consider that for Equation 4.8 to be satisfied, the term

$$\nabla f(\boldsymbol{\beta}(\lambda)) + \lambda\alpha\partial J_{\text{slope}}(\mathbf{0};\mathbf{v})$$

needs to be as close to zero as possible for each group in the $[\cdot]_{\mathcal{G},-0.5}$ operator. The subdifferential term we aim to pick is given in general form by [111]

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

First, note that $\text{cumsum}(|\mathbf{y}|) \preceq \text{cumsum}(\mathbf{v}) \iff |\mathbf{y}| \preceq \mathbf{v}$. Then, considering a general group $j \in [m]$ and a general variable in that group $i \in G_j$ we have the following cases,

1. $\nabla_i f(\boldsymbol{\beta}) > \lambda\alpha v_i$: choose $y_i = -v_i$ so that the term is minimised as $\nabla f(\boldsymbol{\beta}(\lambda)) - \lambda\alpha v_i$.

2. $\nabla_i f(\boldsymbol{\beta}) < -\lambda\alpha v_i$: choose $y_i = v_i$ so that the term is minimised as $\nabla f(\boldsymbol{\beta}(\lambda)) + \lambda\alpha v_i$.

3. $\nabla_i f(\beta) \in [-\lambda\alpha v_i, \lambda\alpha v_i]$: choose $y_i = -\nabla_i f(\beta)/\lambda\alpha$ so that the term becomes zero. This is a valid choice as $\nabla_i f(\beta)/\lambda\alpha \in [-v_i, v_i]$.

Hence, the term that minimises the subdifferential is the soft thresholding operator

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

□

*Remark* 4.1. Lemma 4.1, via the soft-thresholding operator, provides a valuable connection between SGS and SGL, as the operator is used in the gradient update step for SGL [178]. This connection has the potential to lead to new and more efficient optimisation approaches for SGS that are more closely related to those used to solve SGL. This is a research direction explored in the development of the coordinate descent algorithm for SLOPE in Larsson et al. [112].

Using the soft-thresholding operator, the (theoretical) strong group screening rule for SGS is presented in Proposition 4.4. Applying a similar Lipschitz assumption as for the gSLOPE rule gives the gradient approximation strong group screening rule for SGS (Proposition 4.5).

**Proposition 4.4** (Strong group screening rule for SGS)**.** *Let*

$$\tilde{h}(\lambda) := ([S(\nabla f(\hat{\boldsymbol{\beta}}(\lambda)), \lambda\alpha\mathbf{v})]_{\mathcal{G},-0.5})_{\downarrow}.$$

*Taking* $\mathbf{c} = \tilde{h}(\lambda_{k+1})$ *and* $\boldsymbol{\phi} = (1-\alpha)\lambda_{k+1}\mathbf{w}$ *as inputs for Algorithm 3 returns a superset* $\mathcal{S}_g(\lambda_{k+1})$ *of the active set* $\mathcal{A}_g(\lambda_{k+1})$.

**Proposition 4.5** (Gradient approximation strong group screening rule for SGS)**.** *Let* $\tilde{h}(\lambda)$ *be as in Proposition 4.4. Taking* $\mathbf{c} = \tilde{h}(\lambda_k) + \lambda_k(1-\alpha)\mathbf{w} - \lambda_{k+1}(1-\alpha)\mathbf{w}$ *and* $\boldsymbol{\phi} = \lambda_{k+1}(1-\alpha)\mathbf{w}$ *as inputs for Algorithm 3, and assuming that for any* $k \in [l-1]$,

$$\left|\tilde{h}_j(\lambda_{k+1}) - \tilde{h}_j(\lambda_k)\right| \leq (1-\alpha)w_j|\lambda_{k+1} - \lambda_k|,\ \forall j \in [m],$$

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

### 4.5.2 Variable Screening

The input dimensionality for SGS can be further reduced by considering the sparse-group nature of the penalty via a second layer of variable screening. The KKT

conditions (Equation 4.7) for inactive variables in active groups are

$$\nabla_{\mathcal{G}_{\mathcal{A}_g}} f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) \in \lambda_{k+1}\alpha\partial J_{\text{slope}}(\mathbf{0}; \mathbf{v}_{\mathcal{G}_{\mathcal{A}_g}}). \tag{4.9}$$

The gSLOPE subdifferential term disappears as the numerator is zero in Theorem 4.1. The problem reduces to that of SLOPE screening, applied to the variables in groups in $\mathcal{A}_g$ and scaled by $\alpha$. The gradient approximated rule is shown in Proposition 4.7 (see Proposition 4.6 for the non-approximated version). To apply screening on a path for SGS in practice, Proposition 3.1 in Chapter 3 describes the calculation of the first path value.

**Proposition 4.6** (Strong variable screening rule for SGS)**.** *Let* $\bar{h}(\lambda) := |\nabla f(\hat{\boldsymbol{\beta}}(\lambda))|_{\downarrow}$. *Taking* $\mathbf{c} = \bar{h}(\lambda_{k+1})$ *and* $\boldsymbol{\phi} = \lambda_{k+1}\alpha\mathbf{v}$*, for the variables contained in the groups in* $\mathcal{A}_g(\lambda_{k+1})$*, as inputs for Algorithm 3 returns a superset* $\mathcal{S}_v(\lambda_{k+1})$ *of the active set* $\mathcal{A}_v(\lambda_{k+1})$.

**Proposition 4.7** (Gradient approximation strong variable screening rule for SGS)**.** *Let* $\bar{h}(\lambda) := (\nabla f(\hat{\beta}(\lambda)))_{|\downarrow|}$*, where* $(\cdot)_{|\downarrow|}$ *sorts a vector into decreasing absolute form. Taking* $\mathbf{c} = |\bar{h}(\lambda_{k+1})| + \lambda_k\alpha\mathbf{v} - \lambda_{k+1}\alpha\mathbf{v}$ *and* $\boldsymbol{\phi} = \lambda_{k+1}\alpha\mathbf{v}$*, for the variables in the groups in* $\mathcal{A}_g(\lambda_{k+1})$*, as inputs for Algorithm 3, and assuming that for any* $k \in [l-1]$*,*

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

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

*Remark* 4.2. In practice, $\mathcal{A}_g(\lambda_{k+1})$ is not available, as this is precisely what any screening rule attempts to superset. However, Proposition 4.5 guarantees that it is contained in $\mathcal{S}_g(\lambda_{k+1})$, so that this set can be used instead. That is, the variable screening rule is applied to the screened group set. Any violations caused by this replacement are checked for by the KKT checks, which are performed in any case (for any strong rule).

### 4.5.3 KKT Checks

For SGS, the KKT conditions (Equation 4.7) are first checked at the group level for inactive groups. Further variable checks are performed for variables that are either in violating or active groups (to check whether the variables should also be active). The violating variables from these secondary variable checks are added back into $\mathcal{E}_v$.

### Group Checks

A group violation occurs if the KKT conditions do not hold at the group level (Equation 4.8). That is, a violation occurs if a group is discarded but

$$\operatorname{cumsum}\Big(\big([\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) + \lambda_{k+1}\alpha\partial J_{\text{slope}}(\mathbf{0};\mathbf{v})]_{\mathcal{G}_{\mathcal{Z}},-0.5}\big)_{\downarrow} - \lambda_{k+1}(1-\alpha)\mathbf{w}_{\mathcal{Z}}\Big) \succ \mathbf{0}.$$

### Variable Checks

For the set of variables in a violating or active group, denoted by $\mathcal{G}_{\mathcal{K}_g \cup \mathcal{A}_g}$, a variable violation occurs if Equation 4.9 does not hold. That is, if

$$\begin{aligned} &\nabla_{\mathcal{G}_{\mathcal{K}_g \cup \mathcal{A}_g}} f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) \notin \lambda_{k+1}\alpha\partial J_{\text{slope}}(\mathbf{0}; \mathbf{v}_{\mathcal{G}_{\mathcal{K}_g \cup \mathcal{A}_g}}) \\ \implies &\operatorname{cumsum}(|\nabla_{\mathcal{G}_{\mathcal{K}_g \cup \mathcal{A}_g}} f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))| - \lambda_{k+1}\alpha\mathbf{v}_{\mathcal{G}_{\mathcal{K}_g \cup \mathcal{A}_g}}) \succ \mathbf{0}. \end{aligned}$$

Note that, unlike the group screening step in Proposition 4.5, where $\mathcal{A}_g$ is not available at the current path point (and a screened set is used instead), in the case of the KKT checks, the optimisation process has been completed, and thus the active set is available.

## 4.6 Synthetic Data Experiments

The effectiveness of our screening rules is illustrated using synthetic data. References to $\mathcal{E}, \mathcal{A}$ for group and variable metrics denote $\mathcal{E}_g, \mathcal{A}_g$ and $\mathcal{E}_v, \mathcal{A}_v$. For SGS, $\mathcal{E}_g$ represents groups with members in $\mathcal{E}_v$.

### 4.6.1 Simulation Design

A multivariate Gaussian design matrix, $\mathbf{X} \sim \boldsymbol{\mathcal{N}}_p(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^{400\times p}$, was used and the within-group correlation set to $\Sigma_{i,j} = \rho$, where $i$ and $j$ belong to the same group. The correlation and number of features were varied between $\rho \in \{0, 0.3, 0.6, 0.9\}$ and $p \in \{500, 1625, 2750, 3875, 5000\}$, producing 20 simulation cases. Each simulation case was repeated 100 times.

Two models were considered: linear and logistic. For the linear model, the output was generated as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{N}}_n(\mathbf{0}, \mathbf{I}_n)$ and for the logistic model the class probabilities were calculated using $\sigma(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{N}}_n(\mathbf{0}, \mathbf{I}_n))$, where $\sigma$ is the sigmoid function. Primarily, the results for the linear model are presented, and the results for

the logistic model are in Appendix C.5.3. Screening for logistic models was found to be as effective as for linear models (further supported by Table 4.2).

Groups of sizes between 3 and 25 were considered, of which 15% were set to active. Within each active group, 30% of the variables were set to active with signal $\beta_i \sim \mathcal{N}(0, 5)$, for $i \in S_v$.

The simulations were repeated for group-based OSCAR models (Appendix C.4). The rules were found to be very effective for these models, with similar performance as observed for the group-based SLOPE models in this section. The number of KKT violations were of a very similar level.

### 4.6.2 Model Hyperparameters

gSLOPE and SGS were fit along a log-linear path of 50 regularisation parameters, beginning at $\lambda_1 = \lambda_{\max}$ (from Propositions 4.3 and 3.1 for gSLOPE and SGS, respectively), and terminating at $\lambda_{50} = 0.05\lambda_1$. The data was $\ell_2$-standardised, and for the linear model, an intercept was used. The FDR control parameters were set to $q_v = q_g = 0.05$, and $\alpha = 0.95$ was used for SGS. gSLOPE and SGS used their default weights (see Remark 3.4). The models were optimised using ATOS, although the screening rules can be applied with any fitting algorithm.

### 4.6.3 Metrics

As the purpose of screening rules is to reduce the dimensionality of the input space and, as a result, reduce the computational cost, the following metrics are defined and used for evaluation:

- *Improvement factor = no screen time / screen time*, quantifies by how many orders the screening has improved the fitting time (high is best).
- *Input (variable) proportion* $= \mathcal{E}_v/p$, measures how much of the input space was used in the optimisation (low is best). The variable metric is used by default, unless otherwise stated. The group proportion is given by $\mathcal{E}_g/m$.
- *Screened (variable) proportion* $= \mathcal{S}_v/p$, measures how much of the input space was retained after screening (low is best). The variable metric is used by default, unless otherwise stated. The group proportion is given by $\mathcal{S}_g/m$.

Further simulation design and model implementation information can be found in Section A.3.

### 4.6.4 Results

#### Benefits of Bi-level Screening

Figure 4.1 clearly illustrates the advantages of employing the bi-level screening framework. For both $p = 500$ and $p = 5000$, the two layers of screening achieve a substantially greater reduction in input dimensionality compared to applying only group-level screening. This advantage becomes even more pronounced as dimensionality increases between the two scenarios. The inclusion of two screening layers is vital for sparse-group models, as it enables the removal of noise variables from the optimisation process, an aspect similar in nature to the limitations inherent in a group-only norm.

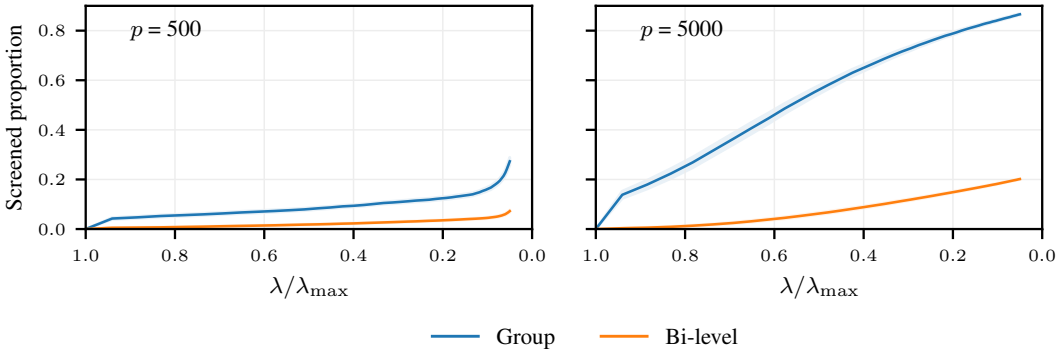



Figure 4.1: The screened proportion for group-only and bi-level screening applied to SGS, for $\rho = 0.3$, plotted along the regularisation path with 95% confidence intervals.

#### Screening Efficiency

By comparing the sizes of the fitting set ($\mathcal{E}$) to the active set ($\mathcal{A}$), the screening rules are found to be efficient in providing dimensionality reduction close to the minimum possible (the active set size) (Figures 4.2 and 4.3). As expected, the sets increase in size as $\lambda$ decreases, and the difference in size between the sets remains stable along the path, decreasing towards the termination point. The size of the fitting set remains significantly below the input size across the whole path, showing the utility of the screening. This is found to be true for any correlation, input dimensionality, and model considered (Appendix C.5.3).

The screening rules show robust dimensionality reduction in both linear and logistic models (Figure 4.4) under varying correlation. As the correlation increases,

the signal concentrates in fewer groups, causing the active group set to decrease in size. SLOPE models generally deal well with highly correlated features, as the sorted norm clusters them together [229].

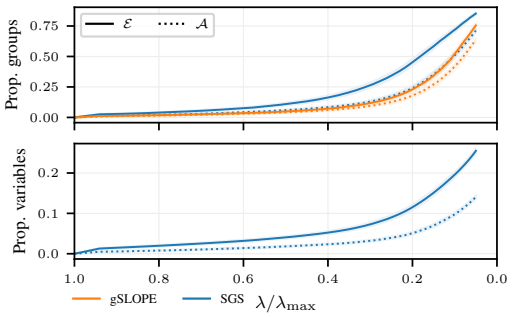



Figure 4.2: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$, relative to the input, for both gSLOPE and SGS as a function of the path for the linear model with $p = 2750, \rho = 0.6, m = 197$. The results are averaged over 100 repetitions, with 95% confidence intervals shown.

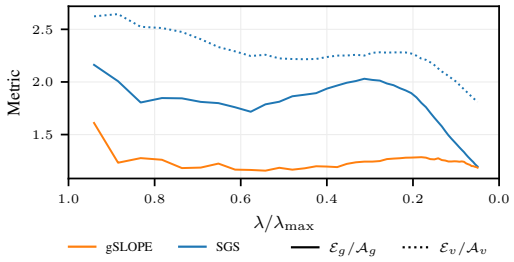



Figure 4.3: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$, relative to the full input, for gSLOPE and SGS, as a function of the regularisation path for the linear model with $p = 2750, \rho = 0.6, m = 197$. The results are averaged over 100 repetitions.

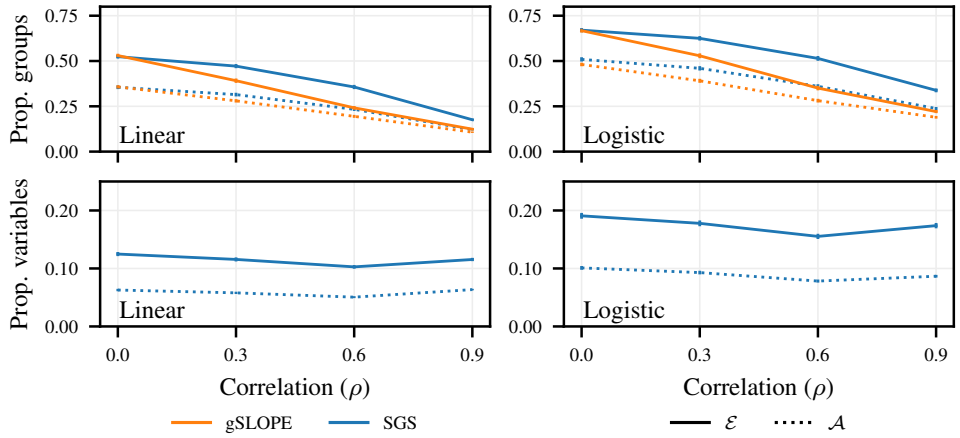



Figure 4.4: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$, relative to the full input, shown for gSLOPE and SGS under linear and logistic models. This is given as a function of the correlation ($\rho$), averaged over all cases of the input dimension ($p$), with 100 repetitions for each dimension considered ($p$), with standard errors shown.

The screening rules are found to efficiently reduce the input dimensionality on average across all cases considered (Table 4.1).

| Method | Type | Synthetic | | | Real | | |
|---|---|---|---|---|---|---|---|
| | | $\mathcal{A}$ | $\mathcal{E}$ | Dim. | $\mathcal{A}$ | $\mathcal{E}$ | Dim. |
| gSLOPE | Linear | 55 (1) | 76 (1) | 210 | 112 (3) | 168 (3) | 240 |
| gSLOPE | Logistic | 71 (1) | 97 (1) | 210 | 49 (2) | 83 (3) | 1634 |
| SGS | Linear | 178 (3) | 364 (6) | 2750 | 378 (37) | 1139 (87) | 14488 |
| SGS | Logistic | 230 (3) | 472 (6) | 2750 | 526 (5) | 979 (13) | 13734 |

Table 4.1: The cardinality of the active ($\mathcal{A}$) and fitting ($\mathcal{E}$) sets for gSLOPE and SGS, averaged across all synthetic and real data cases, split into model type. For gSLOPE, the cardinality is for the group sets, and for SGS, the variable ones. Dim. is the average input dimensionality. Standard errors are shown in grey.

### Runtime Performance

A key metric of performance for a screening rule is the time taken to fit a path of models. Figure 4.5 shows the significant runtime improvements provided by the gSLOPE and SGS screening rules, as a function of increasing input dimensionality. The gain is observed to be substantial under all correlation cases. Applying screening improves the scalability of gSLOPE and SGS models when applied to larger datasets.

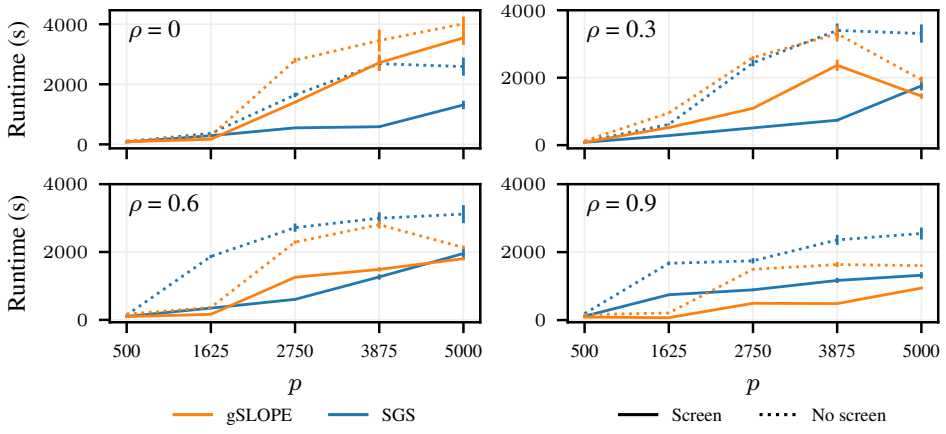



Figure 4.5: Runtime (in seconds) for fitting 50 models along a path, shown for screening against no screening as a function of input dimensionality ($p$), broken down into different correlation cases, for the linear model. The results are averaged over 100 repetitions, with standard errors shown.

The clear benefit and robustness of our screening method, with regard to run-

time, can be seen by aggregating the results of all the simulation cases (Table 4.2). For both models, in the linear and logistic cases, screening substantially improves the computational cost, halving the runtime for SGS models.

| Method | Type | Screen (s) | No screen (s) | Improvement factor |
|---|---|---|---|---|
| gSLOPE | Linear | 1016 (21) | 1623 (27) | 1.88 (0.14) |
| gSLOPE | Logistic | 814 (8) | 1409 (11) | 1.84 (0.14) |
| SGS | Linear | 735 (15) | 1830 (34) | 2.59 (0.31) |
| SGS | Logistic | 407 (2) | 859 (6) | 1.97 (0.17) |

Table 4.2: Runtime (in seconds) for fitting 50 models along a path, shown for screening against no screening, for the linear and logistic models. The results are averaged across all cases of the correlation and dimensionality, with standard errors shown in grey.

### 4.6.5 KKT Violations

As with any strong screening rule, our approach depends on assumptions that may fail. When this happens, KKT checks are used to ensure no active variables are excluded, with violations added to $\mathcal{E}_v$. For SGS, KKT violations occur at a variable level (Section 4.5.3) and for gSLOPE at a group level (Section 4.4.3).

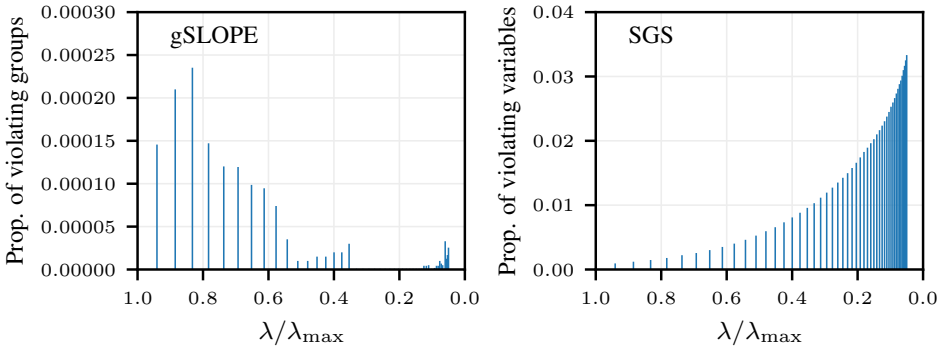



Figure 4.6: The proportion of KKT violations relative to the full input for gSLOPE and SGS, under linear models, averaged over all synthetic data cases. This is shown as a function of the regularisation path.

KKT violations are very rare for gSLOPE (Figure 4.6), occurring on the simulated data infrequently toward the start of the path. For SGS, KKT violations are more common, but still infrequent (Figure 4.6) due to additional assumptions in the

second layer of screening. In Equation 4.8, minimising the subdifferential term leads to tighter screened sets, contributing to these violations. In Figure 4.6, the number of violations increases as a function of the model density for SGS, mirroring the log-linear shape of the regularisation path. This pattern is also seen in the strong rule for SLOPE [111].

Despite these violations and the computational cost of the checks and refitting, the overall screening process yields significant runtime improvements in all synthetic cases considered.

## 4.7 Real Data Experiments

The screening rules are applied to real datasets to assess whether the computational savings observed on synthetic data translate to real-world settings.

### 4.7.1 Datasets

Seven real gene datasets of different response types and dimensionality are used. Two of the datasets, Carbotax and Scheetz, had a continuous response and were fit using a linear model. For these, the groups were generated using K-means clustering applied to the features [125]. Scheetz was also used to test the SLOPE strong rule in Larsson and Wallin [110]. The remaining five, Adenoma, Cancer, Celiac, Colitis, and Tumour, had binary labels, so a logistic model was used. For these, the design matrices contained gene expression data downloaded from NCBI's GEO database [53], so the genes were assigned to pathways (groups) using the C3 regulatory target gene sets from MSigDB [122, 188].

Further details on the datasets are provided in Table 4.3 and Appendix A.3.3.

| **Dataset** | $p$ | $n$ | $m$ | **Group sizes** | **Grouping** | **Response** | **Downloaded on** |
|---|---|---|---|---|---|---|---|
| Carbotax | 10000 | 101 | 200 | $[1, 126]$ | K-means | Continuous | 08/2024 |
| Scheetz | 18975 | 120 | 379 | $[1, 165]$ | K-means | Continuous | 08/2024 |
| Adenoma | 17661 | 64 | 1849 | $[1, 646]$ | Gene sets | Binary | 08/2024 |
| Cancer | 7057 | 60 | 1277 | $[1, 292]$ | Gene sets | Binary | 08/2024 |
| Celiac | 14294 | 132 | 1666 | $[1, 570]$ | Gene sets | Binary | 08/2024 |
| Colitis | 11999 | 127 | 1528 | $[1, 497]$ | Gene sets | Binary | 08/2024 |
| Tumour | 17661 | 52 | 1849 | $[1, 646]$ | Gene sets | Binary | 08/2024 |

Table 4.3: Dataset information for the seven datasets used in the real data analysis.

### 4.7.2 Model Hyperparameters

Both gSLOPE and SGS were fit with their FDR control parameters set to 0.01, and for SGS $\alpha = 0.99$. Each model was applied along a log-linear path of 100 regularisation parameters, with $\lambda_{100} = 0.01\lambda_1$, where $\lambda_1 = \lambda_{\max}$.

Detailed model and algorithmic hyperparameter information is provided in Tables A.4 and A.6.

### 4.7.3 Results

For both gSLOPE and SGS, the screening rules led to considerably faster runtimes across all datasets (Figure 4.7). The screening was more effective for SGS under a linear model and for gSLOPE under a logistic model. Under the synthetic data, the screening for SGS was also more effective for linear models, whereas for gSLOPE, there was little distinction between linear and logistic models (Table 4.2). For the real data, the active set sizes for gSLOPE tended to be smaller for the logistic models, allowing for larger feature reduction (Table 4.1). For SGS, this trend went the other way; the active sets were smaller for the continuous datasets.

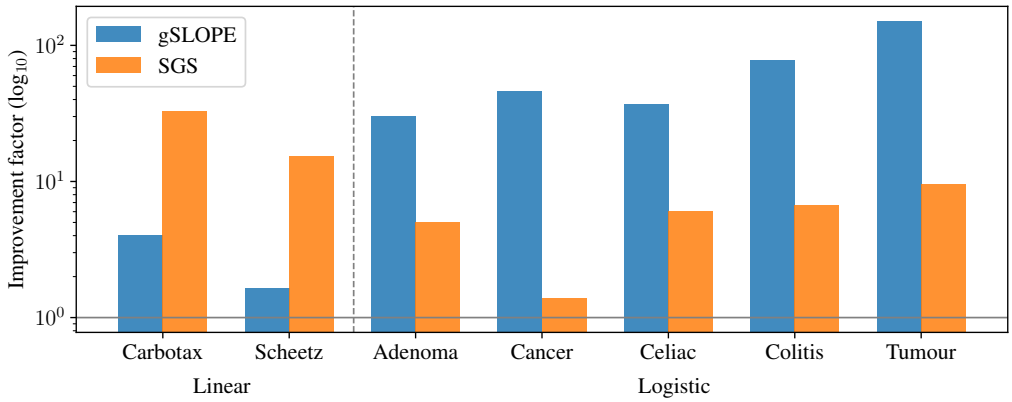



Figure 4.7: The improvement factor ($\log_{10}$ scale) of gSLOPE and SGS applied to the real datasets, for fitting 100 path models, split into response type. The horizontal grey line represents no screening improvement.

An additional reason for the reduced screening efficiency for gSLOPE under linear models was the grouping structure. For the continuous responses, the groups were generated using K-means clustering, leading to fewer groups and less opportunity for dimensionality reduction (Table 4.3). With fewer groups, screening for gSLOPE cannot reduce dimensionality as effectively, since the small number of

groups are more likely to each contain some signal.

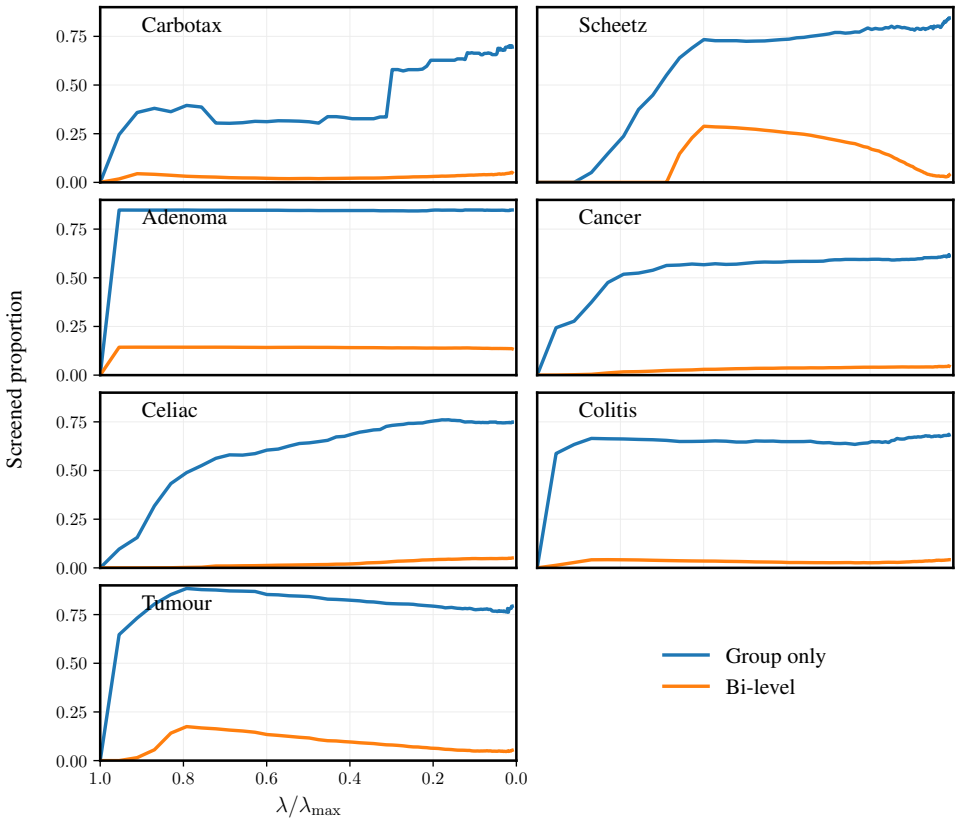



Figure 4.8: The screened proportion for group-only and bi-level screening applied to SGS, plotted along the regularisation path for all datasets.

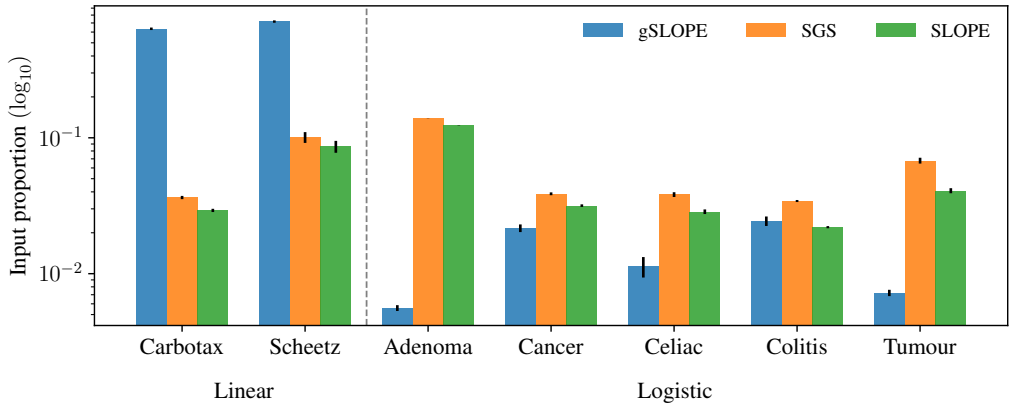



Figure 4.9: The input proportion ($\log_{10}$ scale) of SLOPE, gSLOPE, and SGS applied to the real datasets, for fitting 100 path models, split into response type.

The feature reduction provided by the screening rules led to the alleviation of convergence issues (Table C.4). SGS failed to converge for three datasets (Carbotax, Scheetz, and Adenoma) without screening and none with screening. gSLOPE experienced failed convergences across all datasets without screening. With screening, gSLOPE only failed to converge for three datasets, each showing fewer instances of failure. As gSLOPE applies no variable penalisation, it is forced to activate all variables within an active group. For datasets with large groups, such as those considered here, this leads to a problematic fitting process which can include many noise variables. Our screening rules help gSLOPE partially overcome this issue, providing large computational savings and better solution optimality.

The analysis of the real data further illustrates the benefits of the bi-level screening to the runtime and performance of SGS. Figure 4.8 shows that for all real datasets, the bi-level screening allows the input dimensionality for SGS to be reduced to a much greater extent than by just group screening.

Both screening rules drastically reduce the input dimensionality on all the real datasets (Table 4.1). On average, for the logistic real data models, the SGS screening rules reduced the input to just 7% of the total space. This reduction comes without affecting solution optimality (Appendices C.5.2 and C.5.4). Our rules for gSLOPE and SGS are found to offer similar levels of feature reduction to that of the SLOPE strong rule (Figure 4.9). For additional metrics and comparison to the SLOPE strong rule for the real datasets, see Figures C.12 and C.13.

The screening rules also showed significant improvements to runtime (Figure C.7) and reduction of input proportion (Figure C.6) for group-based OSCAR models applied to the real datasets.

### 4.7.4 KKT Violations

Across all real datasets, gSLOPE only had a KKT violation when applied to Carbotax, and the number of violations was very small (an average proportion of 0.0005 violating groups per path model; see Table C.4). In contrast, SGS had an average proportion of 0.0003 violating variables per path model, showing a pathwise trend similar to that in the synthetic data but with a lower violating proportion (Figure 4.10).

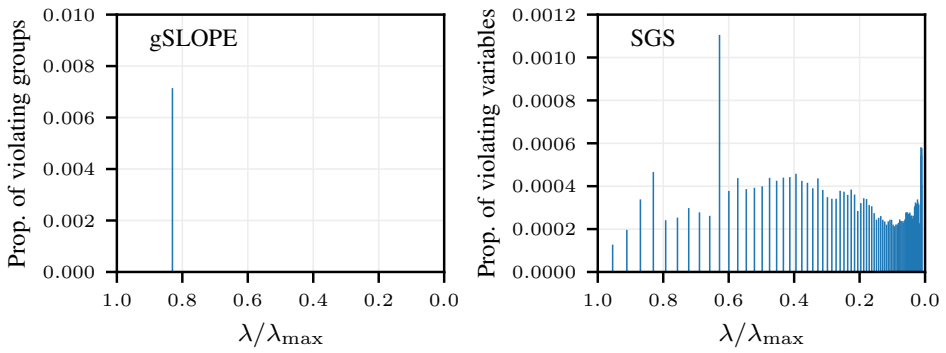



Figure 4.10: The proportion of KKT violations relative to the full input for gSLOPE and SGS averaged over all real datasets. This is shown as a function of the regularisation path.

## 4.8 Discussion

In this chapter, strong screening rules were proposed for group-based SLOPE models using a new sparse-group strong screening framework. The rules were developed for Group SLOPE and Sparse-group SLOPE, neither of which have any previous screening rules. Our proposed screening rules differ from the existing SLOPE strong rule both in construction and in outcome. The screening rule for gSLOPE screens out irrelevant groups before fitting by transforming the gSLOPE subdifferential into the SLOPE one. The screening rules for SGS perform bi-level screening, removing inactive groups and inactive variables in active groups. This is achieved by using the soft thresholding operator to specify the unknown subdifferential. Our rules apply to the wider class of group-based OWL models.

SLOPE models are finding increasing use in genetics and machine learning, with SGS found to have superior disease prediction performance over other penalised methods in Chapter 3. The screening rules developed in this chapter make the group-based versions more accessible by reducing their computational burden. These rules enable gSLOPE and SGS to achieve fitting times comparable to those of their lasso-based counterparts. This will allow practitioners to utilise their FDR properties, facilitating their widespread use across various fields. The screening rules also helped gSLOPE and SGS overcome convergence issues in large datasets, improving solution optimality.

Through a comprehensive analysis of synthetic and real data, we illustrate that

the screening rules lead to dramatic improvements in the runtime and reduction of input dimensionality of gSLOPE and SGS models. This is particularly important in datasets where $p \gg n$, such as genetics ones, which are covered extensively in this thesis and are the main motivation behind the proposal of SLOPE. Throughout the data studies, no scenario was discovered where our screening rules did not perform better than no screening.

The sparse-group framework developed in this chapter is extended to the sparse-group lasso in Chapter 5, showing the wide utility of the framework.

# CHAPTER 5

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

*The content of this chapter is based largely on our workshop paper Feser and Evangelou [67], published at the 16th Annual Workshop on Optimization for Machine Learning (OPT) at NeurIPS (2024), and full paper Feser and Evangelou [68], published in the Proceedings of the 42nd International Conference on Machine Learning (ICML) (2025). The screening rules developed in this chapter are implemented in the* `dfr` *R package [65], available on CRAN.*

This chapter introduces a novel feature reduction method for SGL, Dual Feature Reduction (DFR), that uses the sparse-group screening framework of Chapter 4 to develop strong screening rules for the sparse-group lasso and the adaptive sparse-group lasso. SGL, like SGS, performs both variable and group selection, simultaneously using the strengths of the lasso and group lasso. It is widely used in genetics, a field that regularly involves the analysis of high-dimensional data, due to its sparse-group penalty, which allows it to utilise grouping information. SGL can also be computationally expensive, due to the added shrinkage complexity from its sparse-group norm, and its additional hyperparameter that needs tuning. As such, the motivations behind developing screening approaches for SGS are also relevant for SGL.

The proposed DFR rules reduce the input space before optimisation, without affecting solution optimality, and are the first strong screening rules for SGL and its adaptive variant. DFR applies two layers of screening via the application of dual norms and subdifferentials. Through synthetic and real data studies, it is shown that DFR drastically reduces the computational cost under many different

scenarios, clearly outperforming existing methods.

## 5.1 Introduction

Section 4.1 provides an overview of the screening literature and the relevant background; in this section, we introduce the specific problem of screening for SGL.

As observed in this thesis, applying only group shrinkage can harm convergence and prediction, as all variables in an active group are retained, including noise variables (Chapters 3 and 4, see also Simon et al. [178]). This limitation led to the development of sparse-group models, such as the SGS model introduced in Chapter 3. SGS was in part motivated by the proposal of SGL. SGL can shrink variables in active groups by applying shrinkage on both variables and groups to yield bi-level selection. It has found increased popularity in applications in the machine learning [196, 223] and healthcare [62, 149, 178] communities. SGL has been shown to have tangible benefits by consistently outperforming the lasso and group lasso in selection and prediction tasks [178].

The SGL norm, introduced in Section 2.3.1, is defined as a convex combination of the lasso and group lasso penalties, as shown in Equation 2.6. It has been extended to have adaptive shrinkage through the adaptive sparse-group lasso (aSGL) (introduced in Section 5.3.1). By applying adaptivity to the weights, aSGL has been shown to attain the oracle property in a double-asymptotic framework [151] and has been applied in areas such as portfolio selection in finance [167] and quantile regression [132]. This underscores the need for screening rules for aSGL, so that its strengths can be fully utilised across different domains.

The objective of screening for SGL is as it was for SGS in Chapter 4: use the solution at $\lambda_k$ to generate a set of screened variables at path point $\lambda_{k+1}$, given by $\mathcal{S}_v(\lambda_{k+1})$, that is a superset of the active set $\mathcal{A}_v(\lambda_{k+1})$. The SGL optimisation at $\lambda_{k+1}$ (Equation 2.6) is then performed using only the remaining variables.

### 5.1.1 Screening Approaches for the Sparse-group Lasso

The strengths of SGL come with increased computational cost, due to the additional shrinkage and the tuning of two hyperparameters. Typically, $\alpha$ is set subjectively, and $\lambda$ is tuned using CV along a path. Feature reduction techniques, including screening rules, can help ease the cost of fitting an SGL model along a path by

discarding features before optimisation that would have been inactive at the optimal solution.

An exact screening rule for SGL, called GAP safe (Section 5.1.2), was proposed by Ndiaye et al. [137] using the SAFE framework. GAP safe uses the duality gap to create feasible regions of active features and applies reduction on both the groups and variables. Other reduction methods for SGL include Two-Layer Feature Reduction (TLFre), proposed as an exact method [206], although it was later shown not to be exact [136], and the heuristic sparsegl rule (Section 5.1.2) [121], which applies only group-level reduction. Additional attempts at improving the SGL runtime include using approximate bounds for inactive conditions [95] and a heuristic screening rule limited to multi-response Cox modelling [120].

### 5.1.2 Details of Competitive Approaches

In Sections 5.4 and 5.5, DFR is compared to GAP safe and sparsegl in computational studies. Here, we provide detailed descriptions of both approaches.

#### sparsegl

sparsegl, proposed by Liang et al. [121], is a strong screening rule for SGL that performs only a single layer of group screening. As a result, it does not fully exploit the sparse-group structure of the SGL penalty, unlike DFR, which reduces dimensionality further by using both group- and variable-level screening. The rule is similar to DFR in that it is based on the strong screening framework and applies a Lipschitz assumption.

sparsegl is motivated by the first order condition derived in Simon et al. [178] for SGL, *i.e.*, that a group $j \in [m]$ is inactive if

$$\|S(\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})), \lambda_{k+1}\alpha)\|_2 \leq \sqrt{p_j}(1-\alpha)\lambda_{k+1}.$$

As the gradient at $k+1$ is not available, the following Lipschitz assumption on the $\ell_2$ norm is used:

$$\|S(\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})), \lambda_{k+1}\alpha) - S(\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k)), \lambda_k\alpha)\|_2 \leq \sqrt{p_j}(1-\alpha)|\lambda_{k+1} - \lambda_k|.$$

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

group $j \in [m]$ if

$$\|S(\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k)), \lambda_k \alpha)\|_2 \leq \sqrt{p_j}(1-\alpha)(2\lambda_{k+1} - \lambda_k).$$

This screening rule uses a different Lipschitz assumption at the group level (DFR: Equation 5.9), which in turn leads to a different group-level rule (DFR: Equation 5.10). The DFR Lipschitz assumption aligns more closely with the strong screening framework, as it pertains to the dual norm of the full SGL norm, rather than just the group component.

### GAP safe

GAP Safe is an exact screening method for SGL, proposed by Ndiaye et al. [137] under the linear regression setting, and is based on the SAFE framework. The approach makes use of the subdifferential inclusion equation of Fermat's rule (Definition 4.7). In GAP safe, the safe region is constructed as a sphere, although other regions can also be used (such as domes).

The safe sphere is defined as $\mathcal{B}(\Theta_c, r)$ with centre $\Theta_c$ and radius $r \geq 0$. An ideal region is one where $r$ is small and its centre is close to the dual solution. Using this safe region, the GAP safe rules at $\lambda_{k+1}$ are derived as, for a variable $i \in [p]$ and group $j \in [m]$,

$$\begin{aligned} \text{Variable screening:} \quad & |\mathbf{X}_i^\top \Theta_c| + r\|\mathbf{X}_i\|_2 < \alpha \implies \hat{\beta}_i(\lambda_{k+1}) = 0. \\ \text{Group screening:} \quad & \mathcal{T}_j < (1-\alpha)\sqrt{p_j} \implies \hat{\boldsymbol{\beta}}^{(j)}(\lambda_{k+1}) \equiv \mathbf{0}, \end{aligned}$$

where

$$\mathcal{T}_j = \begin{cases} \|S(\mathbf{X}^{(j)^\top}\Theta_c, \alpha)\|_2 + r\|\mathbf{X}^{(j)}\|_2 & \text{if } \|\mathbf{X}^{(j)^\top}\Theta_c\|_\infty > \alpha, \\ (\|\mathbf{X}^{(j)^\top}\Theta_c\|_\infty + r\|\mathbf{X}^{(j)}\|_2 - \alpha)_+ & \text{otherwise.} \end{cases}$$

The centre $\Theta_c$ and the radius $r$ are derived using the duality gap and are calculated at iteration $t$ in an iterative algorithm as

$$\Theta_t(\hat{\boldsymbol{\beta}}_{[t]}) = \frac{\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{[t]}}{\max(\lambda_{k+1}, J^*_{\text{sgl}}(\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{[t]})))},$$

$$r_t(\hat{\boldsymbol{\beta}}_{[t]}, \Theta_t) = \sqrt{\frac{2P_{\lambda_{k+1},\alpha}(\hat{\boldsymbol{\beta}}_{[t]}) - D_{\lambda_{k+1}}(\hat{\boldsymbol{\theta}}_{[t]})}{\lambda_{k+1}^2}},$$

where $P_{\lambda,\alpha}$ and $D_\lambda$ are the primal (Definition 2.6) and dual objectives (Definition 4.6), and $\hat{\boldsymbol{\beta}}_{[t]}$ and $\hat{\boldsymbol{\theta}}_{[t]}$ are the primal and dual fitted values at iteration $t$. The radius and centre are expensive to evaluate, so they are calculated only every 10 iterations.

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 centre and radius are from $\lambda_k$. GAP safe has various alterations, but in this chapter, we present the two that provided the best results in our studies (dynamic and sequential).

### 5.1.3 Contributions and Organisation of this Chapter

In this chapter, we propose a new screening approach for SGL and adaptive SGL, Dual Feature Reduction (DFR), which extends the sparse-group screening framework introduced in Chapter 4. DFR introduces the first bi-level strong screening rules for SGL (Section 5.2) and the first screening rules for aSGL (Section 5.3).

While DFR uses the framework proposed in Chapter 4, the rules themselves are derived from scratch. The rules in Chapter 4 were derived for SGS, which has a non-separable norm that is designed for FDR control. On the other hand, SGL and aSGL are separable norms. Thus, in contrast to the group-based SLOPE rules in Chapter 4, which were derived using the subdifferential directly, the rules in this chapter are derived using the dual norm, making them more closely aligned to the strong framework. Despite these differences, a connection between SGL and SGS was identified in Chapter 4 (Remark 4.1) through the soft-thresholding operator. This connection enabled the derivation of screening rules for SGS, and in this chapter, the same operator is employed to derive the KKT checks for SGL (Section 5.2.3).

For aSGL, we prove a valuable new connection between its dual norm and the $\epsilon$-norm (Proposition 5.5), sharpened further by additional theoretical results, which enabled the adaptation of screening rules from SGL to aSGL.

DFR applies two layers of screening, eliminating inactive groups first, then filtering out inactive variables within the remaining active groups. By reducing the input dimensionality before optimisation, DFR enables expanded tuning regimes to be performed, including simultaneous tuning of $\lambda$ and $\alpha$. These benefits are achieved without affecting solution optimality. The resulting computational efficiency increases the accessibility of SGL and aSGL models, encouraging wider adoption across fields.

DFR-SGL has key benefits over other competitive approaches. The GAP safe rules for SGL require computation of safe regions, which include a radius and centre,

and the dual norm, as well as iterative screening and fitting. In contrast, DFR needs only the dual norm and screens just once at each path point, making it considerably less expensive. On the other hand, sparsegl performs only group-level screening, so it is not able to reduce the input dimensionality to the same extent as DFR.

These claims are supported by our results: DFR is applied to synthetic and real data in Sections 5.4 and 5.5, where it is found to be the state-of-the-art screening approach for SGL, outperforming other existing methods in terms of computational cost and dimensionality reduction.

## 5.2 Dual Feature Reduction for SGL

To derive the *Dual Feature Reduction* (DFR) approach, the dual norm of SGL is used (Definition 5.1), which is formulated in terms of the $\epsilon$-norm (Definition 5.2).

**Definition 5.1** (Dual norm of SGL)**.** The dual norm of SGL is defined as

$$J^*_{\text{sgl}}(\mathbf{x}) := \sup \left\{ \mathbf{x}^\top \mathbf{z} : J_{\text{sgl}}(\mathbf{z}) \leq 1 \right\}.$$

The SGL norm can be expressed in terms of the dual of the $\epsilon$-norm (Definition 5.2) [137],

$$J_{\text{sgl}}(\boldsymbol{\beta}) = \sum_{j=1}^{m} \tau_j \|\boldsymbol{\beta}^{(j)}\|^*_{\epsilon_j}, \tag{5.1}$$

where $\tau_j = \alpha + (1-\alpha)\sqrt{p_j}$. Using this, the dual norm of SGL applied to a group $j \in [m]$ is formulated as

$$J^*_{\text{sgl}}(\mathbf{x}^{(j)}) = \max_{j=1,\dots,m} \tau_j^{-1} \|\mathbf{x}^{(j)}\|_{\epsilon_j}. \tag{5.2}$$

**Definition 5.2** ($\epsilon$-norm [35])**.** The $\epsilon$-norm, $\|x\|_{\epsilon_j}$, applied to a group $j \in [m]$, is defined as the unique non-negative solution $\xi$ of the equation

$$\sum_{i=1}^{p_j} (|x_i| - (1-\epsilon_j)\xi)_+^2 = (\epsilon_j \xi)^2, \quad \epsilon_j = \frac{\tau_j - \alpha}{\tau_j}. \tag{5.3}$$

### 5.2.1 Group Reduction

To generate a screened group set, the KKT stationarity conditions (Definition 4.1) are used, providing conditions for an inactive group. For SGL, they are given by,

for a group $j \in [m]$ at $\lambda_{k+1}$ (using the $\epsilon$-norm representation in Equation 5.1),

$$\mathbf{0} \in \nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) + \tau_j \lambda_{k+1} \Theta_{j,k+1}, \tag{5.4}$$

where $\Theta_{j,k+1} = \partial \|\hat{\boldsymbol{\beta}}(\lambda_{k+1})\|^*_{\epsilon_j}$ is the subgradient of the dual norm of the $\epsilon$-norm. The subgradient for an inactive group can be expressed by the unit ball of the dual norm [171]:

$$\Theta^0_{j,k+1} := \partial \|0\|^*_{\epsilon_j} = \left\{ \mathbf{x} \in \mathbb{R}^{p_j} : \|x\|_{\epsilon_j} \leq 1 \right\}. \tag{5.5}$$

Plugging the unit ball into Equation 5.4 and applying the $\epsilon$-norm, the subgradient can be cancelled out, giving the condition

$$\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} = \tau_g \lambda_{k+1} \|\Theta^0_{j,k+1}\|_{\epsilon_j} \leq \tau_j \lambda_{k+1}. \tag{5.6}$$

If the gradient at $\lambda_{k+1}$ were available, we could exactly identify the active groups using Equation 5.6; if the equation is satisfied, the group is inactive. This is formalised in Proposition 5.1.

**Proposition 5.1** (Theoretical SGL group screening)**.** *For SGL applied with any $\lambda_{k+1}, k \in [l-1]$, the screened group set,*

$$\mathcal{S}_g(\lambda_{k+1}) = \left\{ j \in [m] : \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} > \tau_j \lambda_{k+1} \right\}, \tag{5.7}$$

*is such that $\mathcal{S}_g(\lambda_{k+1}) = \mathcal{A}_g(\lambda_{k+1})$.*

*Proof of Proposition 5.1.* For $k \in [l-1]$, to prove the two sets are equivalent, we need to prove that for any $j \in [m]$, $j \in \mathcal{A}_g(\lambda_{k+1}) \iff j \in \mathcal{S}_g(\lambda_{k+1})$. We instead prove the contrapositive: $j \notin \mathcal{S}_g(\lambda_{k+1}) \iff j \notin \mathcal{A}_g(\lambda_{k+1})$. So,

$$\begin{aligned}
j \notin \mathcal{S}_g(\lambda_{k+1}) &\iff \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} \leq \tau_j \lambda_{k+1}, && \text{by Equation 5.7} \\
&\iff \nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) \in \tau_j \lambda_{k+1} \Theta^0_{j,k+1}, && \text{by Equation 5.5} \\
&\iff \mathbf{0} \in \nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) + \tau_j \lambda_{k+1} \Theta^0_{j,k+1} \\
&\iff j \notin \mathcal{A}_g(\lambda_{k+1}), && \text{by Equation 5.4.}
\end{aligned}$$

□

### Gradient Approximation

In practice, the gradient term is unavailable, so applying Proposition 5.1 is not possible. In a similar vein to the screening rules in Chapter 4, an approximation $R_j$ is required such that we upper bound the gradient term in Equation 5.6:

$$\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} \leq R_j. \tag{5.8}$$

Then, the screening rule tests whether $R_j \leq \tau_j \lambda_{k+1}$. If this is found to be true, it can be concluded that Equation 5.6 holds and the group must be inactive.

*Remark* 5.1. By establishing an upper bound $R_j$, it is guaranteed that no active groups are inadvertently discarded. Suppose we have an active group $j \in \mathcal{A}_g(\lambda_{k+1})$, then by Equation 5.6

$$\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} > \tau_j \lambda_{k+1} \implies R_j > \tau_j \lambda_{k+1},$$

and the group cannot be discarded via the screening rule. The downside is that fewer inactive groups are screened out; however, this is preferable to incorrectly discarding active groups. To see this, suppose that we have a situation where $R_j > \tau_j \lambda_{k+1}$ but $\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} \leq \tau_j \lambda_{k+1}$. By Equation 5.6, this group should be discarded, but by the approximation, it is added to the screened set.

An approximation $R_j$ can be found by assuming that the gradient is a Lipschitz function of $\lambda_{k+1}$ with respect to the $\epsilon$-norm,

$$\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) - \nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j} \leq \tau_j |\lambda_{k+1} - \lambda_k|, \tag{5.9}$$

which is a similar assumption to that used in the lasso strong rule [193] and for the rules developed in Chapter 4. Using the reverse triangle inequality gives a suitable approximation:

$$\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} \leq \underbrace{\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j} + \tau_j(\lambda_k - \lambda_{k+1})}_{=:R_j}.$$

Therefore, the strong group screening rule for SGL (Proposition 5.2) can be formulated by plugging $R_j$ into Equation 5.8: discard a group $j \in [m]$ if

$$\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j} \leq \tau_j(2\lambda_{k+1} - \lambda_k). \tag{5.10}$$

Since the Lipschitz assumption can fail, KKT checks are performed to prevent violations (Section 5.2.3).

**Proposition 5.2** (DFR-SGL group screening)**.** *For SGL applied with any* $\lambda_{k+1}, k \in [l-1]$*, assuming that*

$$\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) - \nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j} \leq \tau_j |\lambda_{k+1} - \lambda_k|, \ \forall j \in [m], \tag{5.11}$$

*then the screened group set,*

$$\mathcal{S}_g(\lambda_{k+1}) = \left\{ j \in [m] : \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j} > \tau_j(2\lambda_{k+1} - \lambda_k) \right\},$$

*is such that* $\mathcal{A}_g(\lambda_{k+1}) \subset \mathcal{S}_g(\lambda_{k+1})$.

*Proof of Proposition 5.2.* For $k \in [l-1]$, to prove the screened set is a superset of the active set, we need to prove that for any $j \in [m]$, $j \in \mathcal{A}_g(\lambda_{k+1}) \implies j \in \mathcal{S}_g(\lambda_{k+1})$. Similar to the proof of Proposition 5.1 (although the implication here is only in one direction), we instead prove the contrapositive: $j \notin \mathcal{S}_g(\lambda_{k+1}) \implies j \notin \mathcal{A}_g(\lambda_{k+1})$. First, we rewrite the Lipschitz assumption (Equation 5.11) as

$$\begin{aligned} \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} - \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j} &\leq \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) - \nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j} \\ &\leq \tau_j |\lambda_{k+1} - \lambda_k| \\ \implies \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} &\leq \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j} + \tau_j |\lambda_{k+1} - \lambda_k|. \end{aligned} \tag{5.12}$$

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

$$\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j} \leq \tau_j(2\lambda_{k+1} - \lambda_k).$$

Plugging this into Equation 5.12 yields

$$\begin{aligned} & \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} \leq \tau_j(2\lambda_{k+1} - \lambda_k) + \tau_j|\lambda_{k+1} - \lambda_k| \\ \implies & \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j} \leq \tau_j \lambda_{k+1} \\ \implies & \nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) \in \tau_j \lambda_{k+1} \Theta^0_{j,k+1}, && \text{by Equation 5.5} \\ \implies & \mathbf{0} \in \nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) + \tau_j \lambda_{k+1} \Theta^0_{j,k+1} \\ \implies & j \notin \mathcal{A}_g(\lambda_{k+1}), && \text{by Equation 5.4} \end{aligned}$$

□

### 5.2.2 Variable Reduction

Group screening reduces the input dimensionality, but using the bi-level nature of the SGL norm, further reduction is possible by applying a second screening layer to the variables in the screened groups. For an inactive variable in a group, $i \notin \mathcal{A}_v(\lambda_{k+1}), i \in G_j$, the KKT conditions are

$$\mathbf{0} \in \nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) + \lambda_{k+1}\alpha\Phi^0_{i,k+1} + \lambda_{k+1}(1-\alpha)\Psi^{(j)}_{i,k+1}, \tag{5.13}$$

where $\Phi$ and $\Psi$ are the subgradients of the $\ell_1$ and $\ell_2$ norms, respectively. For an active group $j \in [m]$, the subgradient of the $\ell_2$ norm is given by $\hat{\beta}^{(j)}_i / \|\hat{\boldsymbol{\beta}}^{(j)}\|_2$, which is zero for an inactive variable. Therefore,

$$\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) \in \lambda_{k+1}\alpha\Phi^0_{i,k+1} \iff |\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))| \leq \lambda_{k+1}\alpha, \tag{5.14}$$

$$\text{where } \Phi^0_{i,k+1} = \{x \in \mathbb{R} : |x| \leq 1\}. \tag{5.15}$$

As with the SGL group screening, knowledge of the gradient would lead to exact support recovery (Proposition 5.3). Equation 5.14 is similar to the strong screening rule for the lasso, scaled by $\alpha$. Hence, using a scaled version of the Lipschitz assumption used for the lasso strong rule, the variable screening rule for SGL is formulated in Proposition 5.4.

**Proposition 5.3** (Theoretical SGL variable screening)**.** *For SGL applied with any $\lambda_{k+1}, k \in [l-1]$, the screened variable set,*

$$\mathcal{S}_v(\lambda_{k+1}) = \left\{ i \in \mathcal{G}_{\mathcal{A}_g(\lambda_{k+1})} : |\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))| > \lambda_{k+1}\alpha \right\},$$

*is such that $\mathcal{S}_v(\lambda_{k+1}) = \mathcal{A}_v(\lambda_{k+1})$.*

**Proposition 5.4** (DFR-SGL variable screening)**.** *For SGL applied with any $\lambda_{k+1}, k \in [l-1]$, assuming that*

$$|\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) - \nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_k))| \leq \alpha(\lambda_k - \lambda_{k+1}),\ \forall i \in \mathcal{G}_{\mathcal{A}_g(\lambda_{k+1})},$$

*then the screened variable set,*

$$\mathcal{S}_v(\lambda_{k+1}) = \left\{ i \in \mathcal{G}_{\mathcal{A}_g(\lambda_{k+1})} : |\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_k))| > \alpha(2\lambda_{k+1} - \lambda_k) \right\},$$

*is such that $\mathcal{A}_v(\lambda_{k+1}) \subset \mathcal{S}_v(\lambda_{k+1})$.*

*Remark* 5.2. To apply Proposition 5.4 in practice, knowledge of $\mathcal{A}_g(\lambda_{k+1})$ is required, but this is unknown. By Proposition 5.2, $\mathcal{A}_g(\lambda_{k+1}) \subset \mathcal{S}_g(\lambda_{k+1})$, and so the screened set is used in practice. A similar situation was also resolved in Chapter 4 (see Remark 4.2).

### 5.2.3 KKT Checks

The screening rules of DFR use several Lipschitz assumptions (Propositions 5.2 and 5.4), as well as replacing the active group set by the screened group set for the variable screening step (Remark 5.2). The Lipschitz assumptions are standard for strong rules, but the replacement is unique to sparse-group screening. Violations of these assumptions can lead to DFR incorrectly discarding active variables. To ensure that no active variables are discarded, the KKT conditions are checked for each variable after screening (Lemma 5.1). Any variable violating the KKT conditions is added back into the optimisation procedure.

**Lemma 5.1** (SGL KKT violation)**.** *An SGL KKT violation occurs for a variable* $i \in G_j$ *if* $i \in \mathcal{A}_v(\lambda_{k+1}), i \notin \mathcal{S}_v(\lambda_{k+1})$*, and the following condition is met*

$$|S(\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})), \lambda_{k+1}(1-\alpha)\sqrt{p_j})| > \lambda_{k+1}\alpha. \tag{5.16}$$

*Proof.* To determine whether a variable $i \in G_j$ has been correctly discarded, the KKT stationarity conditions are checked. Equation 5.13 describes the condition under which a variable $i \in G_j$ is inactive. Without specifying whether the group $j$ is inactive, this can be expressed as (by Equation 5.15)

$$|\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) + \lambda_{k+1}(1-\alpha)\Psi^{(j)}_{i,k+1}| \leq \lambda_{k+1}\alpha, \tag{5.17}$$

where $\Psi^{(j)}_{k+1} = \{\mathbf{x} \in \mathbb{R}^{\sqrt{p_j}} : \|\mathbf{x}\|_2 \leq 1\}$. To satisfy Equation 5.17, the unknown subdifferential, $\Psi^{(j)}_{i,k+1}$, is taken to be the minimum possible value. For $\mathbf{x} \in \Psi^{(j)}_{k+1}$, we have that

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

Hence, the values in the subdifferential are bounded by $\sqrt{p_j}$. We consider the

following scenarios for Equation 5.17:

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

This allows Equation 5.17 to be expressed using the soft-thresholding operator as

$$|S(\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})), \lambda_{k+1}(1-\alpha)\sqrt{p_j})| \leq \lambda_{k+1}\alpha. \tag{5.18}$$

Hence, a KKT violation occurs when Equation 5.18 does not hold. □

*Remark* 5.3. A related derivation is provided in Simon et al. [178] for establishing conditions to determine whether a group is active in SGL, and in Chapter 4 (Lemma 4.1) for deriving the group screening rule in SGS

### 5.2.4 Path Start

To apply DFR, SGL is fit along a path of values, $\lambda_1 \geq \cdots \geq \lambda_l \geq 0$. The path start, $\lambda_1$, is often chosen to be the exact point at which the first predictor becomes active, $\lambda_{\max}$. By Ndiaye et al. [138] and using the dual norm from Equation 5.2, this value is given by

$$\lambda_{\max} = J^*_{\text{sgl}}(\nabla f(\mathbf{0})) = \max_{j=1,\dots,m} \tau_j^{-1} \|\nabla_j f(\mathbf{0})\|_{\epsilon_j}.$$

### 5.2.5 Algorithm

The DFR algorithm is based on the sparse-group strong screening framework, described in Chapter 4 (Section 4.3). The algorithm has the following key steps:

1. *Group screening*: find $\mathcal{S}_g(\lambda_{k+1})$ using Proposition 5.2.
2. *Variable screening*: find $\mathcal{S}_v(\lambda_{k+1})$ using Proposition 5.4 for $i \in \mathcal{G}_{\mathcal{S}_g(\lambda_{k+1})}$.
3. *Optimisation*: Compute $\hat{\beta}_i(\lambda_{k+1})$ for $i \in \mathcal{E}_v$ using the *optimisation set* $\mathcal{E}_v = \mathcal{S}_v(\lambda_{k+1}) \cup \mathcal{A}_v(\lambda_k)$. Perform KKT checks to identify any violations and add offending variables into $\mathcal{E}_v$ (Section 5.2.3). Repeat this step until no violations.

The algorithm's two primary computational costs are computing the solution, $\hat{\boldsymbol{\beta}}$, and evaluating the $\epsilon$-norm. The cost of computing the solution depends on the fitting algorithm used (as this framework is applicable for any SGL fitting algorithm). Proximal and descent algorithms are popular choices and typically have complexities of $\mathcal{O}(tp^2)$, for $t$ iterations [237]. The latter has a worst-case cost of $\mathcal{O}(p_j \log p_j)$ for each group $j \in [m]$ [137]. In Section 5.4, we implement DFR using both proximal and descent algorithms, and compare their performance.

Tables 5.1 and 5.2 show a runtime breakdown of DFR components for two cases using a proximal algorithm (ATOS), indicating that screening adds minimal overhead while significantly improving fitting efficiency. By comparison, in Figure 5.2 without screening, the average fitting time was 544s for SGL and 603s for aSGL.

| Component | DFR-SGL | DFR-aSGL |
|---|---|---|
| Fitting algorithm | 88% | 86% |
| $\epsilon$-norm evaluation | 3.9% | 3.6% |
| Group screening | 3.9% | 3.6% |
| Variable screening | 0.01% | 0.01% |
| KKT checks | 0.6% | 0.6% |

Table 5.1: Figure 5.2 runtime.

| Component | DFR-SGL | DFR-aSGL |
|---|---|---|
| Fitting algorithm | 77% | 65% |
| $\epsilon$-norm evaluation | 0.46% | 0.57% |
| Group screening | 0.46% | 0.58% |
| Variable screening | 0.01% | 0.02% |
| KKT checks | 0.2% | 0.3% |

Table 5.2: Scheetz runtime (Figure 5.9).

## 5.3 Dual Feature Reduction for aSGL

### 5.3.1 Adaptive Sparse-group Lasso

Using adaptive weights $(\mathbf{v}, \mathbf{w})$, the *Adaptive Sparse-group Lasso* (aSGL) applies adaptive shrinkage in a sparse-group setting with the norm [132]

$$J_{\text{asgl}}(\boldsymbol{\beta}; \mathbf{v}, \mathbf{w}) = \alpha \sum_{i=1}^{p} v_i |\beta_i| + (1-\alpha) \sum_{j=1}^{m} w_j \sqrt{p_j} \|\boldsymbol{\beta}^{(j)}\|_2. \tag{5.19}$$

There are many choices of weights, but in this chapter, the weights are chosen according to Mendez-Civieta et al. [132] as, for a variable $i \in [p]$ and group $j \in [m]$,

$$v_i = \frac{1}{|q_{1i}|^{b_1}}, \; w_j = \frac{1}{\|\mathbf{q_1}^{(j)}\|_2^{b_2}},$$

where $q_1$ is the first principal component from performing principal component analysis on $\mathbf{X}$ and $b_1, b_2$ are manually specified, often in the range $[0, 2]$. The weights

for $b_1 = b_2 = 0.1$ are shown in Figure 5.1.

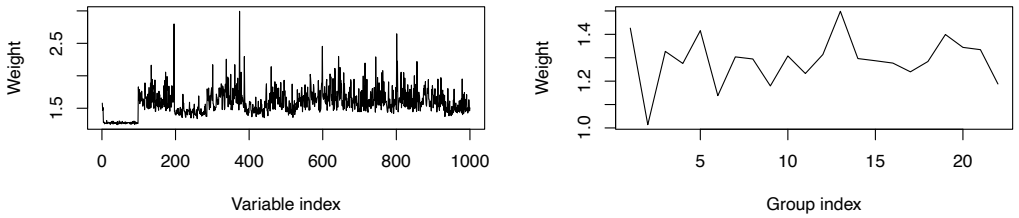



Figure 5.1: The weights $(\mathbf{v}, \mathbf{w})$ for aSGL, used in Figure 5.5 (right), where $p = 1000, n = 200, m = 22, \rho = 0.3, b_1 = b_2 = 0.1$, and $\alpha = 0.95$.

### 5.3.2 Screening for aSGL

DFR is extended to aSGL via the $\epsilon$-norm: Proposition 5.5 expresses the aSGL norm as the $\epsilon$-norm, linking it to SGL. Therefore, the DFR-SGL rules apply for aSGL by replacing $\tau_j$ with $\tau_j'$ and $\epsilon_j$ with $\epsilon_j'$ (see Appendices D.2.1 and D.2.2).

**Proposition 5.5.** *The aSGL norm can be expressed using the $\epsilon$-norm by*

$$
\begin{aligned}
&J_{asgl}(\boldsymbol{\beta}; \mathbf{v}, \mathbf{w}) = \sum_{j=1}^{m} \tau_j' \|\boldsymbol{\beta}^{(j)}\|_{\epsilon_j'}^{*}, \text{ where,} \\
&\tau_j' = \alpha \|\mathbf{v}^{(j)}\|_1 - \frac{\alpha}{\|\boldsymbol{\beta}^{(j)}\|_1} \sum_{i,j \in G_j, i \neq j} v_j |\beta_i| + (1-\alpha) w_j \sqrt{p_j}, \\
&\epsilon_j' = (\tau_j')^{-1} (1-\alpha) w_j \sqrt{p_j}.
\end{aligned} \tag{5.20}
$$

*Proof of Proposition 5.5.* Splitting up the variable norm gives

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

This allows the aSGL norm to be written in terms of the groups

$$\sum_{j=1}^{m}\left[\left(\|\mathbf{v}^{(j)}\|_1-\frac{\sum_{i,j\in G_j,i\neq j}v_j|\beta_i|}{\|\boldsymbol{\beta}^{(j)}\|_1}\right)\alpha\|\boldsymbol{\beta}^{(j)}\|_1+(1-\alpha)w_j\sqrt{p_j}\|\boldsymbol{\beta}^{(j)}\|_2\right]. \tag{5.21}$$

Setting

$$\tau_j'=\alpha\|\mathbf{v}^{(j)}\|_1-\frac{\alpha\sum_{i,j\in G_j,i\neq j}v_j|\beta_i|}{\|\boldsymbol{\beta}^{(j)}\|_1}+(1-\alpha)w_j\sqrt{p_j},$$

simplifies Equation 5.21 to

$$\sum_{j=1}^{m}\tau_j'\left[\left(\frac{\tau_j'-(1-\alpha)w_j\sqrt{p_j}}{\tau_j'}\right)\|\boldsymbol{\beta}^{(j)}\|_1+\left(\frac{(1-\alpha)w_j\sqrt{p_j}}{\tau_j'}\right)\|\boldsymbol{\beta}^{(j)}\|_2\right]. \tag{5.22}$$

Further, setting

$$\epsilon_j'=\frac{(1-\alpha)w_j\sqrt{p_j}}{\tau_j'},$$

allows Equation 5.22 to be written in terms of the $\epsilon$-norm

$$J_{\text{asgl}}(\boldsymbol{\beta};\mathbf{v},\mathbf{w})=\sum_{j=1}^{m}\tau_j'\left[(1-\epsilon_j')\|\boldsymbol{\beta}^{(j)}\|_1+\epsilon_j'\|\boldsymbol{\beta}^{(j)}\|_2\right]=\sum_{j=1}^{m}\tau_j'\|\boldsymbol{\beta}^{(j)}\|_{\epsilon_j'}^{*}.$$

□

The connection between aSGL and the $\epsilon$-norm is further sharpened theoretically: Lemma 5.2 guarantees that the $\tau_j'$ term in Equation 5.20 always exists, even under inactive groups. Lemma 5.3 guarantees that the expression of aSGL as the $\epsilon$-norm correctly reduces to the SGL $\epsilon$-norm representation under constant weights.

**Lemma 5.2.** *For any inactive group $j\notin\mathcal{A}_g$, the limit defining $\tau_j'$ in Equation 5.20 exists and is finite.*

*Proof of Lemma 5.2.* Under $\boldsymbol{\beta}^{(j)}\equiv\mathbf{0}$ for a group $j\notin\mathcal{A}_g$, the middle term in $\tau_j'$ becomes

$$\lim_{\boldsymbol{\beta}^{(j)}\to\mathbf{0}}\left(\frac{\alpha\sum_{i,j\in G_j,i\neq j}v_j|\beta_i|}{\|\boldsymbol{\beta}^{(j)}\|_1}\right)=\frac{\alpha(p_j-1)}{p_j}\sum_{i=1}^{p_j}v_i,$$

so that $\tau_j'$ still exists. This can be observed by using L'Hôpital's rule and noting that,

$$\frac{\partial}{\partial\beta_i}\sum_{i\neq j}v_j|\beta_i|=\sum_{i\neq j}v_j,\quad\frac{\partial}{\partial\beta_i}\|\boldsymbol{\beta}^{(j)}\|_1=1,\quad\text{for } i\in G_j.$$

□

**Lemma 5.3.** *Under* $\mathbf{v} \equiv \mathbf{1}$ *and* $\mathbf{w} \equiv \mathbf{1}$ *in Equation 5.20,* $\tau_j' = \tau_j$ *and* $\epsilon_j' = \epsilon_j$ *for each* $j \in [m]$.

*Proof of Lemma 5.3.* Under $v \equiv \mathbf{1}$ and $w \equiv \mathbf{1}$, note that

$$\begin{aligned}\tau_j' &= \alpha\left(p_j - \frac{\sum_{i,j\in G_j, i\neq j} v_j|\beta_i|}{\|\boldsymbol{\beta}^{(j)}\|_1}\right) + (1-\alpha)\sqrt{p_j} \\ &= \alpha\left(p_j - \frac{(p_j-1)\|\boldsymbol{\beta}^{(j)}\|_1}{\|\boldsymbol{\beta}^{(j)}\|_1}\right) + (1-\alpha)\sqrt{p_j} = \alpha + (1-\alpha)\sqrt{p_j} = \tau_j.\end{aligned}$$

To understand the cross summation term, note that we are summing over each term $p_j - 1$ times, as the matching indices are removed, that is (without loss of generality we consider $G_1$ so that the indexing is reset from 1)

$$\begin{aligned}\sum_{i,j\in G_1, i\neq j} v_j|\beta_i| &= |\beta_1|v_2 + \cdots + |\beta_1|v_{p_1} + |\beta_2|v_1 + \cdots + |\beta_2|v_{p_1} + \cdots + |\beta_{p_1}|v_{p_1-1} \\ &= (p_1-1)|\beta_1| + \cdots + (p_1-1)|\beta_{p_1}|, \text{ by } v_j = 1, \forall j \in G_1, \text{ in SGL} \\ &= (p_1-1)\sum_{i\in G_1}|\beta_i| = (p_1-1)\|\boldsymbol{\beta}^{(1)}\|_1.\end{aligned}$$

Hence, using $w_j = 1$ and $\tau_j = \alpha + (1-\alpha)\sqrt{p_j}$,

$$\epsilon_j' = \frac{(1-\alpha)w_j\sqrt{p_j}}{\tau_j'} = \frac{(1-\alpha)\sqrt{p_j}}{\tau_j} = \frac{\tau_j - \alpha}{\tau_j} = \epsilon_j.$$

□

*Remark* 5.4. The DFR algorithm for aSGL is as described for SGL in Section 5.2.5, replacing the SGL components with the corresponding aSGL equations.

### 5.3.3 KKT checks

The KKT checks for aSGL are similar to those for SGL but include the adaptive penalties (Section 5.2.3): a KKT violation occurs for a variable $i \in G_j$ if

$$|S(\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})), \lambda_{k+1}(1-\alpha)w_j\sqrt{p_j})| > \lambda_{k+1}v_i\alpha. \tag{5.23}$$

### 5.3.4 Path start

In contrast to SGL, where the path start point can be derived from the dual norm, Lemma 5.2 shows that this is not possible for aSGL, as all groups are zero at that point and $\tau_j'$ is defined only in the limiting case when $\boldsymbol{\beta}^{(j)} \equiv \mathbf{0}$. Instead, a similar approach to that of Simon et al. [178] for SGL is applied, where the point is found by solving the piecewise quadratic, for each $j \in [m]$,

$$\left\| S\left(\mathbf{X}^{(j)\top}\mathbf{y}/n, \lambda_j \mathbf{v}^{(j)}\alpha\right)\right\|_2^2 - p_j w_j^2 (1-\alpha)^2 \lambda_j^2 = 0,$$

where $\mathbf{X}^{(j)} \in \mathbb{R}^{n \times p_j}$ is the design matrix restricted to group $j$ and $\mathbf{v}^{(j)} \in \mathbb{R}^{p_j}$ contains the variable weights for group $j$. The quadratic is solved using a root-finding bisection method [29]. Choosing $\lambda_{\max} = \max_{j \in [m]} \lambda_j$ gives the path start point.

## 5.4 Synthetic Data Experiments

In this section, the efficiency and robustness of DFR is evaluated through the analysis of synthetic data that capture different data characteristics, and assessed against other screening techniques. The aim is to compare the dimensionality reduction of the input space and the improvement in runtime.

DFR is compared with the existing SGL screening rules sparsegl and GAP safe. Both methods are described in detail in Section 5.1.2. Table 5.3 shows a summary of all rules considered.

| **Method** | **Type** | **Rules** (discard if true) | |
|---|---|---|---|
| | | Variable | Group |
| DFR-aSGL | Heuristic | $\lvert\nabla_i f(\hat{\beta}(\lambda_k))\rvert \leq \alpha v_i(2\lambda_{k+1} - \lambda_k)$ | $\lVert\nabla_j f(\hat{\beta}(\lambda_k))\rVert_{\epsilon_j'} \leq \gamma_j(2\lambda_{k+1} - \lambda_k)$ |
| DFR-SGL | Heuristic | $\lvert\nabla_i f(\hat{\beta}(\lambda_k))\rvert \leq \alpha(2\lambda_{k+1} - \lambda_k)$ | $\lVert\nabla_j f(\hat{\beta}(\lambda_k))\rVert_{\epsilon_j} \leq \tau_j(2\lambda_{k+1} - \lambda_k)$ |
| sparsegl | Heuristic | - | $\lVert S(\nabla_j f(\hat{\beta}(\lambda_k)), \lambda_k\alpha)\rVert_2 \leq \sqrt{p_j}(1-\alpha)(2\lambda_{k+1} - \lambda_k)$ |
| GAP safe | Exact | $\lvert X_i^\top \Theta_c\rvert + r\lVert X_i\rVert_2 < \alpha$ | $\mathcal{T}_j < (1-\alpha)\sqrt{p_j}$ |

Table 5.3: A summary of the four screening rules for SGL considered.

For both the synthetic and real data (Section 5.5) experiments, the SGL optimisation for DFR and sparsegl is performed using the ATOS algorithm. GAP safe is implemented using Block Coordinate Descent (BCD) [152]. ATOS is used for DFR due to its flexibility, as it can be easily adapted for use with different sparse-group penalties (by simply swapping out the proximal operators). However, DFR is

compatible with any fitting algorithm, including BCD. Accordingly, a comparison between BCD and ATOS for DFR is also conducted.

### 5.4.1 Simulation Design

The synthetic data was generated using a linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X} \sim \boldsymbol{\mathcal{N}}_p(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^{200\times 1000}$, with noise $\boldsymbol{\epsilon} \sim \boldsymbol{\mathcal{N}}_n(\mathbf{0}, \mathbf{I}_n)$, and where $\boldsymbol{\beta}$ is a sparse vector with the signal sampled from $\beta_i \sim \mathcal{N}(0, 4), i \in S_v$ (signal strength of zero). For $\mathbf{X}$, correlation was applied inside each group, such that $\Sigma_{i,j} = \rho = 0.3$ for $i$ and $j$ in the same group. The variables were placed in even groups of sizes 20, with 0.8 group sparsity proportion and 0.8 variable sparsity proportion inside the active groups.

The preceding text outlines the baseline simulation setup. In this section, we vary certain parameters to generate different cases, including signal strength, $\alpha$, input dimensionality, interaction data, and uneven group structures. These variations allow us to assess different aspects of DFR, such as comparisons with GAP safe, improvements under increasing dimensionality, robustness, and suitability for hyperparameter tuning. Each simulation case was repeated 100 times and the results are averaged across these repetitions, unless otherwise stated.

### 5.4.2 Model Hyperparameters

The models were fit along a 50-length log-linear path, starting at $\lambda_1 = \lambda_{\max}$ (as defined by Section 5.2.4 for SGL and Section 5.3.4 for aSGL) and terminating at $\lambda_{50} = 0.1\lambda_1$. The SGL models used $\alpha = 0.95$ as a default setting.

### 5.4.3 Metrics

The same primary metrics as in Chapter 4 are used for evaluation: *improvement factor* and *input proportion* (see Section 4.6.3 for a description).

Further simulation design and model implementation information can be found in Section A.3.

### 5.4.4 Results

#### Comparison to GAP Safe Rules

GAP safe is the most direct competitor to DFR, as it is the only other existing SGL rule that performs bi-level screening, and it provides a basis for a broader discussion

of strong versus safe screening rules. For comparison with GAP safe, both the signal strength and $\alpha$ are varied, allowing evaluation with respect to key data-generating and model parameters. The strength of the signal $\beta_i \sim \mathcal{N}(s, 4)$, was varied between $s \in [0, 20]$ and $\alpha$ was varied in $\alpha \in [0.1, 0.9]$.

Compared to the GAP safe rules, DFR demonstrates a significantly higher improvement factor across variations in both $\alpha$ (Figure 5.2) and signal strength (Figure 5.3). This holds when compared to both the dynamic and sequential GAP rules. In fact, despite the input proportion of DFR and GAP safe being of similar levels, the cost of calculating safe regions appears to nullify any gain in dimensionality reduction. This comparison shows that the two reduction approaches (strong vs safe) arrive at very similar results (the screened sets), but DFR achieves this with far greater computational efficiency.

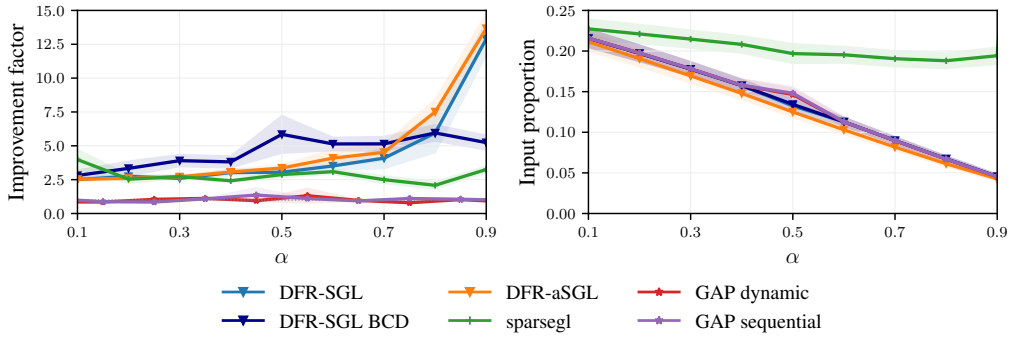



Figure 5.2: The improvement factor (left) and input proportion (right) for the screening methods applied to synthetic data, as a function of the $\alpha$ under even groups, with 95% confidence intervals.

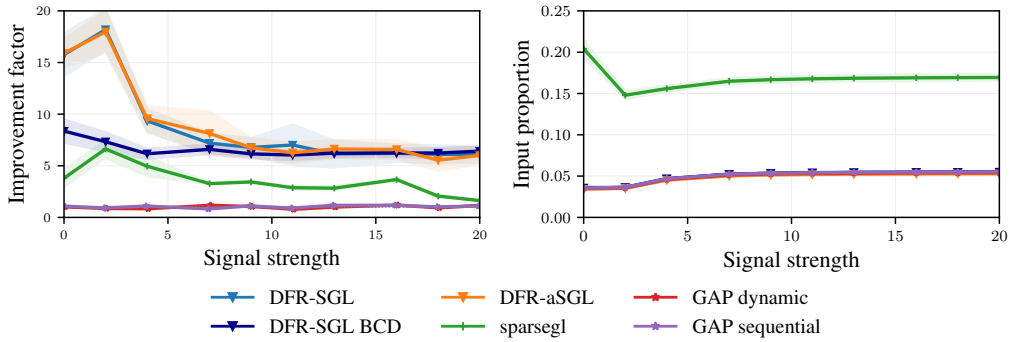



Figure 5.3: The improvement factor (left) and input proportion (right) for the screening methods applied to synthetic data, as a function of the signal strength under even groups, with 95% confidence intervals.

Across all values of $\alpha$, DFR consistently achieves substantial input space reduction, with screening efficiency decreasing approximately linearly (Figure 5.2). When $\alpha$ is close to zero, the SGL model picks more variables within each group as active, thereby limiting the potential reduction of the input space. In these settings, the contribution of the second screening layer diminishes, as evidenced in the similar performance of all methods. However, approaching the commonly used value of $\alpha = 0.95$, the advantages of DFR become increasingly clear. This is especially apparent when comparing to sparsegl, which has a significantly higher input proportion close to $\alpha = 0.95$.

The same performance hierarchy is observed across both cases: DFR performs best, followed by sparsegl, and then GAP safe.

*Remark* 5.5. Due to the poor performance of the GAP safe methods and the challenges encountered during their implementation, they are not considered further in this chapter. GAP safe is limited to linear regression and exhibited computational issues when handling uneven group sizes, which we consider in this section. Consequently, the remaining comparisons focus primarily on sparsegl.

### Comparison to BCD

The GAP safe rules are implemented using BCD, while DFR uses ATOS. In Figures 5.2 and 5.3, DFR was also implemented using BCD, demonstrating similar performance to ATOS. This shows that the choice of fitting algorithm has little impact on the gain from screening, and that DFR outperforms GAP safe regardless of fitting algorithm. In general, we apply DFR using ATOS because of the added flexibility it provides, which allows DFR to be extended to aSGL far more easily than would be possible with BCD. This is also the case for any future extensions of DFR to other sparse-group models.

### Increased Dimensionality

The benefits of DFR over sparsegl are observed under varying input dimensionality (Figure 5.4), peaking at the case $p = 800$; the input dimensionality was varied between $p \in [200, 5000]$.

With groups of a fixed size of 20, their relative size to $p$ changes, suggesting an optimal grouping regime around $p = 800$ for screening. For small $p$, only a few large groups exist, limiting the ability to eliminate groups through screening, while

for large $p$, the abundance of small groups reduces the overall efficiency of group screening.

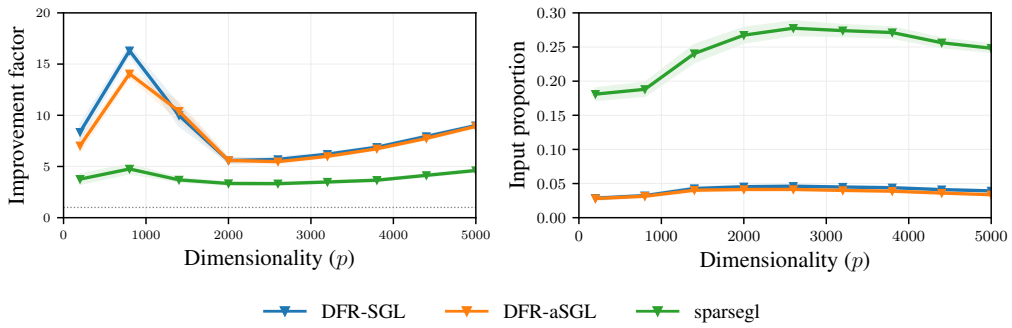



Figure 5.4: The improvement factor (left) and input proportion (right) for the strong rules applied to synthetic data, as a function of $p$ under even groups, with 95% confidence intervals.

### Robustness

*Remark* 5.6. For the remainder of this section, the variables were organised into $m = 22$ uneven groups with sizes ranging from $[3, 100]$. This configuration introduces an additional level of heterogeneity into the grouping structure, reflecting more realistic scenarios where group sizes are rarely uniform. Uneven groupings also pose a greater challenge for screening and optimisation procedures, as smaller groups can be more easily discarded while larger groups increase computational burden. Evaluating DFR under these conditions, therefore, provides a stronger indication of its robustness and efficiency across diverse problem settings.

The previous experiments demonstrated that DFR is robust under varying signal strength and $\alpha$. We observe that DFR is also robust to the data-generating parameters of signal sparsity (variable and group sparsity proportions varied together) and group correlation in $\mathbf{X}$ (Figures 5.5 and 5.6); the signal proportion was varied between $[0.05, 0.4]$ and correlation was varied in $\rho \in [0, 0.9]$.

A clear advantage of DFR over sparsegl is seen under sparse signal settings, where screening rules tend to be more effective; sparser signals present screening rules with more opportunities to discard inactive features. However, as the signal saturates, these opportunities disappear, leading to similar performance across methods.

Under varying correlation, DFR is more successful at reducing the input space compared to sparsegl, particularly under minor correlation. Under higher correla-

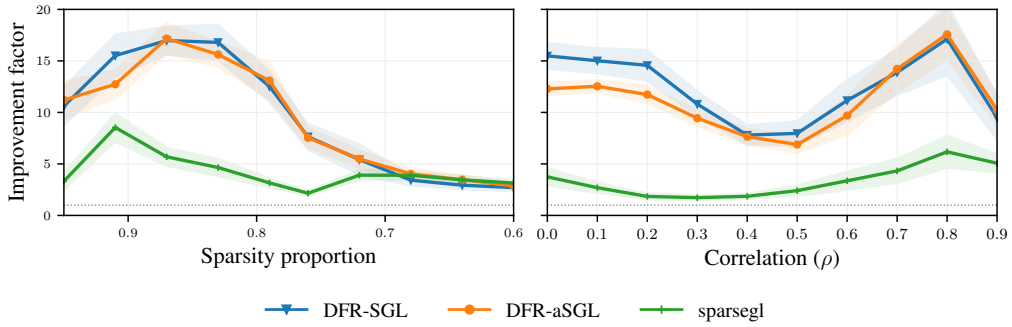



Figure 5.5: The improvement factor for the strong rules applied to synthetic data, as a function of the sparsity proportion (left) and data correlation (right) under uneven groups, with 95% confidence intervals.

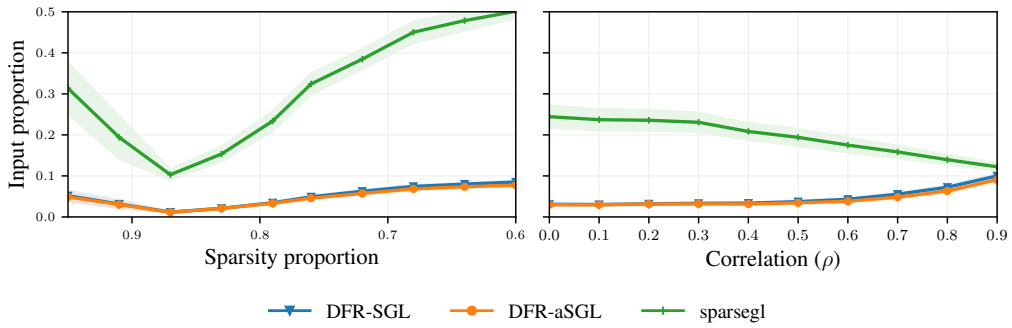



Figure 5.6: The input proportion for the strong rules applied to synthetic data, as a function of the sparsity proportion (left) and data correlation (right) under uneven groups, with 95% confidence intervals.

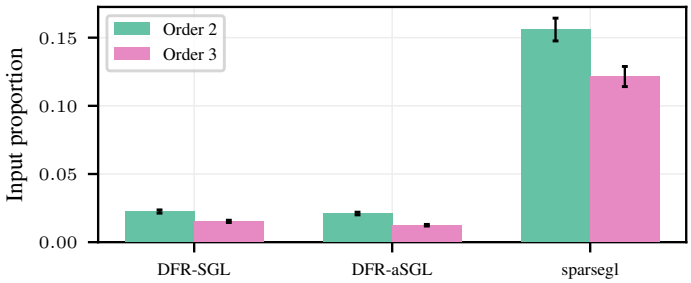
tion, the models become less sparse, which results in reduced screening importance.

Similar to the case with even groups, for uneven groups, DFR remains relatively unaffected by the signal strength and the choice of $\alpha$, consistently achieving effective reduction (Figures D.1 and D.2).

**Interactions.** Another robustness check of DFR is to assess how including interaction terms affects the efficiency, thereby also evaluating further the impact of increased dimensionality on screening (Figure 5.7). All possible interactions of orders 2 and 3 within each group were included in the design matrix and assigned their own group. The interaction terms are assigned to their own groups to ensure a fair comparison between DFR and sparsegl. Placing them in separate groups prevents DFR from gaining an advantage. If interactions were instead included in the original

variable groups, sparsegl would be unable to discard them individually, potentially exaggerating DFR's relative performance. No interaction hierarchy is imposed, as this would alter the penalisation structure and invalidate the screening rules.

In this setting, DFR demonstrates large computational improvements. This is particularly true over sparsegl, which yields only marginal improvements under third-order interactions. These savings make it more feasible for sparse-group models to be used in interaction detection problems. Such problems are frequently encountered in the field of genetics, where gene-gene and gene-environment relationships are vital discoveries [47, 228].




(a) Input proportion.

| Method | Interaction | |
|---|---|---|
| | 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) |

(b) Improvement factor.

Figure 5.7: The strong rules applied to synthetic interaction data, under the linear model, with standard errors. The parameters of the data were changed for these experiments from the baseline to account for the increased dimensionality of interaction data: they were set as $p = 400, n = 80$, and $m = 52$ groups of sizes in $[3, 15]$. The interaction input dimensionality was $p_{O_2} = 2111$ and $p_{O_3} = 7338$, with no interaction hierarchy imposed. The sparsity proportion of interaction variables was set to 0.7 (with the same signal as the marginal effects).

### Hyperparameter Tuning and Cross-Validation

DFR-aSGL demonstrated robust performance across a range of hyperparameter values for $b_1$ and $b_2$, which define the shape of the adaptive weights (Figure 5.8; see Section 5.3.1 for a description of the weights); the weights were varied together in $b_1, b_2 \in [0, 2]$.

The efficiency and stability of DFR across different hyperparameters $(\alpha, b_1, b_2)$ make it a practical tool for enabling full hyperparameter tuning, via ap-

| 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) |

Table 5.4: Improvement factor of strong rules on synthetic data using 10-fold CV, with standard errors shown in grey.

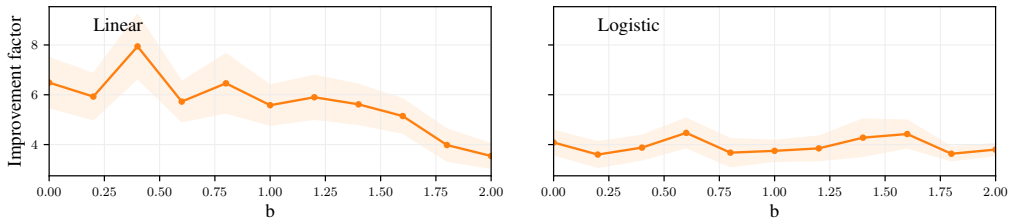



Figure 5.8: The improvement factor of DFR-aSGL under different weights $b_1 = b_2$, shown for the linear (left) and logistic (right) models, with 95% confidence intervals.

proaches like CV, a process that is rarely done due to computational cost. Applying DFR with CV yields significant computational savings (Table 5.4). These findings highlight DFR's value in facilitating expanded tuning regimes for SGL and aSGL.

#### Logistic Model

DFR was also applied to logistic models under the same cases (Figures D.5, D.6, D.7, and D.8; see Appendix D.3.2 for further results and description of the data generation). Across all cases considered for the logistic model, DFR outperforms sparsegl in terms of improvement factor and input proportion by a considerable margin. DFR is also found to be effective and robust, for both SGL and aSGL.

### 5.4.5 KKT Violations

KKT violations for DFR-SGL are extremely rare. Across all linear model experiments, only a single KKT violation was observed (Table D.5). This indicates that the assumptions of DFR-SGL are only rarely violated.

Violations are somewhat more frequent for DFR-aSGL and sparsegl, though they remain relatively uncommon. DFR-aSGL monitors violations at the variable level, while sparsegl does so at the group level, which increases the likelihood of a violation for DFR-aSGL.

The higher number of violations seen for sparsegl in comparison to DFR-SGL suggests that the group Lipschitz assumption underlying DFR-SGL provides a more robust and sensible basis for screening. One explanation for the increased KKT violations of DFR-aSGL stems from the nature of its Lipschitz assumptions. Unlike SGL, the adaptive penalties in aSGL introduce additional dependencies on hyperparameters into the Lipschitz assumptions. This extra dependence on hyperparameters has the potential to lead to violations.

## 5.5 Real Data Experiments

DFR is applied to real datasets to examine whether the computational gains observed on synthetic data translate to real data.

### 5.5.1 Datasets

Six real datasets are used, differing in response type, grouping structure, and input dimensionality. DFR is compared to sparsegl; see Remark 5.5 for the rationale behind excluding GAP safe in this section.

Three of the datasets, BRCA1, Scheetz, and Trust-experts, have continuous responses and are fit using an SGL linear model. BRCA1 and Scheetz were analysed with regard to screening rules in Larsson and Wallin [110], while Trust-experts was examined in Liang et al. [121]. The other three datasets, Adenoma, Celiac, and Tumour, have binary responses, so an SGL logistic model is used. Scheetz, Adenoma, Celiac, and Tumour were used to test the group-based SLOPE screening rules in Chapter 4 (Section 4.7). The Trust-experts dataset is low-dimensional, and the other five are high-dimensional.

Further details on the datasets are provided in Table 5.5 and Appendix A.3.3.

| **Dataset** | $p$ | $n$ | $m$ | **Group sizes** | **Grouping** | **Response** | **Downloaded on** |
|---|---|---|---|---|---|---|---|
| BRCA1 | 17322 | 536 | 243 | $[1, 6505]$ | SVD | Continuous | 05/2024 |
| Scheetz | 18975 | 120 | 85 | $[1, 6274]$ | SVD | Continuous | 05/2024 |
| Trust-experts | 101 | 9759 | 7 | $[4, 51]$ | Factors | Continuous | 05/2024 |
| Adenoma | 18559 | 64 | 313 | $[1, 741]$ | Gene sets | Binary | 05/2024 |
| Celiac | 14657 | 132 | 276 | $[1, 617]$ | Gene sets | Binary | 05/2024 |
| Tumour | 18559 | 52 | 313 | $[1, 741]$ | Gene sets | Binary | 05/2024 |

Table 5.5: Dataset information for the six datasets used in the real data analysis.

### 5.5.2 Model Hyperparameters

The models were fit along a 100-length log-linear path, starting at $\lambda_{\max}$ and terminating at $0.2\lambda_{\max}$. SGL and aSGL models used $\alpha = 0.99$ and aSGL used $b_1 = b_2 = 0.1$.

Detailed model and algorithmic hyperparameter information is provided in Tables A.4 and A.6.

### 5.5.3 Results

For all datasets, DFR outperforms sparsegl at reducing the computational cost and input dimensionality (Figure 5.9). DFR keeps the input proportion low along the whole path (Figure 5.10). Although DFR is motivated by high-dimensional problems, where screening has the greatest impact, it also improves fitting time in low-dimensional datasets (Trust-experts).

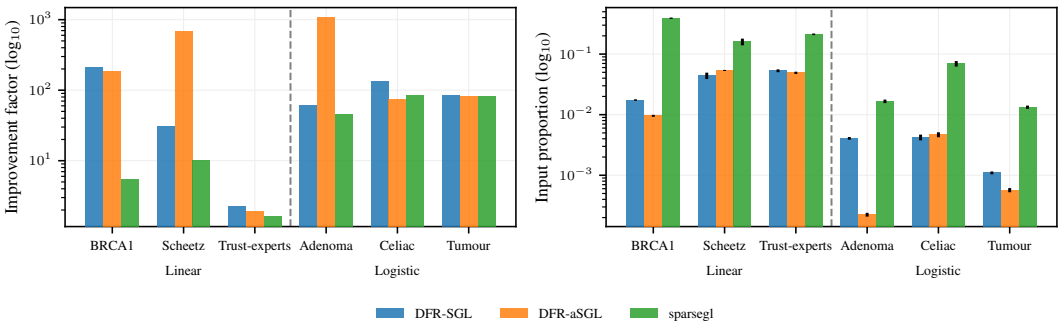



Figure 5.9: The improvement factor (left) and input proportion (right) ($\log_{10}$ scale) of the strong rules applied to the six real datasets, split by model type.

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 greater difficulty converging without screening compared to SGL, which made DFR-aSGL particularly advantageous relative to DFR-SGL. For Adenoma, the active set for aSGL was smaller (Table D.11), a consequence of the increased penalisation introduced by adaptivity. Notably, even with the benefit of a smaller active set, DFR-aSGL still achieved a more efficient reduction of the optimisation set, *with respect to the active set*.

DFR is found to mitigate convergence issues for both SGL and aSGL (Table D.12). Across all datasets, DFR achieved full convergence with no failures. In contrast, sparsegl failed to converge at several path points for both the Adenoma and Scheetz datasets. Since sparsegl screens only at the group level, once a group enters the optimisation set, the entire group must be fit, even if it includes noise variables (potentially hindering convergence). Similarly, when no screening is applied, SGL fails to converge for Adenoma, Scheetz, and Tumour. By drastically reducing the input space, DFR helps resolve convergence issues associated with high-dimensional data, thereby improving both computational efficiency and solution optimality.

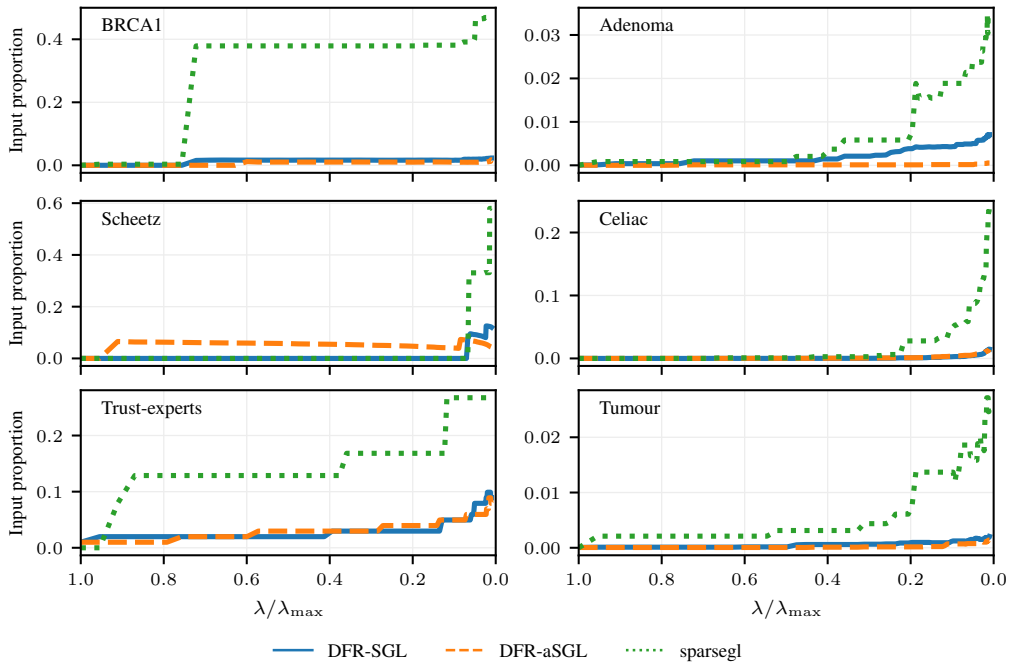



Figure 5.10: The input proportion as a function of the shrinkage path for the strong rules applied to the real datasets.

### 5.5.4 KKT Violations

On the real data, sparsegl and DFR-SGL had no KKT violations (Tables D.10 and D.11). DFR-aSGL observed variable KKT violations on all datasets, but the number of violations was very low; occurring on average at a rate of 0.045 per path model.

## 5.6 Discussion

In this chapter, we introduce a novel screening approach for the sparse-group lasso and adaptive sparse-group lasso, called *Dual Feature Reduction* (DFR), derived using the dual norms of SGL and aSGL. DFR introduces the first bi-level strong screening rules for SGL and the first screening rules for aSGL. DFR is computationally much simpler than the GAP safe rules, which require iterative screening and fitting. In comparison, DFR screens only once per path point, introducing minimal computational overhead.

DFR begins with group-level screening to eliminate inactive groups, followed by variable-level reduction that removes inactive variables within the remaining groups.

By discarding variables that are inactive at the optimal solution, DFR significantly reduces computational cost, allowing the SGL family of models to scale more efficiently with increasing dimensionality and handle larger, more complex datasets. This gain comes at no cost, as the optimal solution is still achieved (Appendices D.3.1, D.3.2, and D.3.3). In fact, by reducing the input, instances were observed where DFR helped SGL and aSGL overcome convergence issues.

DFR demonstrated robustness across various data and model parameters, consistently achieving drastic feature reduction in all scenarios considered. This translated into significant computational savings on both synthetic and real data. DFR outperformed all other screening approaches, establishing it as the state-of-the-art screening method for SGL and highlighting the benefit of bi-level screening.

SGL models have widespread applications, and our DFR rules enable them to be applied far more efficiently, drastically reducing the additional computational cost arising from a sparse-group norm. This efficiency also makes it feasible to apply SGL to high-dimensional data with CV, allowing for data-driven tuning of $\alpha$, which is often chosen subjectively in practice due to the computational burden of a grid search.

# CHAPTER 6

## Background Materials for "Model Selection for SLOPE Models: A Bayesian Perspective"

In Chapter 7 we tackle the problem of model selection for SLOPE models. To do this, we develop new Bayesian methodology. This chapter presents relevant model selection and Bayesian background materials for that chapter. The background material is extensive; therefore, a separate chapter has been written.

## 6.1 Model Selection

There are many approaches for model selection for penalised models, and here we broadly group them into *estimation* and *tuning* regimes. To tune $\lambda$, we typically fit a path of $l$ models $\mathcal{M} = \{\hat{\boldsymbol{\beta}}(\lambda_1), \dots, \hat{\boldsymbol{\beta}}(\lambda_l)\}$, where $\lambda_1 \geq \cdots \geq \lambda_l > 0$, and choose the optimal model, $\hat{\mathcal{M}}$, from the path according to some metric. In an ideal world, the solution path would be *path consistent*, meaning the true model is contained along the path [85]. Zhao and Yu [236] show that conditions exist for the lasso to be path consistent, but no similar results for SLOPE exist.

On the other hand, estimation of $\lambda$ provides a model $\hat{\mathcal{M}} = \hat{\boldsymbol{\beta}}(\hat{\lambda})$. Both approaches aim to obtain a final optimal solution $\hat{\boldsymbol{\beta}}$ and a corresponding set of selected variables $\hat{S}_v$, which, in the context of SLOPE models, directly determines the FDR.

We briefly discuss relevant approaches here, and the approaches used in our synthetic study in Chapter 7 are discussed in-depth in Section 6.2.

### 6.1.1 Tuning Along a Path

The most common approach to tuning is via CV with predictive performance used as a model selection metric. While predictive performance is often used as a selection criterion, in genetics, the primary focus is typically on model recovery to enable FDR control. These two objectives may not necessarily align in selecting the same model, especially as CV tends to select overfitting models [118, 221]. When the objective is to control the FDR, identifying an appropriate criterion for CV is not straightforward. In genetics, the type I error rate is commonly used (although Barber and Candès [15] advocate for the use of a type-S error), yet this criterion does not provide finite-sample guarantees [24]. Even though not directly related to our problem, it is interesting to note that the lasso lacks selection consistency when $\lambda$ is chosen using CV [118].

CV can also introduce bias and often selects the most saturated model on a path, which leads to high FDR levels in SLOPE models [134]. There are many CV variants, such as stable [159], modified [224], nested [19], and generalised [79] CV (see Arlot and Celisse [5] for a summary). In this thesis, we applied 10-fold cross-validation and selected the 1se model [184].

An approach that builds on CV is the Knockoffs procedure (Section 6.2.2), proposed by Barber and Candès [14]. The Knockoffs procedure generates synthetic copies of the input data that are independent of the response, but with the same structure as the original data and act as negative controls. The Knockoffs method applies CV using both the original and knockoff variables, and then applies the Knockoffs filter to the selected model, $\hat{\mathcal{M}}$, to remove noise variables for controlling the FDR.

A notable feature of path-based tuning is that only one model is ultimately selected, while potentially valuable information from the rest of the path is discarded. On the other hand, stability selection [131] utilises information from across the full path to stabilise the variable selection procedure of any penalised model, providing family-wise error rate guarantees. Ahmed et al. [2] extend the stability selection approach for FDR control; however, our experiments show that it produces unsatisfactory FDR control compared to other competitive methods and is computationally expensive (Appendix E.6.1). Other extensions to stability selection include kappa selection [190] and complementary pairs stability selection [175].

### 6.1.2 Noise Estimation

Model selection methods can generally be considered as estimators of some quantity. For penalised regression models, estimation approaches tend to directly estimate the noise $\sigma^2$ and use it to fit the final selected model. Popular noise estimation procedures include Fan et al. [61], which uses a two-stage data splitting technique to calculate an estimator, and Dicker [51], which derive new consistent method-of-moments-based estimators using no sparsity assumptions on $\boldsymbol{\beta}$. See Reid et al. [154] for a comprehensive study of variance estimators.

An estimation approach tailored toward penalised regression is scaled regression, described in Section 6.2.1. This method jointly estimates the coefficients and noise through an iterative procedure that scales the regularisation parameter using the mean residual square. In Chapter 7, we modify it to be adaptive for SGS (Section 7.6.1).

### 6.1.3 Bayesian Approaches

An alternative estimation approach is to learn the noise together with the other model parameters in a Bayesian penalised regression model (Section 6.3.1). A common strategy for this is to use spike-and-slab priors (Section 6.3.2).

The (continuous) spike-and-slab framework was extended to SLOPE models in Jiang et al. [99] to form ABSLOPE (Section 7.2). It was only developed for the SLOPE model and not for group-based ones. Bayesian group-based lasso models were developed in Xu and Ghosh [217] using (point-mass) spike-and-slab priors. The (continuous) spike-and-slab approach has also been applied to the group lasso in Bai et al. [10]. Additional Bayesian model selection techniques have been proposed in the literature [40, 101, 104, 200, 199, 210].

Section 6.3 outlines the key Bayesian topics relevant to Chapter 7.

### 6.1.4 Additional Approaches

Other model selection approaches that do not necessarily fall into the categories discussed above include post-selection inference [117, 153, 195], which provides statistical inference after variable selection; information criteria [4, 172, 183] which use information theory to guide model choice; and methods for direct FDR estimation [127, 182]. There are also modifications of the lasso aimed at improving model selection, including the rank lasso [207], ET-LASSO [219], and TREX [115].

A short simulation study, varying the correlation and using the lasso, compared some of the approaches described above to identify those with potential FDR control for further investigation (Appendix E.6.1). Among the approaches considered, we selected Knockoffs for its strong performance in the study, and CV, due to its widespread use, for further investigation in our synthetic data analysis (Section 7.7).

For additional approaches to model selection and parameter tuning in high-dimensional regression, see relevant summaries [89, 116, 213].

## 6.2 Relevant Approaches

In this section, we provide detailed descriptions of the key approaches to model selection for SLOPE models used in our synthetic study in Chapter 7, expanding on the brief overview presented above.

### 6.2.1 Scaled Regression

*Scaled Regression* jointly estimates the coefficients and noise in a penalised regression model. It uses unbiased estimators when $n > p$ and iterative procedures when $p \geq n$ [189]. Originally introduced for the lasso, it was later extended to SLOPE in Bogdan et al. [24]. The noise is calculated using the estimator

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)\|_2^2}{n - |\hat{S}_v^\lambda|},$$

where $|\hat{S}_v^\lambda|$ is the number of non-zero elements in $\hat{\boldsymbol{\beta}}(\lambda)$ (the degrees of freedom). The regression coefficients are calculated by fitting the penalised regression model with regularisation parameter $\lambda = \hat{\sigma}/n$ (the $1/n$ factor is from the loss function). This process is repeated until $\hat{S}_v^\lambda$ stabilises between iterations. The procedure is described fully in Algorithm 4. The algorithm can be adapted to gSLOPE and SGS by using their fitted regression coefficients in place of the SLOPE estimates in Step 5.

**Algorithm 4** Scaled regression for SLOPE models [24]

**Require:** $\mathbf{y}$, $\mathbf{X}$

1: Initialise: $\hat{S}_v^+ \leftarrow \varnothing$
2: **while** $\hat{S}_v^+ \neq \hat{S}_v$ **do**
3: $\quad \hat{S}_v \leftarrow \hat{S}_v^+$
4: $\quad$ Set $\hat{\sigma}^2 = \frac{\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)\|_2^2}{n-|\hat{S}_v^\lambda|}$
5: $\quad$ Compute the solution $\hat{\boldsymbol{\beta}}$ via a SLOPE model with $\lambda = \hat{\sigma}/n$
6: $\quad$ Set $\hat{S}_v^+ \leftarrow \text{supp}(\hat{\boldsymbol{\beta}})$
7: **end while**

### 6.2.2 Knockoffs

*Knockoffs* has been gaining popularity in the multiple testing literature, due to its flexibility and powerful properties. Knockoffs provides FDR control by generating knockoff copies of variables that act as negative controls. By comparing each variable to its knockoff, the procedure approximates the number of false positives and retains only variables that show a clear advantage over their knockoff counterparts.

The original Knockoffs procedure was limited to low-dimensional settings [14]. Barber and Candès [15] extended it to high-dimensional data using the fixed-X paradigm, which was later generalised to the model-X framework [38]. This framework also allows for a Bayesian construction, which the authors briefly explored.

**Construction.** Formally, a set of *knockoff variables* $\mathbf{X}_{\text{KO}} = (\mathbf{X}_{\text{KO}})_1, \ldots, (\mathbf{X}_{\text{KO}})_p$ for random variables $\mathbf{X} = \mathbf{X}_1, \ldots, \mathbf{X}_p$ satisfy [38]:

1. *Exchangeability*: $(\mathbf{X}, \mathbf{X}_{\text{KO}})_{\text{swap}(M)} \overset{d}{=} (\mathbf{X}, \mathbf{X}_{\text{KO}})$ for any $M \subset 1, \ldots, p$,

2. *Independence*: $\mathbf{X}_{\text{KO}} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{X}$,

where $(\mathbf{X}, \mathbf{X}_{\text{KO}})_{\text{swap}(M)}$ denotes the matrix obtained by swapping $\mathbf{X}_i$ and $(\mathbf{X}_{\text{KO}})_i$ for each $i \in M$. Knockoff variables can be generated using various methods, including Markov models [174], Metropolis-Hastings [18], deep learning [164], and non-parametric approaches [23]. In Chapter 7, the second-order model-X construction is used, which matches the first two moments of $(\mathbf{X}, \mathbf{X}_{\text{KO}})$ and $(\mathbf{X}, \mathbf{X}_{\text{KO}})_{\text{swap}(M)}$ [38].

**Feature statistic.** The knockoff variables are combined with the originals to form the concatenated design matrix $[\mathbf{X}\ \mathbf{X}_{\text{KO}}]$, which is used as input to a penalised regression model (typically with a lasso penalty) to produce $[\hat{\boldsymbol{\beta}}\ \hat{\boldsymbol{\beta}}_{\text{KO}}]$. A feature statistic is then computed for each variable. The coefficient-difference statistic is often used, defined for $i \in [p]$ by

$$W_i = |\hat{\beta}_i| - |(\hat{\beta}_{\text{KO}})_i|. \tag{6.1}$$

See Weinstein et al. [212] for a comparison of different statistics.

The feature statistic must satisfy the *coin flip property*, which specifies that swapping a variable with its knockoff must reverse the sign of its statistic, and the *sufficiency property*, meaning it depends on the design only through the covariance and on the response through marginal correlations [15]. Because knockoffs serve as controls, a null variable is equally likely to be selected as its knockoff, making $W_i$ symmetric around zero [155].

**Filter.** The Knockoffs filter applies a data-adaptive significance threshold to the feature statistic,

$$T = \min\left\{t > 0 : \frac{\#\{i \in [p] : W_i \leq -t\}}{\#\{i \in [p] : W_i \geq t\}} \leq q_{\text{KO}}\right\}, \tag{6.2}$$

where $q_{\text{KO}} \in (0, 1)$ is the target FDR level, such that the filter generates a set of active variables $\hat{S}_v = \{i \in [p] : W_i \geq T\}$. FDR control follows from exchangeability, which makes the signs of the null statistics $W_i$ independent coin flips. Consequently, the threshold $T$ gives a conservative estimate of false discoveries [14].

**Limitations and improvements.** The Knockoffs framework has some key limitations. Doubling the data dimensionality can be computationally expensive, which has motivated the development of screening rules for Knockoffs [124, 144]. Additionally, the knockoff generation process can be random, producing different results across runs. Consequently, several approaches have been proposed to *derandomise* Knockoffs by aggregating results from multiple runs [156], including using e-values [155].

### Knockoffs with SLOPE

SLOPE and Knockoffs both control the FDR, but under different assumptions. SLOPE assumes an orthogonal design matrix, limiting the FDR properties to low dimensions. The model-X Knockoffs framework has no dimensionality restrictions, but assumes the feature distribution $F_X$ is known and the conditional distribution $F_{Y|X}$ can be arbitrary and unknown. These assumptions are reasonable in genetics, where $F_{Y|X}$ is difficult to estimate due to complex correlation structures, whereas $F_X$ is often known or easily estimated, as it can be experimentally controlled [174]. Knockoffs also assumes exchangeability of the knockoff variables and that the coin flip and sufficiency properties hold. Violations can make the feature statistic asymmetric, compromising FDR control [23].

The flexibility of Knockoffs stems from its use of penalised regression, allowing any estimator that satisfies the assumptions. As such, we explore combining Knockoffs with SLOPE to achieve FDR control in general settings. Humayoo and Cheng [93] showed that SLOPE with Knockoffs can improve FDR control and power over the lasso in simulations, though several questions remain.

These are: Does the SLOPE statistic satisfy the coin flip property? This has not been addressed in Humayoo and Cheng [93] or elsewhere. How does the SLOPE FDR parameter ($q_v$) interact with the Knockoffs parameter ($q_{\text{KO}}$)? Ren and Barber [155] show, for eBH and Knockoffs, that setting the Knockoffs parameter lower than the eBH one is preferable for FDR control. Finally, how should the SLOPE design be optimised for Knockoffs, and are new penalty sequences or knockoff constructions needed? This is especially relevant since SLOPE's steep decay and clustering can reduce contrast between $\mathbf{X}$ and $\mathbf{X}_{\text{KO}}$.

Although these questions are beyond the scope of this thesis, the following section explores in more detail how the Knockoffs framework can be applied to SGS.

### Sparse-group Knockoffs

Knockoffs have been applied to group regression in Dai and Barber [46], where multi-task regression is reformulated as a group lasso problem and a group-specific Knockoffs construction is proposed. For the feature statistic, the norm is modified to be

$$J_{\text{glasso, KO}}(\mathbf{b}) = \sum_{j=1}^{m} \|\mathbf{b}^{(j)}\|_2 + \sum_{j=1}^{m} \|(\mathbf{b}_{\text{KO}})^{(j)}\|_2.$$

The path statistic is then used, defined as the point along the path at which a group enters the model. To allow for direct comparison to the variable feature statistic, in this thesis, we adopt the group effects statistic for $j \in [m]$,

$$W = \|\hat{\boldsymbol{\beta}}^{(j)}\|_2 - \|(\hat{\boldsymbol{\beta}}_{\text{KO}})^{(j)}\|_2. \tag{6.3}$$

The Knockoffs construction and filter proceed as in the variable case. The knockoff variables are assigned to their own groups, so that each original group has a corresponding knockoff group.

Although multi-layer Knockoffs filters exist [16, 74, 82, 103], Knockoffs have not been specifically studied in the context of sparse-group regression models. Two main concerns in this setting are the construction of knockoffs and the coin flip property. Constructing knockoffs compatible with both variable and group penalties is challenging: the coin flip property does not hold at a variable level under a group knockoff construction [82]. Our investigation for the sparse-group lasso (SGL) and SGS shows that the coin flip property holds only if the knockoffs are placed in the same group as the originals, which may dilute group penalisation, and holds less often for SGS due to its dual sorting procedures (Figure E.7).

In this chapter, we present an initial implementation of sparse-group Knockoffs using SGS. SGS provides both variable- (Equation 6.1) and group-level (Equation 6.3) feature statistics, with corresponding thresholds used to select active variables and groups, enabling bi-level FDR control. To satisfy the coin flip property, the knockoff variables are placed into the same groups as the originals; this is in contrast to the group-only setup where they form separate groups. This gives the SGS Knockoffs problem as

$$[\hat{\boldsymbol{\beta}}(\lambda)\ \hat{\boldsymbol{\beta}}_{\text{KO}}(\lambda)] \in \underset{\mathbf{b}\in\mathbb{R}^{2p}}{\arg\min} \left\{ \frac{1}{2n}\|\mathbf{y} - [\mathbf{X}\ \mathbf{X}_{\text{KO}}]\mathbf{b}\|_2^2 + \lambda\alpha \sum_{i=1}^{2p} v_i |b|_{(i)} \right. \\ \left. + \lambda(1-\alpha) \sum_{j=1}^{2m} w_j \sqrt{p_j} \|\mathbf{b}^{(j)}\|_2 \right\}.$$

The screening rules developed in Chapter 4 reduce the computational cost of Knockoffs, but the synthetic data results in Section 7.7 show that it is the most computationally intensive model selection approach for SGS.

### 6.2.3 Two-step

Two-step procedures combine the positive elements of multiple models into a stronger overall model. Chételat et al. [42] proposed a two-step method for prediction, using the lasso followed by least squares refitting to debias coefficients. The approach addresses CV's computational inefficiency and lack of finite-sample guarantees. Similarly, Yang et al. [220] introduced the two-stage prior lasso (TSPLASSO), performing two lasso stages for gene selection and sample classification. For overviews of multi-step regression, see relevant summaries [39, 211, 232], and for FDR-controlling two-step methods see Sarkar [169].

#### Two-step SLOPE

The *Two-Step SLOPE* (TS-SLOPE) procedure has been shown to recover the true model with high probability [81, 87]. The procedure is defined by the steps

1. Obtain an initial estimate $\hat{\boldsymbol{\beta}}_{[1]}$ from applying SLOPE.
2. Obtain a truncated estimate $\hat{\boldsymbol{\beta}}_{[2]} = \text{prox}_{\text{slope}}(\hat{\boldsymbol{\beta}}_{[1]})$, where $\text{prox}_{\text{slope}}$ is the proximal operator of the SLOPE penalty (given in Algorithm 8).

The final estimate is shown to recover the pattern of the true signal, although is also shown to be heavily biased [87].

This approach can be extended to gSLOPE and SGS using the model-specific proximal operators. The gSLOPE proximal operator is given in Algorithm 8. For SGS, the proximal operators are applied one after another, so that Step 2 becomes

$$\hat{\boldsymbol{\beta}}_{[2]} = \text{prox}_{\text{slope}}(\text{prox}_{\text{gslope}}(\hat{\boldsymbol{\beta}}_{[1]})).$$

Applying the operators in the reverse order was also tested and yielded nearly identical results.

## 6.3 Bayesian Models

### 6.3.1 Bayesian Penalised Regression

Switching from a frequentist viewpoint of penalised regression to a Bayesian perspective provides many benefits. In the Bayesian framework, the priors placed on

model parameters can incorporate prior knowledge from the practitioner, allowing the model to reflect existing expertise or domain insights. One key advantage for our problem of model selection is the ability to learn the noise level directly as part of the inference process. This means that the noise level is estimated jointly with the coefficients $\boldsymbol{\beta}$, ensuring that both are informed by the same model structure and data.

The first step to a Bayesian way of thinking for penalised regression came from the observation that the lasso estimator corresponds to the posterior mode when each regression coefficient is assigned an independent double exponential (Laplace) prior [192]. Formally, any penalised estimator admits a Bayesian interpretation given by

$$\pi(\boldsymbol{\beta} \mid \lambda) = \prod_{j=1}^{p} \pi(\beta_j \mid \lambda), \quad \text{where the penalty function satisfies}$$
$$J(\boldsymbol{\beta}; \lambda) = -\log \pi(\boldsymbol{\beta} \mid \lambda) = -\sum_{j=1}^{p} \log \pi(\beta_j \mid \lambda).$$

Applied to the lasso, we have $\pi(\boldsymbol{\beta} \mid \lambda) = \prod_{j=1}^{p} \frac{\lambda}{2} e^{-\lambda|\beta_j|}$ [162]. Like the frequentist SLOPE model, the Bayesian lasso is non-separable, allowing for the sharing of information across variables. However, the Bayesian lasso has been shown to struggle with bias and suboptimal shrinkage of coefficients [78, 161].

### 6.3.2 Spike-and-slab Framework

The spike-and-slab framework applied to the lasso overcomes issues with bias and suboptimal shrinkage by modelling the noise and signal using a mixture prior [9]. The *continuous* spike-and-slab prior is of the form [77]

$$\pi\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^2\right) = \prod_{i=1}^{p} \left[(1-\gamma_i)\,\mathcal{N}\left(0, \sigma^2\tau_0^2\right) + \gamma_i \mathcal{N}\left(0, \sigma^2\tau_1^2\right)\right]$$
$$\pi(\boldsymbol{\gamma} \mid \theta) = \prod_{i=1}^{p} \theta^{\gamma_i}(1-\theta)^{1-\gamma_i},\ \theta \sim \pi(\theta),\ \sigma^2 \sim \pi\left(\sigma^2\right),$$

where $\tau_0^2$ models the noise (the *spike*) and $\tau_1^2$ captures the signal (the *slab*), such that $0 < \tau_0^2 \ll \tau_1^2$, $\boldsymbol{\gamma} \in \{0,1\}^p$ is a binary inclusion vector over $2^p$ models, and $\theta \in (0,1)$ is a mixture proportion. A key advantage of the spike-and-slab model is

its ability to automatically adapt to the underlying sparsity structure of the data through the mixing parameter $\theta$ [9]. In addition, spike-and-slab formulations yield posterior inclusion probabilities, $\boldsymbol{\gamma}$, which provide interpretable measures of variable importance and quantify uncertainty in model selection. Spike-and-slab priors also integrate naturally into hierarchical Bayesian frameworks, offering flexibility for model design. This flexibility underlies the development of our proposed models in Chapter 7. The priors enable simultaneous variable selection and parameter estimation, as well as mitigating over-shrinkage of signals by placing sufficient prior mass on large effect sizes [99].

A non-continuous alternative exists in which a point-mass is used to model the noise and a heavy-tailed distribution for the signal [133]. Additionally, the spike-and-slab lasso (SSL) was proposed by Ročková and George [162], replacing the Gaussian density with the lasso Laplace density $\psi(\boldsymbol{\beta} \mid \lambda) = (\lambda/2)e^{-\lambda|\boldsymbol{\beta}|}$, such that $\lambda$ is chosen differently for the spike/slab components.

# CHAPTER 7

## Model Selection for SLOPE Models: A Bayesian Perspective

*The AS-SGS content is based on the manuscript Feser and Evangelou [66], used for Chapter 3, but has been moved to this chapter to better align with the present topic. The rest of the content of this chapter is based on a manuscript that we are preparing for submission. Background material for this chapter is presented in Chapter 6.*

SLOPE models can control the FDR under orthogonal settings, where the noise is known and the regularisation parameter can be set accordingly. However, orthogonal settings are rare in practice. Under general settings with unknown noise, CV is typically used for model selection along the SLOPE regularisation path, but it prioritises predictive performance over FDR control. In this chapter, we address the problem of model selection for SLOPE and group-based SLOPE models, with the goal of identifying the most effective approach for applying SLOPE to general problems while ensuring FDR control.

To achieve this goal, we propose new group-based Bayesian models: Bayesian Group SLOPE (BGSLOPE) and Bayesian Sparse-group SLOPE (BSGS). The models unify group-based SLOPE norms and spike-and-slab priors. BSGS introduces the first continuous spike-and-slab framework designed specifically for sparse-group models. The models offer uncertainty quantification, are adaptive to unknown sparsity structures, and are found to be very stable to the choice of hyperparameters.

Additionally, we propose a two-step procedure, Two-step Orthogonal (TSO), that transforms a general problem to an orthogonal setting, thereby allowing the FDR properties of SLOPE to be applied, along with an adaptively scaled regression method (AS-SGS).

Synthetic and real data analyses showed that the proposed Bayesian models reliably controlled the FDR, achieved higher power, and delivered superior predictive performance compared to competing approaches.

## 7.1 Introduction

For most penalised regression approaches, the regularisation parameter, $\lambda$, determines the sparsity level in the fitted model and is proportional to the noise of the data-generating process. Indeed, most suggested schemes for $\lambda$ are as a function of $\sigma^2$ [189]. The true noise level is rarely known in practice, so estimation or tuning is required. This is particularly problematic for SLOPE models with regard to FDR control, as the sparsity of a fitted SLOPE model determines the FDR level. SLOPE theory shows that the FDR is controlled under orthogonal $\mathbf{X}$ with $\lambda = 1$ (as we know the noise level). Our results in Section 3.4 demonstrate FDR control theoretically and experimentally for SGS under orthogonal designs, with standard Gaussian errors. See Bogdan et al. [24] and Brzyski et al. [30] for similar results for SLOPE and gSLOPE, respectively.

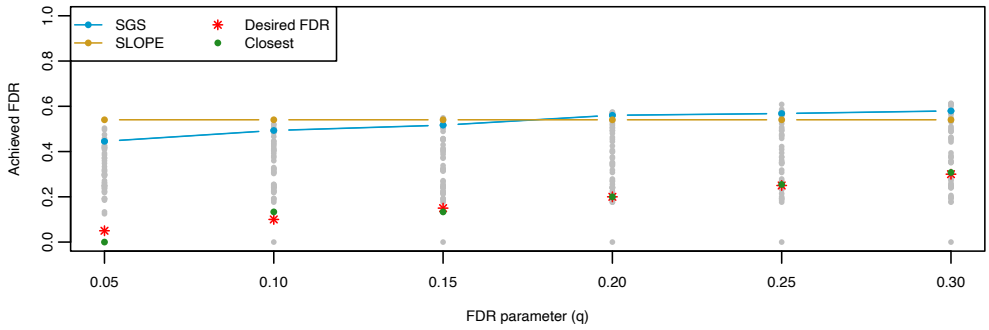



Figure 7.1: vFDR levels for SLOPE and SGS. The lines show the vFDR levels for the 1se CV models. Each grey dot represents the vFDR for a $\lambda$ value along the path of SGS models, with the green dots showing the value closest to the desired level. The design matrix $\mathbf{X} \sim \boldsymbol{\mathcal{N}}_p(\mathbf{0}, \boldsymbol{\Sigma})$ was used with $n = 100$, $p = 250$, $m = 50$, and $\rho = 0.3$.

However, the theory does not extend beyond this case, and so the model selection problem is acute for SLOPE models under general settings (where the noise level is unknown). Indeed, as illustrated by the synthetic study in Section 3.5, SLOPE models do not maintain FDR control under general designs when using CV for

tuning. The reason for this can be clearly seen in Figure 7.1. SLOPE and SGS were applied using CV to a synthetic dataset, varying the FDR parameter $q$ (for SGS, $q_v = q_g = q$). Along the path for SGS, models exist that achieve vFDR levels close to the desired level for most choices of $q$, but the chosen CV models had amongst the highest FDR levels. This suggests that CV is not a suitable method for model selection for SLOPE under general settings.

In this chapter, we consider the problem of model selection for SLOPE models under general settings, with a particular focus on a Bayesian perspective. To do this, we consider the linear regression setup from Section 2.1 from a probabilistic view, so that $\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2 \sim \boldsymbol{\mathcal{N}}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$.

### 7.1.1 Contributions and Organisation of this Chapter

In this chapter, we develop Bayesian models for group-based SLOPE models for the purpose of model selection under general settings. The models simultaneously learn the noise along with the other model parameters. To do this, the gSLOPE and SGS norms are incorporated into a spike-and-slab framework to form the Bayesian gSLOPE (BGSLOPE) (Section 7.3.1) and the Bayesian SGS (BSGS) (Section 7.3.2). The BSGS model introduces a novel sparse-group spike-and-slab framework that can be extended to other sparse-group formulations, such as the sparse-group lasso, as well as the wider class of OWL models (see Section 7.3.3). Our models inherit many of the advantages of Bayesian formulations, including uncertainty quantification and adaptivity to the underlying sparsity structure. They are also shown to be extremely stable to the choice of prior hyperparameters (Section 7.5).

To further investigate model selection in SLOPE models beyond our developed Bayesian methodology, we have also introduced a new two-step procedure (Two-step Orthogonal, Section 7.6.2) and extended existing approaches for group-based SLOPE models. These include the development of scaled regression methods (Sections 6.2.1 and 7.6.1) and the introduction of the first sparse-group Knockoffs framework in the literature (Section 6.2.2).

We compare our Bayesian methodology and a Bayesian SLOPE model (Section 7.2) with other existing approaches using both synthetic (Section 7.7) and real (Section 7.8) data studies. The studies provide an analysis of the key model selection strategies for SLOPE models, providing practitioners with practical insights for choosing an appropriate method tailored to a given problem. The primary metric of interest is FDR control, but we also assess the predictive performance of the

methods in our studies, since this is often of interest. The studies offer the first comprehensive comparison of all SLOPE model selection methods in the literature.

Our studies show that our Bayesian methodology achieves FDR control under the majority of general settings and obtains higher power than other competitive approaches. Additionally, on both the synthetic and real data, the Bayesian approaches achieved the strongest predictive performance. It is clear from our results that the best approach to applying SLOPE models under general settings is in a Bayesian framework. In addition to the strong performance of the Bayesian methods, TSO also demonstrated effective FDR control, achieved with very fast runtimes.

## 7.2 SLOPE in a Bayesian Context

SLOPE received a Bayesian treatment in Sepehri [173] to form the *Bayesian SLOPE* model, given by

$$\pi(\boldsymbol{\beta} \mid \sigma^2; \mathbf{v}) = Ce^{\frac{-1}{\sigma}\sum_{i=1}^{p} v_i|\beta|_{(i)}} = C\prod_{i=1}^{p} \exp\left\{-\frac{1}{\sigma} v_{r_v(\boldsymbol{\beta},i)}|\beta_i|\right\}, \tag{7.1}$$

where $C$ is a normalising constant. This prior was incorporated into the spike-and-slab framework in the *Adaptive Bayesian SLOPE* (ABSLOPE) [99]. ABSLOPE combines the benefits of SLOPE and spike-and-slab models: as SSL is non-separable and adaptive to unknown sparsity, there is a natural marriage between the spike-and-slab framework and SLOPE.

The ABSLOPE prior is given by

$$\pi\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, c, \sigma^2; \mathbf{v}\right) \propto c^{\sum_{i=1}^{p} \mathbf{1}(\gamma_i=1)} \prod_{i=1}^{p} \exp\left\{-a_i \left|\beta_i\right| \frac{1}{\sigma} v_{r_v(\mathbf{A}\boldsymbol{\beta},i)}\right\}, \tag{7.2}$$

where $\boldsymbol{\gamma} \in \{0,1\}^p$ is the variable inclusion vector and $\mathbf{A} = \text{diag}(a_1, \ldots, a_p)$ is a diagonal matrix with elements $a_i = c\gamma_i + (1-\gamma_i)$, where $c \in (0,1)$ represents the ratio of the average signal strengths between the active and non-active variables. The spike-and-slab effect comes via $c$ and can be seen more clearly by writing the

prior as

$$\pi\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, c, \sigma^2; \mathbf{v}\right) \propto c^{\sum_{i=1}^{p} \mathbf{1}(\gamma_i=1)} \prod_{i=1}^{p} \exp\left\{ -\frac{1}{\sigma}\left[ \underbrace{c\gamma_i \left|\beta_i\right| v_{r_v(\mathbf{A}\boldsymbol{\beta},i)}}_{\text{slab}} + \underbrace{(1-\gamma_i) \left|\beta_i\right| v_{r_v(\mathbf{A}\boldsymbol{\beta},i)}}_{\text{spike}} \right] \right\},$$

so that as $c < 1$, the signal (slab) variables are penalised less; this is the purpose of the matrix $\mathbf{A}$.

The ABSLOPE prior (Equation 7.2) is shown to be equivalent to the Bayesian SLOPE prior (Equation 7.1) under weighted design matrix $\mathbf{X}\mathbf{A}^{-1}$. As a result, the MAP under ABSLOPE can be obtained as the solution to frequentist SLOPE with the weighted design matrix [99]. ABSLOPE was developed to model missing data by also modelling $\mathbf{X}$. While this is not a problem of interest in this chapter, we note that our methodology readily extends to cover this case.

ABSLOPE is a hierarchical Bayesian model with priors

$$\pi(\boldsymbol{\gamma} \mid \theta) = \prod_{i=1}^{p} \theta^{\gamma_i}(1-\theta)^{1-\gamma_i}, \quad \theta \sim \text{Beta}(d_1, d_2), \quad c \sim \mathcal{U}[0,1], \quad \pi(\sigma^2) \propto \sigma^{-2},$$

where $\theta = \mathbb{P}(\gamma_j = 1 \mid \theta)$ represents the sparsity level of the coefficients, with small values leading to sparse models and $d_1, d_2 > 0$ defining the prior Beta distribution. Further, an improper uninformative prior is placed on $\sigma^2$ when the noise is unknown, which is our case of interest.

ABSLOPE is optimised using the SAEM algorithm, proposed by Lavielle [114] and presented in detail in Section 7.4.

## 7.3 Group-based Bayesian SLOPE Models

We extend the spike-and-slab framework (Section 6.3.2) to group-based SLOPE models to give *Bayesian Group SLOPE* (BGSLOPE) and *Bayesian Sparse-group SLOPE* (BSGS). Our models are designed to be consistent with the ABSLOPE framework and are also fit using the SAEM algorithm, described in Section 7.4.

There is currently limited literature on applying sparse-group models with spike-and-slab priors. As discussed in Section 6.1.3, Xu and Ghosh [217] developed the

Bayesian Group Lasso and Bayesian Sparse-Group Lasso using spike-and-slab priors. Our approaches differ in that we focus on group-based SLOPE models, which employ non-separable penalties specifically designed for FDR control. On the other hand, their models were based on lasso-type norms. Consequently, a key difference between the two approaches lies in the underlying model formulations. Our Bayesian methods inherit the useful properties of SLOPE models: non-separability, adaptivity, and the clustering property.

The two approaches share a hierarchical spike-and-slab formulation. However, Xu and Ghosh [217] employ point-mass spike-and-slab priors, whereas we use continuous variants that are less computationally prohibitive in high-dimensional settings [9]. In the sparse-group context, their Bayesian SGL and our BSGS models each utilise two sets of binary indicators to enable bi-level selection. Our formulation represents the first continuous spike-and-slab framework for sparse-group models.

The non-separability of SLOPE further distinguishes our models: coefficients are updated jointly, allowing information to be shared across variables [161], unlike the independent priors used by Xu and Ghosh [217], which update each variable separately. This aspect of their framework renders it incompatible with non-separable penalties (such as SLOPE), necessitating the development of new frameworks, as presented in this section for gSLOPE and SGS. From a computational standpoint, Xu and Ghosh [217] rely on Gibbs sampling, whereas we adopt an EM algorithm (Section 7.4). Other Bayesian sparse-group models include those proposed by Steiger and Vingron [181], Cai et al. [36], Chen et al. [41].

We develop our Bayesian methodology within the linear regression framework, as the SLOPE FDR properties (including the penalty formulation) are derived under this model. While the methodology can be extended to broader settings by modifying the likelihood term, linear regression remains the primary focus of this chapter.

The group-based SLOPE models are designed to provide FDR control through their weight choices, and allowing the noise level to be informed by this should provide better FDR control under general settings (in contrast to CV, where the noise level is estimated with regard to predictive performance).

Beyond the potential for improved FDR control, the Bayesian formulation provides additional benefits such as uncertainty quantification (in our case through the inclusion probabilities) and extensions to hierarchical models that offer flexibility in representing complex data structures. These hierarchical formulations facilitate

modelling of missing data, latent variables, and information sharing across groups, the last of which is particularly useful for our proposed models [76].

### 7.3.1 Bayesian gSLOPE (BGSLOPE)

This chapter presents the first Bayesian implementation for group SLOPE. To derive the Bayesian gSLOPE model, the penalty (Equation 2.4) is redefined as

$$\sum_{j=1}^{m} \sqrt{p_j} w_{r_g(\boldsymbol{\beta},j)} \|\boldsymbol{\beta}^{(j)}\|_2,$$

where $r_g(\boldsymbol{\beta}, g) \in [m]$ is the rank of $\boldsymbol{\beta}^{(j)}$ among the group elements (using the $\sqrt{p_j}\|\boldsymbol{\beta}^{(j)}\|_2$ ordering). gSLOPE can be seen as the posterior mode under the following Laplace prior

$$\pi(\boldsymbol{\beta} \mid \sigma^2; \mathbf{w}) = C \prod_{j=1}^{m} \exp\left\{ -\frac{1}{\sigma} w_{r_g(\boldsymbol{\beta},j)} \sqrt{p_j} \|\boldsymbol{\beta}^{(j)}\|_2 \right\}, \tag{7.3}$$

with the normalising constant $C$ given in Theorem E.1.

Extending the ABSLOPE framework to gSLOPE, the *Bayesian Group SLOPE* (BGSLOPE) is defined as the hierarchical model

$$\pi\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, c, \sigma^2; \mathbf{w}\right) \propto c^{\sum_{j=1}^{m} p_j \mathbf{1}(\gamma_j = 1)} \prod_{j=1}^{m} \exp\left\{ -\frac{1}{\sigma} \tilde{a}_j \|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta},j)} \right\}, \tag{7.4}$$

$$\pi(\boldsymbol{\gamma} \mid \theta) = \prod_{j=1}^{m} \theta^{\gamma_j} (1-\theta)^{1-\gamma_j}, \quad \theta \sim \text{Beta}(d_1, d_2), \quad \pi(\sigma^2) \propto \sigma^{-2}, \quad c \sim \mathcal{U}[0,1],$$

where

- $\boldsymbol{\gamma} \in \{0,1\}^m$ is the group inclusion vector under an independent Bernoulli product prior.
- $\theta$ is a mixture parameter representing the sparsity of the fitted groups (with $d_1, d_2 > 0$ prior hyperparameters). As smaller values of $\theta$ correspond to sparser underlying models, setting $d_1 \ll d_2$ places most of the prior mass near zero, thereby encouraging sparsity in the model.
- $c$ ensures the signal groups receive less penalisation (also seen as the signal strength ratio between active and non-active groups).

- The weight matrix is given by $\tilde{\mathbf{A}} = \text{diag}(\tilde{a}_1, \ldots, \tilde{a}_m) \in \mathbb{R}^{p \times p}$ with entries $\tilde{A}_{ii} = \tilde{a}_j, \forall i \in G_j$, where $\tilde{a}_j = c\gamma_j + (1 - \gamma_j)$. Variables in the same group have the same entry.

The BGSLOPE dependency graph is shown in Figure 7.2. BGSLOPE, through the inclusion parameter, $\boldsymbol{\gamma}$, provides measures of both uncertainty quantification and group-level importance, aspects that are not available in the frequentist gSLOPE. Additionally, as the parameter $\theta$ governs the degree of shrinkage applied to each group, it allows BGSLOPE to adapt to varying levels of group sparsity in the data.

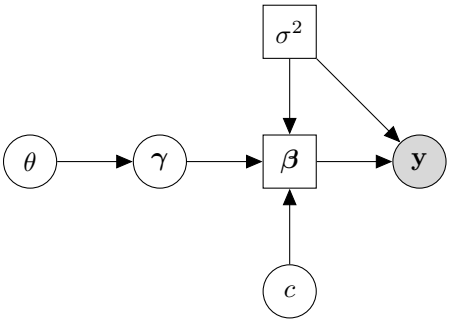



Figure 7.2: BGSLOPE dependency graph.

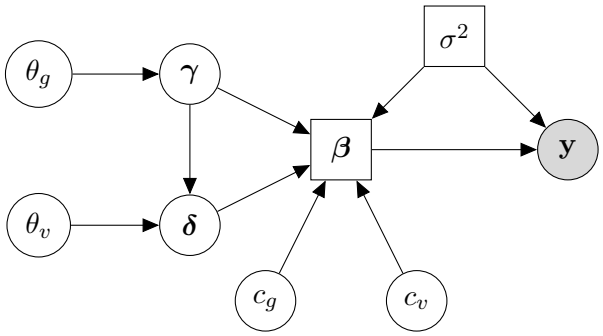



Figure 7.3: BSGS dependency graph.

Proposition 7.1 demonstrates that the MAP estimate under the BGSLOPE prior is equivalent to the solution under the gSLOPE Laplace prior, when using the matrix $\mathbf{X}\tilde{\mathbf{A}}^{-1}$ as input. By extension, this also makes it equivalent to the frequentist gSLOPE formulation, after reparameterisation. The proof is provided in Appendix E.1.1.

**Proposition 7.1.** *If the coefficients* $\mathbf{z} = (z_1, \ldots, z_p)$ *follow the gSLOPE Laplace prior (Equation 7.3), then* $\boldsymbol{\beta} = \tilde{\mathbf{A}}^{-1}\mathbf{z}$ *follows the BGSLOPE prior (Equation 7.4).*

### 7.3.2 Bayesian SGS (BSGS)

SGS (Equation 3.1) applies two levels of penalisation and sorting to produce bi-level selection, which makes its Bayesian formulation more involved. It has a Laplace prior given by

$$\pi(\boldsymbol{\beta} \mid \sigma^2; \mathbf{w}, \mathbf{v}) \propto \exp\left\{ -\frac{1}{\sigma}\alpha \sum_{i=1}^{p} v_{r_v(\boldsymbol{\beta},i)} |\beta_i| - \frac{1}{\sigma}(1-\alpha) \sum_{j=1}^{m} w_{r_g(\boldsymbol{\beta},j)} \sqrt{p_j} \|\boldsymbol{\beta}^{(j)}\|_2 \right\}, \tag{7.5}$$

where, as for ABSLOPE and BGSLOPE, the sorting occurs through the weights with rank functions $r_v, r_g$.

As with BGSLOPE, the development of the Bayesian SGS model follows the AB-SLOPE (continuous spike-and-slab) framework. The *Bayesian Sparse-group SLOPE* (BSGS) is given by the hierarchical model

$$\begin{aligned}
\pi\left(\boldsymbol{\beta} \mid \boldsymbol{\delta}, \boldsymbol{\gamma}, c_g, c_v, \sigma^2; \mathbf{w}, \mathbf{v}\right) \propto & \, c_g^{\sum_{j=1}^m p_j \mathbf{1}(\gamma_j=1)} c_v^{\sum_{i=1}^p \mathbf{1}(\delta_i=1)} \\
& \times \prod_{i=1}^p \exp\left\{-\frac{1}{\sigma}|\beta_i|\hat{a}_i \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\} \\
& \times \prod_{j=1}^m \exp\left\{-\frac{1}{\sigma}\|\hat{\mathbf{A}}^{(j)}\beta^{(j)}\|_2 \sqrt{p_j}(1-\alpha) w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)}\right\}, \\
\pi(\boldsymbol{\gamma} \mid \theta_g) = & \prod_{j=1}^m \theta_g^{\gamma_j}(1-\theta_g)^{1-\gamma_j}, \\
\pi(\boldsymbol{\delta} \mid \boldsymbol{\gamma}, \theta_v) = & \prod_{j=1}^m \prod_{i \in G_j} \left[\theta_v^{\delta_i}(1-\theta_v)^{1-\delta_i}\mathbf{1}(\gamma_j=1) + \mathbf{1}(\gamma_j=0, \delta_i=0)\right], \\
\theta_g \sim \text{Beta}(d_1, d_2), \quad & \theta_v \sim \text{Beta}(e_1, e_2), \quad \pi(\sigma^2) \propto \sigma^{-2}, \quad c_g, c_v \sim \mathcal{U}[0,1].
\end{aligned} \tag{7.6}$$

where

- $\boldsymbol{\gamma} \in \{0,1\}^m, \boldsymbol{\delta} \in \{0,1\}^p$ are the group/variable inclusion vectors.
- $\theta_g, \theta_v \in (0,1)$ are the group/variable mixture proportions (with $d_1, d_2, e_1, e_2 > 0$ prior hyperparameters). As with BGSLOPE, choosing $d_1 \ll d_2$ and $e_1 \ll e_2$ leads to sparser models.
- $c_g, c_v \in (0,1)$ allow the signal group/variables to be penalised less.
- The weight matrix is denoted as $\hat{\mathbf{A}} = \text{diag}(\hat{a}_1, \ldots, \hat{a}_p)$ with elements, for $i \in G_j$,

$$\hat{a}_i = c_g\gamma_j(c_v\delta_i + 1 - \delta_i) + (1-\gamma_j) = \begin{cases} c_g c_v & \text{if } \gamma_j = 1 \text{ and } \delta_i = 1, \\ c_g & \text{if } \gamma_j = 1 \text{ and } \delta_i = 0, \\ 1 & \text{if } \gamma_j = 0. \end{cases}$$

The weight matrix penalises the active groups less via $c_g$ and the active variables less by $c_g c_v$. Additionally, the hierarchical construction of $\boldsymbol{\delta}$ being dependent on $\boldsymbol{\gamma}$

enforces all variables in a non-active group to also be non-active. To see this, note that if $\gamma_j = 1$, the prior on $\boldsymbol{\delta}$ is an independent Bernoulli product (as for ABSLOPE and BGSLOPE) and if $\gamma_j = 0$, the prior is a point-mass at zero.

BSGS provides uncertainty quantification and feature importance at two levels, group and variable, through the inclusion parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, respectively. Moreover, the dual mixing proportions $\theta_g$ and $\theta_v$ enable BSGS to adapt simultaneously to the underlying sparsity structures at both levels.

The BSGS dependency graph is shown in Figure 7.3. Proposition 7.2 demonstrates that the posterior mode from the BSGS prior is equivalent to the estimates obtained via the SGS Laplace prior under the input $\mathbf{X}\hat{\mathbf{A}}^{-1}$. This also makes it equivalent to the frequentist model. The proof is given in Appendix E.1.2.

**Proposition 7.2.** *If the coefficients* $\mathbf{z} = (z_1, \dots, z_p)$ *follow the SGS Laplace prior (Equation 7.5), then* $\boldsymbol{\beta} = \hat{\mathbf{A}}^{-1}\mathbf{z}$ *follows the BSGS prior (Equation 7.6).*

*Remark* 7.1. It is possible to scale the signal strength ratio using $\alpha$ and replace

$$c_g^{\sum_{j=1}^m p_j \mathbf{1}(\gamma_j=1)} c_v^{\sum_{i=1}^p \mathbf{1}(\delta_i=1)} \quad \to \quad c_g^{(1-\alpha)\sum_{j=1}^m p_j \mathbf{1}(\gamma_j=1)} c_v^{\alpha \sum_{i=1}^p \mathbf{1}(\delta_i=1)}$$

in Equation 7.6. This scaling allows the hyperparameters to have a direct interpretational connection to ABSLOPE and BGSLOPE. However, as this adds complexity and interpretation between models is not a key priority, this is not pursued further.

### 7.3.3 Extensions

Our methodology is developed for SLOPE models but easily extends to additional models. Both BGSLOPE and BSGS can be applied to the wider class of OWL models by replacing the corresponding penalties. The framework developed for BSGS applies to any sparse-group model, including SGL. To incorporate SGL, DFR can be used for screening during SAEM optimisation (Chapter 5), although this is beyond the scope of this chapter.

## 7.4 Stochastic Approximation EM (SAEM) Algorithm

Both BGSLOPE and BSGS are fitted using SAEM; we describe it here in the context of BGSLOPE. We wish to estimate $\Theta = (\boldsymbol{\beta}, \sigma)$ in the presence of latent variables $\Lambda = (\boldsymbol{\gamma}, c, \theta)$. Many spike-and-slab models employ *expectation–maximization* (EM)

algorithms to achieve this objective; for example, the SSL model is fitted via the EMVS algorithm [160].

There are two steps to an EM algorithm (for step $t$):

1. **Expectation step (E step)**: Calculate the expected value of the log-likelihood,

$$Q(\Theta \mid \Theta_{[t]}) = \mathbb{E}_{\Lambda \sim \pi(\cdot \mid \mathbf{y}, \Theta_{[t]})} \left[ \log \pi(\mathbf{y}, \Lambda \mid \Theta_{[t]}) \right],$$

   where $\Theta_{[t]}$ denotes the value of the parameters at step $t$.

2. **Maximisation step (M step)**: Find the parameters that maximise $Q$,

$$\Theta_{[t+1]} = \arg\max_{\Theta} Q(\Theta \mid \Theta_{[t]}).$$

However, for SLOPE models, $Q$ is not tractable due to the sorting procedures. Instead, the *Stochastic Approximation EM* (SAEM) algorithm is used, in which the E step is replaced by a simulation step, followed by a stochastic approximation step [114]. That is, the E step above is replaced by

- *Simulation step*: Sample the latent variables $\Lambda$ from $\pi(\Lambda \mid \mathbf{y}, \Theta_{[t-1]})$ using a Gibbs sampler.

- *Stochastic approximation*: Calculate $Q$ as

$$Q(\Theta \mid \Theta_{[t]}) = Q(\Theta \mid \Theta_{[t-1]}) + \eta_t \left( \log \pi(\mathbf{y}, \Lambda \mid \Theta_{[t]}) - Q(\Theta \mid \Theta_{[t-1]}) \right),$$

  where $\eta_t$ is the step size, which is chosen as a decreasing sequence $\eta_t = 1/t$ that guarantees almost sure convergence [50].

SAEM estimates $\boldsymbol{\beta}$ and simulates $\boldsymbol{\gamma}$ to allow for selection. It has two loosely defined stages: first, a quick search to find the solution neighbourhood, followed by a steady fine-tuning stage; these steps can be seen in the illustrative example in Section 7.4.5. SAEM is generally not sensitive to initialisation, as demonstrated in the study presented in Section 7.5, which is in agreement with findings in the literature [114]. The stochastic approximation is applied only after the 20th iteration, so that $\eta_t = 1$ if $t \leq 20$.

To determine the final set of active variables/groups, median thresholding is applied to the inclusion parameters over the last $T = 20$ iterations of SAEM: a

variable $i \in [p]$ is active if $\mathbb{P}(\delta_i = 1 \mid \mathbf{y}) \approx \frac{1}{T}\sum_{t\in\mathcal{T}}(\delta_{[t]})_i > 0.5$ and a group $j \in [m]$ is active if $\mathbb{P}(\gamma_j = 1 \mid \mathbf{y}) \approx \frac{1}{T}\sum_{t\in\mathcal{T}}(\gamma_{[t]})_j > 0.5$, where $\mathcal{T}$ denotes the index set of the last $T$ iterations. In BSGS, group thresholding is not necessary due to the hierarchical structure between $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$.

SAEM is described fully for BGSLOPE in Algorithm 5 and for BSGS in Algorithm 9. The SAEM updates for BGSLOPE and BSGS are derived next. This is followed by the development of the BSGS-$\alpha$ model, which additionally updates the $\alpha$ parameter, and the SLOBE models, which approximate the conditional distribution by its conditional expectation to accelerate the algorithm.

**Algorithm 5** SAEM for BGSLOPE

**Input:** Initial values $(\boldsymbol{\beta}_{[0]}, \sigma_{[0]}, c_{[0]}, \theta_{[0]})$, maximum number of iterations $T > 0$, tolerance $\epsilon > 0$
**while** $t \leq T$ **and** $\|\boldsymbol{\beta}_{[t+1]} - \boldsymbol{\beta}_{[t]}\|_2^2 > \epsilon$ **do**
  **Simulation step:**
  Sample $(\gamma_j)_{[t]}$ from Equation 7.8 for $j = 1, \ldots, m$
  Sample $\theta_{[t]}$ from Equation 7.9
  Sample $c_{[t]}$ from Equation 7.10
  **Stochastic approximation step:**
  Calculate $\boldsymbol{\beta}_{[t]}^{\text{MLE}}$ via Equation 7.12
  Calculate $\sigma_{[t]}^{\text{MLE}}$ from Equation 7.13
  Update parameters

$$\boldsymbol{\beta}_{[t+1]} = \boldsymbol{\beta}_{[t]} + \eta_t(\boldsymbol{\beta}_{[t]}^{\text{MLE}} - \boldsymbol{\beta}_{[t]}), \quad \sigma_{[t+1]} = \sigma_{[t]} + \eta_t(\sigma_{[t]}^{\text{MLE}} - \sigma_{[t]})$$

  Update step size $\eta_t = \begin{cases} 1, & \text{if } t \leq 20, \\ \frac{1}{t-20}, & \text{if } t > 20 \end{cases}$
  $t = t + 1$
**end while**
**Output:** $(\boldsymbol{\beta}_{[t+1]}, \boldsymbol{\gamma}_{[t]})$, such that $\hat{\boldsymbol{\gamma}}_{[t]} \leftarrow \frac{1}{20}\sum_{k=t-19}^{t}\boldsymbol{\gamma}_{[k]}$ (if $t < 20$ all iterations of $\boldsymbol{\gamma}$ are used), giving the fitted coefficients $\hat{\boldsymbol{\beta}}_{[t+1]} = \boldsymbol{\beta}_{[t+1]} \cdot \mathbf{1}\left(\hat{\boldsymbol{\gamma}}_{[t]} > 0.5\right)$

### 7.4.1 Updates for BGSLOPE

For BGSLOPE, the penalised log-likelihood, which is used to derive the maximisation steps, is given by,

$$\begin{aligned}\log \pi(\mathbf{y}, \Lambda \mid \Theta_{[t]}) &= \log \pi\left(\mathbf{y}, \boldsymbol{\gamma}, c; \boldsymbol{\beta}, \theta, \sigma^2\right) + \log \pi\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, c, \sigma^2; \mathbf{w}\right) \\ &= \log \pi\left(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\beta}, \sigma^2\right) + \log \pi(\boldsymbol{\gamma} \mid \theta) + \log \pi(\sigma^2) \\ &\quad + \log \pi\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, c, \sigma^2; \mathbf{w}\right) \\ &= -(n+2)\log\sigma - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1)\log\theta \\ &\quad + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 0)\log(1-\theta) + \log c \sum_{j=1}^{m} p_j \mathbf{1}(\gamma_j = 1) \\ &\quad - \frac{1}{\sigma}\sum_{j=1}^{m} \tilde{a}_j \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)} \|\boldsymbol{\beta}^{(j)}\|_2,\end{aligned} \tag{7.7}$$

where the $c$ log prior term is omitted since it is a standard uniform.

#### Simulation Step

The following Gibbs sampler regime is used for the simulation step, for $j \in [m]$,

$$\gamma_j \sim \text{Bernoulli}\left(\frac{L_1}{L_1 + L_2}\right), \text{ where,} \tag{7.8}$$

$$L_1 = \theta c^{p_j} \exp\left\{\frac{-c}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)}\right\},$$

$$L_2 = (1-\theta) \exp\left\{\frac{-1}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)}\right\},$$

$$\theta \sim \text{Beta}\left(d_1 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1), d_2 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 0)\right), \tag{7.9}$$

$$c \sim \text{Gamma}\left(1 + \sum_{j=1}^{m} p_j 1(\gamma_j = 1), \frac{1}{\sigma}\sum_{j=1}^{m} \|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)} \mathbf{1}(\gamma_j = 1)\right), \tag{7.10}$$

truncated at $[0, 1]$.

The derivations of these are given in Appendix E.2.1.

### Stochastic Approximation and Maximisation Steps

After simulating the latent variables, we use these to complete the maximisation step and compute $\Theta_{[t+1]}$. For a general step size $\eta_t$, we update the model parameters as

$$\boldsymbol{\beta}_{[t+1]} = \boldsymbol{\beta}_{[t]} + \eta_t(\boldsymbol{\beta}_{[t]}^{\text{MLE}} - \boldsymbol{\beta}_{[t]}), \quad \sigma_{[t+1]} = \sigma_{[t]} + \eta_t(\sigma_{[t]}^{\text{MLE}} - \sigma_{[t]}), \tag{7.11}$$

where $\boldsymbol{\beta}_{[t]}^{\text{MLE}}$ and $\sigma_{[t]}^{\text{MLE}}$ are the *Maximum Likelihood Estimation* (MLE) estimators of the log-likelihood (Equation 7.7), using samples of the latent variables at step $t$. Considering only the terms involving the parameters of interest in the log-likelihood, they are given by

$$\boldsymbol{\beta}_{[t]}^{\text{MLE}} = \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \frac{\sigma_{[t-1]}}{n} \sum_{j=1}^{m} (\tilde{a}_j)_{[t]} \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}_{[t]}\mathbf{b},j)} \|\mathbf{b}^{(j)}\|_2 \right\}, \tag{7.12}$$

$$\sigma_{[t]}^{\text{MLE}} = \frac{K_2 + \sqrt{K_2^2 + 4K_1(n+2)}}{2(n+2)}, \tag{7.13}$$

$$\text{where } K_1 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{[t]}\|_2^2, \; K_2 = \sum_{j=1}^{m} (\tilde{a}_j)_{[t]} \|\boldsymbol{\beta}_{[t]}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]},j)}.$$

See Appendix E.2.2 for the derivation of the $\sigma^{\text{MLE}}$ update and the remark below for the $\boldsymbol{\beta}^{\text{MLE}}$ update.

*Remark* 7.2. To calculate the update for $\boldsymbol{\beta}^{\text{MLE}}$, consider the transformation $\mathbf{z} = \tilde{\mathbf{A}}\boldsymbol{\beta}^{\text{MLE}}$, so that Equation 7.12 becomes

$$\mathbf{z}_{[t]} = \underset{\mathbf{z}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\tilde{\mathbf{A}}_{[t]}^{-1}\mathbf{z}\|_2^2 + \frac{\sigma_{[t-1]}}{n} \sum_{j=1}^{m} \sqrt{p_j} w_{r_g(\mathbf{z},j)} \|\mathbf{z}^{(j)}\|_2 \right\}.$$

This derivation follows essentially the same steps as in Proposition 7.1. Consequently, the update reduces to the standard gSLOPE problem with input matrix $\mathbf{X}\tilde{\mathbf{A}}_{[t]}^{-1}$ and regularisation parameter $\lambda = \sigma_{[t-1]}/n$ to obtain $\mathbf{z}_{[t]}$, and the MLE is then computed as $\boldsymbol{\beta}_{[t]}^{\text{MLE}} = \tilde{\mathbf{A}}_{[t]}^{-1}\mathbf{z}_{[t]}$. Since $\tilde{\mathbf{A}}$ is diagonal, computing the inverse is very inexpensive. This transformation also resolves any issues with sorting the modified penalty sequence as the sorting occurs via $\mathbf{z}$. The optimisation for $\mathbf{z}$ is carried out using ATOS, employing the basic strong rule for gSLOPE developed in Chapter 4.

### 7.4.2 Updates for BSGS

For BSGS, we define the latent variables as $\Lambda' = (\boldsymbol{\gamma}, \boldsymbol{\delta}, c_g, c_v, \theta_g, \theta_v)$ and the model parameters are as for BGSLOPE, $\Theta = (\boldsymbol{\beta}, \sigma)$, giving the penalised log-likelihood,

$$\begin{aligned}
\log \pi(\mathbf{y}, \Lambda' \mid \Theta_{[t]}) &= \log \pi\left(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\beta}, \sigma^2\right) + \log \pi(\boldsymbol{\gamma} \mid \theta_g) + \log \pi(\boldsymbol{\delta} \mid \boldsymbol{\gamma}, \theta_v) + \log \pi(\sigma^2) \\
&\quad + \log \pi\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, c_g, c_v, \sigma^2; \mathbf{w}, \mathbf{v}\right) \qquad (7.14) \\
&= -(n+2)\log\sigma - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1)\log\theta_g + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 0)\log(1-\theta_g) \\
&\quad + \sum_{j=1}^{m}\sum_{i \in G_j} \log\left[\mathbf{1}(\gamma_j = 1)\left(\theta_v^{\delta_i}(1-\theta_v)^{1-\delta_i}\right) + \mathbf{1}(\gamma_j = 0, \delta_i = 0)\right] \\
&\quad + \log c_g \sum_{j=1}^{m} p_j \mathbf{1}(\gamma_j = 1) + \log c_v \sum_{i=1}^{p} \mathbf{1}(\delta_i = 1) \\
&\quad - \sum_{i=1}^{p} \frac{1}{\sigma}|\beta_i|\hat{a}_i v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)} - \sum_{j=1}^{m} \frac{1}{\sigma}\|\hat{\mathbf{A}}^{(j)}\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)}.
\end{aligned}$$

#### Simulation Step

A Gibbs sampler is used for the simulation step (the derivations of these updates are provided in Appendix E.2.1). For the inclusion parameters, we sample from Bernoulli distributions

$$\gamma_j \sim \text{Bernoulli}\left(\frac{L_1'}{L_1' + L_2'}\right), \; j \in [m], \text{ where,} \qquad (7.15)$$

$$\begin{aligned}
L_1' &= \theta_g c_g^{p_j} c_v^{\sum_{i \in G_j} \mathbf{1}\{\delta_i = 1\}} \exp\left\{-\frac{c_g}{\sigma}\|\boldsymbol{\kappa}^{(j)}\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j}(1-\alpha) w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)}\right\} \qquad (7.16) \\
&\quad \times \prod_{i \in G_j} \theta_v^{\delta_i}(1-\theta_v)^{1-\delta_i} \exp\left\{-\frac{c_g}{\sigma}|\beta_i| c_v^{\mathbf{1}(\delta_i = 1)} \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\}, \\
L_2' &= (1-\theta_g)\exp\left\{-\frac{1}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j}(1-\alpha) w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)}\right\} \qquad (7.17) \\
&\quad \times \prod_{i \in G_j} \exp\left\{-\frac{1}{\sigma}|\beta_i| \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\},
\end{aligned}$$

$$\delta_i \sim \text{Bernoulli}\left(\frac{\tilde{L}_1}{\tilde{L}_1 + \tilde{L}_2}\right), \; i \in [p], \text{ where,} \qquad (7.18)$$

$$\tilde{L}_1 = \theta_v c_v \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g c_v \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\}, \; \tilde{L}_2 = (1-\theta_v)\exp\left\{-\frac{1}{\sigma}|\beta_i| c_g \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\}.$$

The mixing parameters are sampled via Beta distributions, given by

$$\theta_g \sim \text{Beta}\left(d_1 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1), d_2 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 0)\right), \tag{7.19}$$

$$\theta_v \sim \text{Beta}\Bigg(me_1 - m + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1) \sum_{i \in G_j} \mathbf{1}(\delta_i = 1) + 1, \\ me_2 - m + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1) \sum_{i \in G_j} \mathbf{1}(\delta_i = 0) + 1\Bigg). \tag{7.20}$$

Finally, the signal strength ratios are sampled as

$$c_g \sim \text{Gamma}\Bigg(1 + \sum_{j=1}^{m} p_j \mathbf{1}(\gamma_j = 1), \tag{7.21}$$

$$\frac{1}{\sigma}\left[\sum_{i=1}^{p} c_v |\beta_i| \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)} \mathbf{1}(\delta_i = 1) + \sum_{j=1}^{m} \|\boldsymbol{\kappa}^{(j)} \boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} (1-\alpha) w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)} \mathbf{1}(\gamma_j = 1)\right]\Bigg),$$

truncated at $[0, 1]$,

$c_v$ sampled via Metropolis-Hastings (MH) with proposal Gamma$(2, 2)$, (7.22)

where $\boldsymbol{\kappa}^{(j)} = \text{diag}(c_v^{\mathbf{1}(\delta_i = 1)}), \forall i \in G_j$.

### Stochastic Approximation and Maximisation Steps

After the simulation step, the model parameters are given by Equation 7.11 with the MLE estimators

$$\boldsymbol{\beta}_{[t]}^{\text{MLE}} = \underset{\mathbf{b} \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \frac{\sigma_{[t-1]}}{n} \Bigg( \alpha \sum_{i=1}^{p} |b_i| (\hat{a}_i)_{[t]} v_{r_v(\hat{\mathbf{A}}_{[t]}\mathbf{b}, i)} \\ + (1-\alpha) \sum_{j=1}^{m} \|\hat{\mathbf{A}}_{[t]}^{(j)} \mathbf{b}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}_{[t]}\mathbf{b}, j)} \Bigg) \right\}, \tag{7.23}$$

$$\sigma_{[t]}^{\text{MLE}} = \frac{K_2' + \sqrt{(K_2')^2 + 4K_1'(n+2)}}{2(n+2)}, \text{ where,} \tag{7.24}$$

$$K_1' = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{[t]}\|_2^2,$$

$$K_2' = \alpha \sum_{i=1}^{p} |(\beta_i)_{[t]}| (\hat{a}_i)_{[t]} v_{r_v(\hat{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]}, i)} + (1-\alpha) \sum_{j=1}^{m} \|\hat{\mathbf{A}}_{[t]}^{(j)} \boldsymbol{\beta}_{[t]}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]}, j)}.$$

The $\boldsymbol{\beta}$ update is the SGS optimisation task. As with BGSLOPE, this is calculated via a transformation with input $\mathbf{X}\hat{\mathbf{A}}_{[t]}^{-1}$ and regularisation parameter $\lambda = \sigma_{[t-1]}/n$; see Remark 7.2. It is performed using ATOS with the basic versions of the SGS screening rules from Chapter 4. See Appendix E.3.2 for the derivation of the $\sigma^{\text{MLE}}$ update.

### 7.4.3 BSGS-$\alpha$

It is also possible to learn the $\alpha$ model parameter for BSGS, which defines the balance between the two types of penalisation, by placing a uniform prior on it, $\alpha \sim \mathcal{U}[0, 1]$. The posterior for $\alpha$ is not available in closed-form, so it is sampled via MH using a Beta$(10, 0.5)$ proposal, which skews $\alpha$ towards 1 (see Appendix E.3.3). We denote this model by BSGS-$\alpha$. To the best of our knowledge, there are no other instances in the literature of $\alpha$ being updated in a Bayesian model.

### 7.4.4 SLOBE Models

The SAEM algorithm for BGSLOPE and BSGS can be accelerated by approximating samples from the conditional distributions of the latent parameters using their conditional expectations. That is,

$$\Lambda_{[t]} \leftarrow \mathbb{E}[\Lambda_{[t]} \mid \mathbf{y}, \Theta_{[t-1]}].$$

Besides speeding up computation, this reduces latent variable variability, which helps diminish algorithmic noise in high dimensions.

ABSLOPE was modified in this way to form SLOBE in Jiang et al. [99]. Here, we apply the SLOBE acceleration to BGSLOPE and BSGS, to form GSLOBE and SGSLOBE, respectively. The SAEM procedures for GSLOBE and SGSLOBE are as in Algorithms 5 and 9, with the simulation step replaced by the conditional expectation approximations.

#### GSLOBE

Taking the conditional expectations of the simulation updates for BGSLOPE (Equations 7.8–7.10):

1. Approximate each $\gamma_j, j \in [m]$, by the active probability of a Bernoulli distri-

bution, where $\boldsymbol{\gamma_{-j}}$ is the vector $\boldsymbol{\gamma}$ with the $j$th entry removed,

$$\mathbb{E}[\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}, c, \boldsymbol{\beta}, \sigma, \theta, \tilde{\mathbf{A}}]$$
$$= \frac{\theta c^{p_j} \exp\left\{\frac{-c}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)}\right\}}{(1-\theta)\exp\left\{\frac{-1}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)}\right\} + \theta c^{p_j} \exp\left\{\frac{-c}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)}\right\}}.$$

2. Approximate $\theta$ by the mean of a Beta distribution

$$\mathbb{E}[\theta \mid \boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\beta}, \sigma, c, \tilde{\mathbf{A}}] = \mathbb{E}[\theta \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma, \tilde{\mathbf{A}}] = \frac{d_1 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1)}{d_1 + d_2 + m}.$$

3. Approximate $c$ by the mean of a truncated Gamma distribution

$$\mathbb{E}[c \mid \boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\beta}, \sigma, \theta, \tilde{\mathbf{A}}] = \frac{\int_0^1 z^x \exp\{-x'z\}\, dz}{\int_0^1 z^{x-1} \exp\{-x'z\}\, dz}, \quad \text{where,}$$
$$x = 1 + \sum_{j=1}^{m} p_j \mathbf{1}(\gamma_j = 1),\ x' = \frac{1}{\sigma} \sum_{j=1}^{m} \|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r(\tilde{\mathbf{A}}\boldsymbol{\beta}, g)} \mathbf{1}(\gamma_j = 1).$$

**SGSLOBE**

Taking the conditional expectations of the simulation updates for BSGS (Equations 7.15–7.21; the update for $c_v$ remains the same via MH):

1. Approximate the inclusion probabilities by the active probability of a Bernoulli distribution (where $L_1'$ and $L_2'$ are as described in Equations 7.16 and 7.17), for each $j \in [m]$ and $i \in [p]$,

$$\mathbb{E}[\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}, c_g, c_v, \boldsymbol{\beta}, \sigma, \theta_g, \theta_v, \boldsymbol{\delta}, \hat{\mathbf{A}}] = \frac{L_1'}{L_1' + L_2'},$$
$$\mathbb{E}[\delta_i = 1 \mid \boldsymbol{\delta}_{-i}, c_g, c_v, \boldsymbol{\beta}, \sigma, \theta_g, \theta_v, \boldsymbol{\gamma}, \hat{\mathbf{A}}]$$
$$= \left( \frac{\theta_v c_v \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g c_v \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\}}{\theta_v c_v \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g c_v \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\} + (1-\theta_v)\exp\left\{-\frac{1}{\sigma}|\beta_i| c_g \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\}} \right).$$

2. Approximate $\theta_g$ and $\theta_v$ by the means of a Beta distribution

$$\mathbb{E}[\theta_g \mid \boldsymbol{\gamma}, \mathbf{y}, \theta_v, \boldsymbol{\beta}, \sigma, c_v, c_g, \boldsymbol{\delta}, \hat{\mathbf{A}}] = \mathbb{E}[\theta_g \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma, \hat{\mathbf{A}}, \boldsymbol{\delta}] = \frac{d_1 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1)}{d_1 + d_2 + m},$$

$$\mathbb{E}[\theta_v \mid \boldsymbol{\gamma}, \mathbf{y}, \theta_g, \boldsymbol{\beta}, \sigma, c_v, c_g, \boldsymbol{\delta}, \hat{\mathbf{A}}]$$

$$= \mathbb{E}[\theta_v \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma, \hat{\mathbf{A}}, \boldsymbol{\delta}] = \frac{me_1 - m + x + 1}{me_1 + x + 2 + me_2 - 2m + x'}, \text{ where,}$$

$$x = \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1) \sum_{i \in G_j} \mathbf{1}(\delta_i = 1), \; x' = \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1) \sum_{i \in G_j} \mathbf{1}(\delta_i = 0).$$

3. Approximate $c_g$ by the mean of a truncated Gamma distribution

$$\mathbb{E}[c_g \mid \boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\beta}, \sigma, \theta, \hat{\mathbf{A}}] = \frac{\int_0^1 z^x \exp\{-x'z\}\, dz}{\int_0^1 z^{x-1} \exp\{-x'z\}\, dz}, \quad \text{where,}$$

$$x = 1 + \sum_{j=1}^{m} p_j \mathbf{1}(\gamma_j = 1),$$

$$x' = \frac{1}{\sigma}\left[\sum_{i=1}^{p} c_v |\beta_i| v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)} \mathbf{1}(\delta_i = 1) + \sum_{j=1}^{m} \|\boldsymbol{\kappa}^{(j)} \boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j}\, w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)} \mathbf{1}(\gamma_j = 1)\right].$$

### 7.4.5 Illustrative Example

BSGS and BGSLOPE are applied to a toy example to demonstrate the fitting process, with data generated as per the baseline setup from Section 7.7.1: a block-correlated multivariate Gaussian design is used to generate linear responses with group-structured sparsity.

BGSLOPE and BSGS were applied with the initial $\boldsymbol{\beta}$ values generated from a lasso model and prior hyperparameters for the $\theta$ distributions given by $d_1 = d_2 = 0.1n = 4$ for BGSLOPE and $d_1 = e_1 = 0.003n = 1.2, d_2 = e_2 = 0.015n = 6$ for BSGS (these initialisations are explored in Section 7.5). BSGS used $\alpha = 0.95$ and other hyperparameters were set as described in Section 7.7.2. Figure 7.4 shows that the parameters very quickly enter into a period of stability. The reduced variability of the SLOBE models is highlighted well in the plot for $\theta_g$, where the conditional mean is clearly seen. BSGS and BGSLOPE appear equally accurate in recovering the true parameter values. Figure E.2 displays the likelihood traces for each model, which shows the SLOBE models achieve almost identical likelihood values.

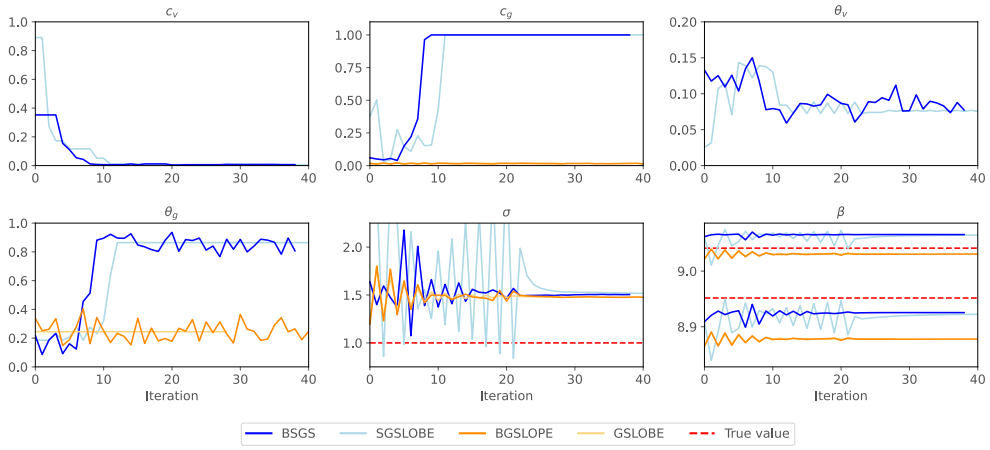



Figure 7.4: Bayesian latent variables and model parameters for BSGS, SGSLOBE, BGSLOPE, and GSLOBE, applied to the toy example, shown for the first 40 iterations. The true values for $\sigma, \boldsymbol{\beta}$ are shown in red. Two active $\boldsymbol{\beta}$ values are shown, and the $c, \theta$ values for BGSLOPE and GSLOBE are placed into the $c_g, \theta_g$ plots. BSGS converged in 39 iterations, SGSLOBE in 37, BGSLOPE in 145, and GSLOBE in 29.

SAEM generally has two phases, as described in Lavielle [114]. The first phase attempts to find a neighbourhood in only a few iterations. This can be seen around iteration 8 for BSGS by the quick leaps to new neighbourhoods, where the values stay until convergence. The second phase consists of convergence to the maximum in a similar manner to algorithms such as gradient descent; this phase can be seen in iterations 10–20 for $\boldsymbol{\beta}$ and $\sigma$. After these two phases, only minor updates occur until the convergence criteria is met. The reduction in fluctuations of $\boldsymbol{\beta}, \sigma$ observed after $t = 20$ is a result of the stochastic approximation beginning at the 20th iteration.

Figure E.1 displays the traceplots for BSGS-$\alpha$, which obtained a final value of $\alpha = 0.68$ and showed increased variability across the model parameters compared to BSGS. This led to slower convergence (202 iterations).

For BGSLOPE, the simplified GSLOBE model came at almost no cost in consistency between models. On the other hand, due to the additional parameters being estimated for BSGS, SGSLOBE was found to deviate often from BSGS. This was observed to an even greater extent when $\alpha$ was also estimated in BSGS-$\alpha$.

## 7.5 Initialisations and Sensitivity Analysis

Setting the initial parameter values for a Bayesian model is an important part of obtaining an accurate final model. The parameters for BGSLOPE and BSGS are initialised as

- $\sigma^0 = \frac{\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^0\|_2}{\sqrt{n-\hat{S}_v^0}}$, where $|\hat{S}_v^0|$ denotes the number of non-zero variables from $\boldsymbol{\beta}^0$. For BGSLOPE, this is taken as the number of non-zero groups instead, $|\hat{S}_g^0|$. This is a widely used estimator of the noise [51, 61, 189], and is also used for ABSLOPE.

- $\theta$. Using the mean of the posterior Beta distribution:
  - BGSLOPE: $\theta^0 = \frac{d_1+|\hat{S}_g^0|}{d_1+d_2+m}$.
  - BSGS: $\theta_v^0 = \frac{me_1-m+|\hat{S}_v^0|+1}{2+me_1+me_2+-2m+|\hat{S}_v^0|+\sum_{j\in\hat{S}_g^0}|\hat{S}_0^v\cap G_j|}$, where $\hat{S}_0^v \cap G_j$ are the non-zero variables in group $j$, and $\theta_g^0 = \frac{d_1+|\hat{S}_g^0|}{d_1+d_2+m}$.

- $c$. Using the means of the posterior Gamma distributions:
  - BGSLOPE: $c^0 = \min\left\{\frac{\sigma^0}{w_{r(\boldsymbol{\beta}^0,m)}}\frac{1+\sum_{j\in\hat{S}_g^0}p_j}{\sum_{j=1}^m\|(\boldsymbol{\beta}^0)^{(j)}\|_2\sqrt{p_j}}, 1\right\}$.
  - BSGS: $c_g^0 = \min\left\{\frac{\sigma^0\left(1+\sum_{j\in\hat{S}_g^0}p_j\right)}{v_{r_v(\boldsymbol{\beta}^0,p)}\sum_{i\in\hat{S}_v^0}c_v|\beta_i^0|+w_{r_g(\boldsymbol{\beta}^0,m)}\sum_{j\in\hat{S}_g^0}\|\boldsymbol{\kappa}^{(j)}(\boldsymbol{\beta}^0)^{(j)}\|_2\sqrt{p_j}}, 1\right\}$, and $c_v \sim \mathcal{U}[0,1]$.

We explore the impact of the choice of the initialisations for $\boldsymbol{\beta}$ and the prior hyperparameters for $\theta$. In ABSLOPE, $\boldsymbol{\beta}^0$ is initialised using the Lasso. Other initialisations considered in this section include ridge, the group lasso, and elastic net. SGL was also tested but provided no improvement over the lasso while adding substantial computational cost.

In ABSLOPE, the following schemes are considered for the Beta prior hyperparameters: i). $d_1 = 0.01n, d_2 = 0.01n$, ii). $d_1 = 2p, d_2 = 1 - 2/p$, iii). $d_1 = 1, d_2 = p$. We considered the following schemes:

- *Scheme 1.* BGSLOPE: $d_1 = 0.01n, d_2 = 0.01n$, BSGS: $d_1 = e_1 = 0.003n, d_2 = e_2 = 0.015n$.

- *Scheme 2.* BGSLOPE: $d_1 = 2/m, d_2 = 1 - 2/m$, BSGS: $d_1 = e_1 = 1, d_2 = e_2 = 1$.
- *Scheme 3.* BGSLOPE: $d_1 = 1, d_2 = m$, BSGS: $d_1 = e_1 = 1, d_2 = e_2 = p$.
- *Scheme 4.* BGSLOPE: $d_1 = m, d_2 = 1$, BSGS: $d_1 = e_1 = p, d_2 = e_2 = 1$.

These initialisations are compared in a simulation study to assess the stability of BGSLOPE and BSGS models applied to a Gaussian design matrix with a continuous response. The baseline setup from Section 7.7.1 is used, varying the correlation ($\rho_w$), signal strength ($s$), and dimensionality ($p$) as described there. The results are averaged across these three cases, with each case using 100 simulation repetitions. BSGS used $\alpha = 0.95$ and other hyperparameters were set as described in Section 7.7.2.

Figure 7.5 shows the mean squared error (MSE) of the fitted coefficients $\boldsymbol{\beta}$ to the true values for BSGS and BSGS-$\alpha$ under different $\boldsymbol{\beta}$ and prior choice initialisations. Figure E.3 shows the mean absolute error (MAE) of $\sigma$.

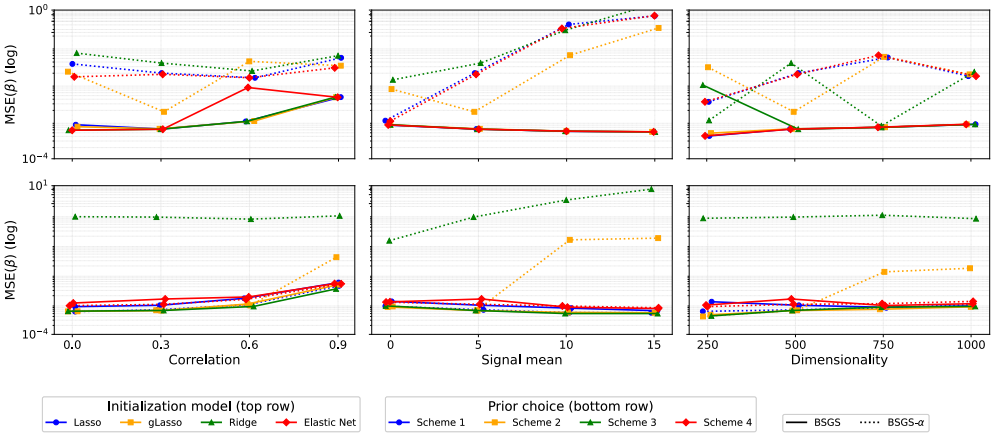



Figure 7.5: MSE($\boldsymbol{\beta}$) for BSGS and BSGS-$\alpha$ under different $\boldsymbol{\beta}$ initialisation models (top row) and Beta prior choices (bottom row), with a small amount of jitter added to allow the differences to be seen.

For all initialisation models, BSGS is found to be very robust, showing minimal differences, aside from the minor deviation for elastic net under $\rho = 0.6$. SGSLOBE (Figure E.4), BGSLOPE (Figure E.5), and GSLOBE (Figure E.5) are all found to be extremely robust under all choices. Indeed, the only models lacking stability are BSGS-$\alpha$ (Figure 7.5) and SGSLOBE-$\alpha$ (Figure E.4), demonstrating that the additional variability introduced by estimating $\alpha$ increases sensitivity to initialisations.

On average across the cases considered, BSGS-$\alpha$ selected $\alpha = 0.70$ and SGSLOBE-$\alpha$ selected $\alpha = 0.75$ (Figure E.6 shows the full distribution of values).

For the remainder of the chapter, we adopt lasso initialisation for $\boldsymbol{\beta}$ and Scheme 1 for the Beta priors. Across all cases considered, no single initialisation model consistently outperformed the others (Table E.2), and although the elastic net converged slightly faster on average (Table E.1), the differences were minimal. The lasso is preferred for its simplicity, robustness to unseen scenarios, and alignment with ABSLOPE, which also uses lasso initialisation. Scheme 1 provides the best balance with respect to parameter bias compared to other schemes (Table E.3). Overall, model performance is not sensitive to the choice of initialisation or prior scheme, and we do not expect these choices to significantly impact our subsequent analyses.

*Remark* 7.3 (Limitations of Scheme 1). As with any choice of hyperparameter scheme, there are some natural limitations. Under Scheme 1 for BGSLOPE, we have $\mathbb{E}[\theta] = 0.5$, which is independent of $n$. This is a desirable property, as it imposes no prior preference for either sparsity or density in the model. Such an assumption is reasonable in settings where no domain-specific knowledge is available. Under increasing sample size, the posterior has the following property:

$$\lim_{n\to\infty} \mathbb{E}[\theta \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma, \tilde{\mathbf{A}}] = \lim_{n\to\infty} \frac{0.01n + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1)}{0.02n + m} = 0.5,$$

so that the posterior mean converges back to the prior mean. In this limit, the data contributes no information. The same behaviour occurs in BSGS, where the posterior means of $\theta_g$ and $\theta_v$ revert to their prior values as $n \to \infty$.

In practice, this limitation is not critical, as BGSLOPE is primarily designed for high-dimensional genetic applications, where large-$n$ settings are uncommon. In the real-data analysis on the Trust-experts dataset ($n = 9759$), BGSLOPE outperformed competing methods in prediction, suggesting that this theoretical limitation has little practical impact.

## 7.6 Other Key Proposals

In this section, we detail two new proposals for model selection, specifically designed for SLOPE models: Adaptively Scaled SGS (AS-SGS) and Two-step Orthogonal (TSO). The models build on the methods described in Section 6.2.

### 7.6.1 Adaptively Scaled SGS (AS-SGS)

We propose an adaptive version of scaled SLOPE (Section 6.2.1) that, in addition to updating the noise and regression parameters, iteratively updates the penalty sequences.

Without loss of generality for Theorem 3.1, under orthogonality, the assumption of $\lambda = 1$ was made to derive the penalty sequences that give bi-level FDR control for SGS. If no such assumption is made, the sequences are derived as:

$$v_i^{\max}(\lambda) = \max_{j\in[m]} \left\{ \frac{\Phi^{-1}\left(1 - \frac{q_v i}{2p}\right) - \frac{1}{3}(1-\alpha)\lambda a_j w_j}{\alpha\lambda} \right\}, \; i \in [p], \tag{7.25}$$

$$w_i^{\max}(\lambda) = \max_{j\in[m]} \left\{ \frac{F_{\text{FN}}^{-1}\left(1 - \frac{q_g i}{m}\right) - \alpha\lambda \sum_{k\in G_j} v_k}{(1-\alpha)\lambda p_j} \right\}, \; i \in [m]. \tag{7.26}$$

Hence, an update of $\lambda$ would result in an adaptive update of the penalty sequences. This can be incorporated into the scaled regression framework to create an iterative penalty and noise procedure, described in Algorithm 6 and termed *Adaptively Scaled SGS* (AS-SGS). AS-SGS introduces an additional step compared to scaled regression, in which the penalty sequences are recalculated using the updated noise estimate.

**Algorithm 6** Adaptively Scaled SGS (AS-SGS)

**Require:** $\mathbf{y}$, $\mathbf{X}$

1: Initialise: $\hat{S}_v^+ \leftarrow \varnothing$
2: **while** $\hat{S}_v^+ \neq \hat{S}_v$ **do**
3: Set $\hat{S}_v \leftarrow \hat{S}_v^+$
4: Set $\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)\|_2^2}{n - |\hat{S}_v^\lambda|}$
5: Compute $\mathbf{v}^{\max}(\lambda)$ and $\mathbf{w}^{\max}(\lambda)$ (Equations 7.25 and 7.26)
6: Compute $\hat{\boldsymbol{\beta}}$ using SGS with $\lambda = \hat{\sigma}/n$, $\mathbf{v}^{\max}(\lambda)$, and $\mathbf{w}^{\max}(\lambda)$
7: Set $\hat{S}_v^+ \leftarrow \text{supp}(\hat{\boldsymbol{\beta}})$
8: **end while**

An interesting consequence of this proposed noise estimation approach is that it appears agnostic to the choice of $\alpha$. By applying AS-SGS to simulated data, we observed that the solutions produced along a path of $\alpha$ values were identical to each other (Figure 7.6). This is not the case for scaled regression applied with SGS.

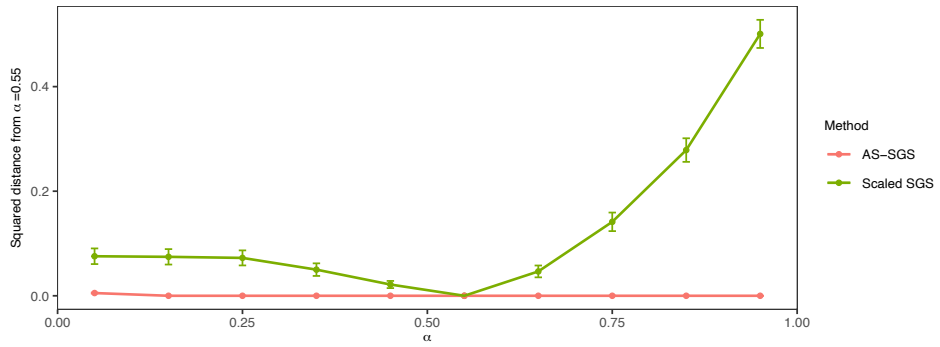



Figure 7.6: Mean squared distance of $\hat{\boldsymbol{\beta}}(\alpha)$ to $\hat{\boldsymbol{\beta}}(0.55)$ as a function of $\alpha \in (0, 1)$, shown for AS-SGS and scaled SGS. The data was generated as described in Remark 3.1 with the following minor differences: the signal of the active variables was set at $\beta_i = 5, i \in S_v$, and the results were averaged over six different sparsity cases; varying from the null true model to a true model with group sparsity proportion at 0.84. Each sparsity case had 100 simulation repetitions.

The AS-SGS model can be simplified by denoting

$$\tilde{v}_i^{\max} = \max_{j \in [m]} \left\{ \Phi^{-1}\left(1 - \frac{q_v i}{2p}\right) - \frac{1}{3}(1-\alpha)\lambda a_j w_j \right\}, \; i \in [p], \tag{7.27}$$

$$\tilde{w}_i^{\max} = \max_{j \in [m]} \left\{ \frac{F_{\text{FN}}^{-1}\left(1 - \frac{q_g i}{m}\right) - \alpha\lambda \sum_{k \in G_j} v_k}{p_j} \right\}, \; i \in [m], \tag{7.28}$$

leading to the penalty (via cancellation of $\alpha$ and $\lambda$): $\sum_{i=1}^{p} \tilde{v}_i |b|_{(i)} + \sum_{j=1}^{m} \|b^{(j)}\|_2 \sqrt{p_j} \tilde{w}_j$.

### SLOPE and gSLOPE

For SLOPE, if we similarly remove the assumption $\lambda = 1$, the penalty sequence is given by $v_i(\lambda) = \frac{1}{\lambda}\Phi^{-1}(1 - q_v i/2p)$. However, as the regularisation parameter enters the penalty sequences in a linear term, it cancels out so that

$$J_{\text{slope}}(\boldsymbol{\beta}; \mathbf{v}) = \lambda \sum_{i=1}^{p} \frac{1}{\lambda} v_i |\beta|_{(i)} = \sum_{i=1}^{p} v_i |\beta|_{(i)},$$

which reduces to the regular SLOPE penalty. The same is true for gSLOPE, and as a result, adaptively scaled regression is only applicable to SGS.

### 7.6.2 Two-step Orthogonal

We propose a new two-step procedure designed to transform a given setting into an orthogonal one, thereby enabling FDR control; see Section 6.2.3 for a background of two-step models. Procedures such as Gram-Schmidt can orthogonalise a matrix, although they rely on low-dimensional data. Assuming that the true number of signal variables is less than the number of observations ($|S_v| \leq n$), if we had access to $\mathbf{X}_S$, we could run the SLOPE procedure on this reduced input space to recover FDR control. We propose a new two-step procedure for SLOPE models with this aim, called *Two-step Orthogonal* (TSO):

1. Estimate $\hat{S}_v$ such that $|\hat{S}_v| \leq n$. We estimate this using the lasso with CV (picking the 1se model).

2. Orthogonalise $\mathbf{X}_{\hat{S}}$ using Gram-Schmidt to generate the orthogonalised design matrix $\tilde{\mathbf{X}}_{\hat{S}}$.

3. Fit a SLOPE model using $\tilde{\mathbf{X}}_{\hat{S}}$ and $\lambda = 1/n$, where the $1/n$ factor is from the loss function.

For gSLOPE and SGS, the procedure is identical, replacing Step 3 with optimising using the gSLOPE and SGS models instead.

The penalty sequences for SLOPE models are designed for orthogonal settings, but for low-dimensional Gaussian designs, a modified *Gaussian sequence* was derived by Bogdan et al. [24]; shown in Figure 7.7. The sequence is given by

$$v_i^{\text{GA}} = v_i^{\text{BH}} \sqrt{1 + f(i-1) \sum_{j<i} (v_j^{\text{GA}})^2}, \text{ where } f(i) = 1/(n-i-1).$$

The sequence flattens further down the ordering, thereby penalising small coefficients more to reduce false discoveries at the cost of some power. However, it reduces to the lasso sequence when $p > n$ [111], so it is not used in other SLOPE implementations in this thesis. Since Step 1 of TSO produces a low-dimensional setting, the Gaussian sequence is applied there, and Step 2 can be omitted. All TSO applications with SLOPE in this chapter use the Gaussian sequence unless otherwise specified (TSO-BH refers to TSO with the BH sequence). For gSLOPE and SGS, the penalty sequences remain unchanged.

Future work on TSO should more carefully consider the model used in Step 1 to estimate $\hat{S}_v$, as this ultimately determines the power recovered by TSO. The lasso was chosen due to its computational simplicity and well-understood performance as a variable selection tool. Additionally, the lasso can select at most $n$ predictors, ensuring the condition $\hat{S}_v \leq n$ is always satisfied [192].

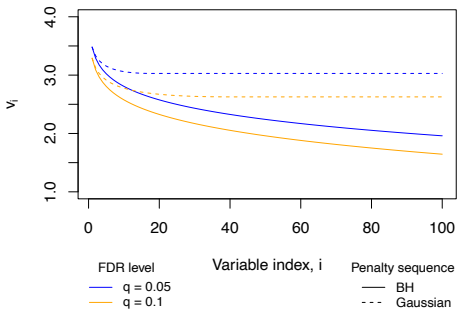



Figure 7.7: BH and Gaussian SLOPE sequences for different values of $q$ for $p = 100, n = 500$.

An ideal tool for this step would retain as much signal as possible by having very few false negatives (*i.e.*, a low type II error rate). Su et al. [187] describe the lasso's trade-off between type I and II errors, noting that low type II error is generally unattainable, although asymptotic theory shows it selects most true signals while including few, if any, noise variables. Replacements to the lasso here would ideally be designed for type II error control. This would lead to a model that combines type I (via SLOPE) and II error control procedures with the aim of type I error control.

For further results on the model consistency and trade-offs for the lasso see Wainwright [201], Wang et al. [203], Zhao and Yu [236]. Our empirical results show that the lasso is a sensible choice, enabling TSO to control the FDR while providing particularly fast runtime (Section 7.7).

## 7.7 Synthetic Data Experiments

In this section, we provide a comparison of the various model selection approaches for SLOPE, gSLOPE, and SGS. In particular, we focus on comparing our Bayesian approaches to other competitive approaches to gain an insight into the best performing model selection method, with the main emphasis being on FDR control.

### 7.7.1 Simulation Design

A multivariate Gaussian design matrix $\mathbf{X} \sim \boldsymbol{\mathcal{N}}_p(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^{400 \times p}$ was generated with block correlation structure, such that within-group correlation was $\Sigma_{i_1,i_2} = \rho_w$, for each pair $i_1, i_2 \in G_j, \forall j \in [m]$, for groups of sizes $[3, 25]$. For variables not in the

same group, across-group correlation was set to $\Sigma_{i_1,i_2} = \rho_a$ if $i_1 \in G_j$, $i_2 \in G_k$ such that $j \neq k$.

The response was generated using a linear model as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{N}}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. The group sparsity proportion was set to 0.8, with variable sparsity proportion of $\xi_v$ within an active group, and signal $\beta_i \sim \mathcal{N}(s, 10), i \in S_v$. The following data-generating parameters were varied:

- (Variable) Sparsity proportion: $\xi_v \in \{0.9, 0.7, 0.5, 0.3\}$.
- Signal strength: $s \in \{0, 5, 10, 15\}$.
- Dimensionality: $p \in \{200, 500, 750, 1000\}$.
- Noise: $\sigma \in \{0, 1, 2, 3\}$.
- Correlations: $\rho_w \in \{0, 0.3, 0.6, 0.9\}$ and $\rho_a \in \{0, 0.1, 0.2, 0.3\}$.

This forms six simulation cases. The baseline parameters were set to $\xi_v = 0.7, s = 5, p = 500, \sigma = 1, \rho_w = 0.3$, and $\rho_a = 0$ when other parameters were varied (see Table A.5). Each unique simulation setup was repeated 100 times.

Three cases are considered (with two subcases each):

- Case 1: Impact of signal, broken down into signal strength and sparsity.
- Case 2: Impact of data-generating parameters, broken down into dimensionality and noise.
- Case 3: Impact of correlation, broken down into within- and across-group correlation.
- The following cases are also considered as additional investigations:
  - Case i: Impact of $\alpha$ on SGS methods.
  - Case ii: Even groups.

### 7.7.2 Model Hyperparameters

All models were implemented with their default penalty sequences (see Remark 3.4), using $q_v = q_g = 0.1$ and all SGS models were applied with $\alpha = 0.95$. The Bayesian models used the lasso for $\boldsymbol{\beta}$ initialisation and Scheme 1 for the Beta prior choice (as described in Section 7.5). Strong screening rules were used for any optimisation

steps (SLOPE: [111], gSLOPE and SGS: Chapter 4). For all models employing a path, a log-linear path of 20 regularisation values was used, ranging from $\lambda_{\max}$ (the value at which the first variable enters the model) to $0.1\lambda_{\max}$.

The models are summarised in Table 7.1, and the algorithmic and model parameters in Tables A.4 and A.6. The oracle model refers to implementing a frequentist SLOPE model with knowledge of the true noise level.

### 7.7.3 Metrics

To measure the predictive performance, a test response, $\mathbf{y}_{\text{test}} = \mathbf{X}_{\text{train}}\boldsymbol{\beta}_{\text{train}} + \boldsymbol{\mathcal{N}}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, was generated. The following metrics are used: FDR (Definition 2.2), power (Definition A.1), out-of-sample (OOS) $\ell_2$ prediction error, $F_1$ score (Definition A.2), MSE($\boldsymbol{\beta}$), and MAE($\sigma$). Not all methods estimate the noise, so they are excluded from any plots containing MAE($\sigma$). We note that the bias of the regression coefficients, measured by MSE($\boldsymbol{\beta}$), is not critical, since penalised regression models are inherently biased and the coefficients can be debiased, for example, by using OLS.

In most cases, we present the best performing models in this section, with the other models presented in Appendix E.7.1 (to help plot readability). Figures E.8 and E.9 present the $F_1$ scores and MAE($\sigma$) values for all models under all cases considered. For SLOPE and SGS models, we primarily present the variable metrics, and for gSLOPE, the group ones.

Further simulation design and model implementation information can be found in Section A.3.

### 7.7.4 Summary of Approaches

Table 7.1 summarises all approaches considered in the synthetic study, with our proposals highlighted.

| Acronym | Full name | Description | Section |
|---|---|---|---|
| **Our Bayesian proposals** | | | |
| BGSLOPE | Bayesian Group SLOPE | One of the main proposals of the chapter. | 7.3.1 |
| GSLOBE | Group SLOBE | Using the conditional mean for the updates in BGSLOPE. | 7.4.4 |
| BSGS | Bayesian Sparse-group SLOPE | One of the main proposals of the chapter. | 7.3.2 |
| SGSLOBE | Sparse-group SLOBE | Using the conditional mean for the updates in BSGS. | 7.4.4 |
| BSGS-$\alpha$ | – | BSGS with $\alpha$ estimated via SAEM. | 7.4.2 |
| SGSLOBE-$\alpha$ | – | SGSLOBE with $\alpha$ estimated via SAEM. | 7.4.4 |
| **Other Bayesian methods** | | | |
| ABSLOPE | Adaptive Bayesian SLOPE | The spike-and-slab Bayesian SLOPE model. | 7.2 |
| SLOBE | – | Using the conditional mean for the updates in ABSLOPE. | 7.4.4 |
| **Other methods** | | | |
| AS-SGS | Adaptively Scaled SGS | Adaptive version of scaled regression, applicable only to SGS. | 7.6.1 |
| CV | Cross-validation | Picking the optimal $\lambda$ using cross-validation, chosen as the 1se model. | – |
| Knockoffs | – | The Knockoffs filter applied using SLOPE models. | 6.2.2 |
| Oracle | – | Fitting a SLOPE model with $\lambda$ equal to the (oracle) noise level. | – |
| Scaled regression | – | Iteratively estimates the noise and regression coefficients. | 6.2.1 |
| TSO | Two-step Orthogonal | Uses lasso and Gram–Schmidt to obtain an orthogonal low-dimensional input, then applies SLOPE. | 6.2.3 |
| TS-SLOPE | Two-step SLOPE | Applies SLOPE twice using the proximal operator. | 7.6.2 |

Table 7.1: Summary of the approaches used in the synthetic study. Our proposals have been highlighted in blue.

### 7.7.5 Results

#### Impact of Signal

Figure 7.8 shows that the Bayesian models maintain extremely low FDR levels (far below the specified level of 0.1) while still maintaining the highest power for all values of signal strength and sparsity proportion. From all other models considered, only TSO is also able to maintain FDR control for all values considered, but drops power in comparison to the Bayesian methods. The Bayesian methods perform well for all SLOPE models. For BGSLOPE, there is almost perfect recovery. Generally, most models are relatively unaffected by changes in the signal, with the Bayesian models in particular showing strong robustness.

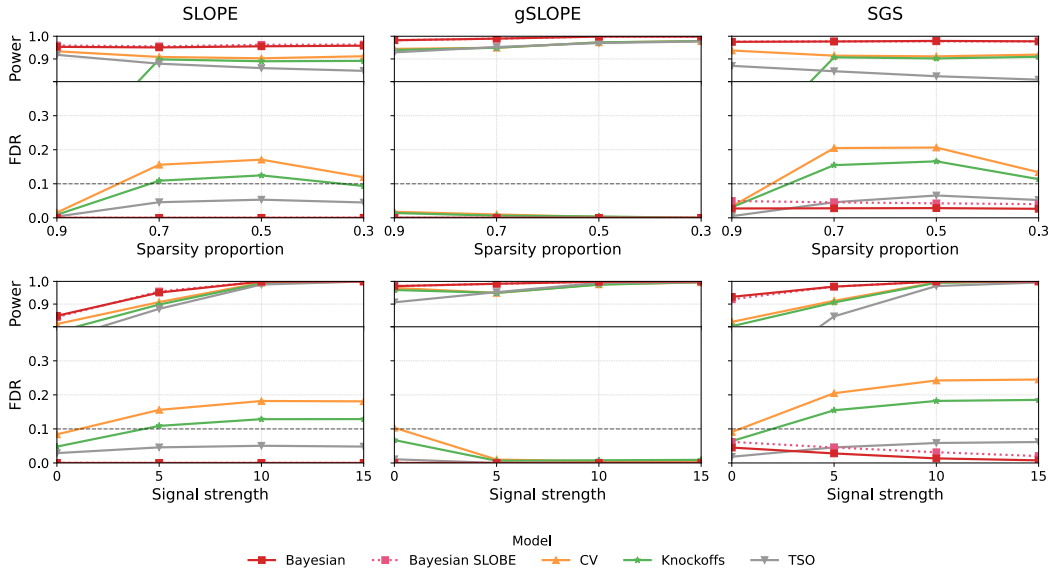



Figure 7.8: FDR (bottom plots) and power (top plots) for the best performing model selection approaches, as functions of the sparsity proportion (top row) and signal strength (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

Aside from controlling the FDR, Figure 7.9 provides further evidence for the strong performance of the Bayesian models under changing signals: the Bayesian methods were found to have the lowest OOS error. The Bayesian methods achieve this consistent FDR control and predictive performance while recovering the least biased $\boldsymbol{\beta}$ estimates (Figure 7.10). The Bayesian methods also clearly achieve the highest $F_1$ score under varying signal, showcasing the best selection performance (Figure E.8).

Figure E.10 presents the FDR and power for the other methods: BSGS-$\alpha$ is found to also control the FDR but suffers a drop off in power under very saturated signals. The other approaches considered here all fail to control the FDR across all model types.

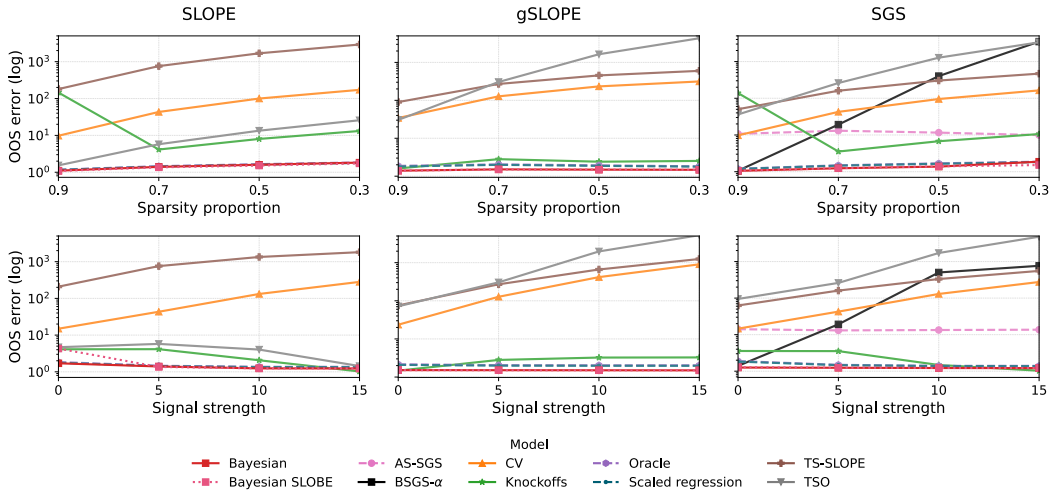



Figure 7.9: OOS error (log scale) for the best performing model selection approaches, as a function of the sparsity proportion (top row) and signal strength (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

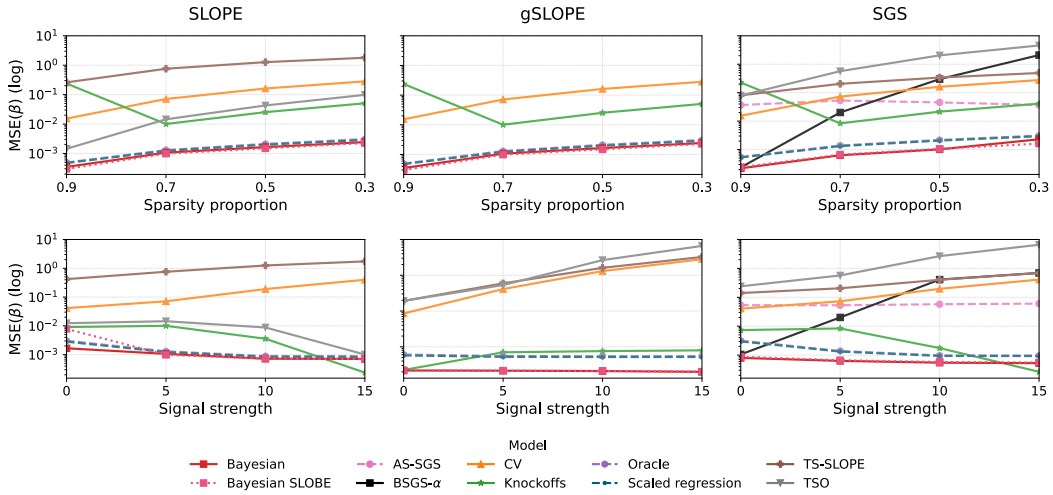



Figure 7.10: MSE($\boldsymbol{\beta}$) (log scale) for all model selection approaches, as a function of the sparsity proportion (top row) and signal strength (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

### Impact of Data-generating Parameters

The impact of changing the dimensionality and noise parameters in the data-generating process is assessed in Figure 7.11. Under increasing dimensionality, the Bayesian methods can clearly control the FDR while maintaining the highest power, showing further robustness to dimensionality. TSO is the only other approach that has somewhat satisfactory FDR control, although it no longer controls the FDR above $p = 750$.

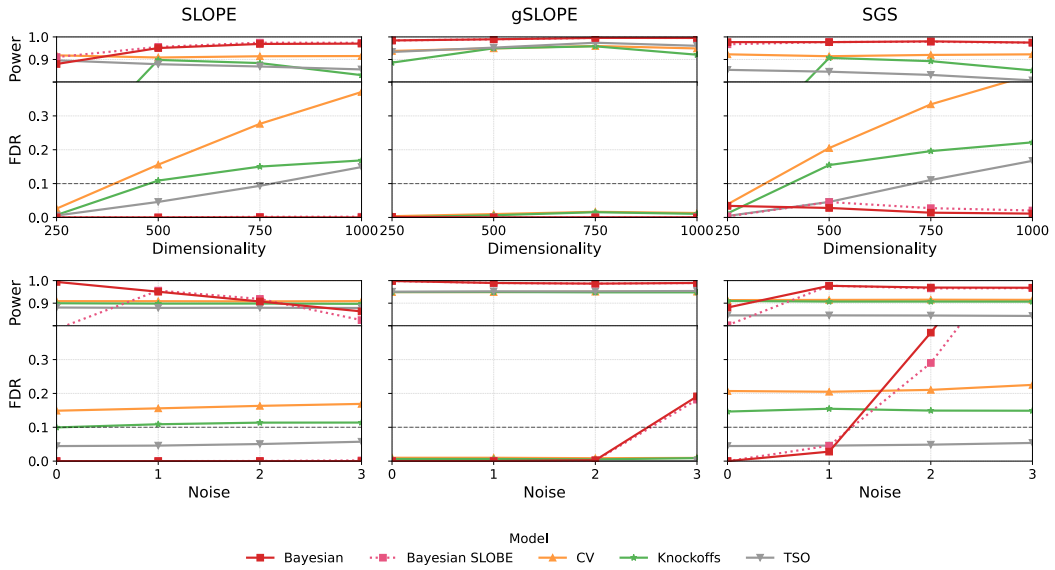



Figure 7.11: FDR (bottom plots) and power (top plots) for the best performing model selection approaches, as functions of the dimensionality (top row) and noise (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

Under increased noise, BGSLOPE and BSGS show the only occasions of failing to control the FDR. For BGSLOPE, the FDR is no longer controlled at $\sigma = 3$ and for BSGS at $\sigma \geq 2$, although both methods maintain consistently high power. The other methods considered here are very stable to changing noise levels, with TSO controlling the FDR under all values of the noise considered. The Bayesian methods are expected to be more sensitive to changing noise levels, given that they attempt to estimate it directly. Notably, BGSLOPE and BSGS perform well in estimating the noise as it increases. This suggests that the inability of the group-based Bayesian methods to control the FDR in highly noisy settings is not due to poor noise estimation, but rather reflects their difficulty in separating the signal

from the noise.

This behaviour is in sharp contrast to ABSLOPE, which does not scale its estimate of the noise appropriately (Figure 7.12). This failure to scale correctly leads to a drop in power, but the FDR level remains consistent (and below the threshold) (Figure 7.11). Therefore, with increasing noise levels, ABSLOPE becomes conservative, whereas BSGS and BGSLOPE become comparatively more permissive.

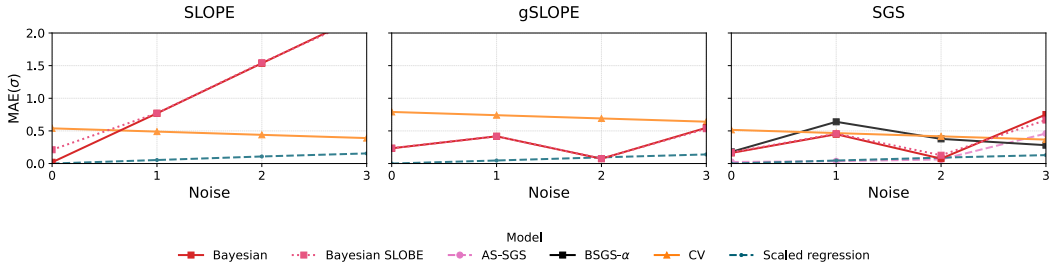



Figure 7.12: MAE($\sigma$) as a function of the noise for all model selection approaches that estimate the noise, split into the type of model (SLOPE, gSLOPE, SGS).

The high FDR levels of the Bayesian models under large noise demonstrate an inherent difficulty, as well as an opportunity, in using Bayesian models. Under high noise, the Bayesian methods are picking saturated models. To counter this, for BSGS, the priors on the mixture proportions $\theta_g, \theta_v$ can be changed to encourage more sparse models. As an example, changing the priors to $\theta_g, \theta_v \sim \text{Beta}(1, 1000)$ regains FDR control (Figure 7.13). This shows the additional flexibility gained from a Bayesian model, allowing a practitioner to have more control over the power-FDR trade-off while encoding their beliefs and knowledge of a problem setting.

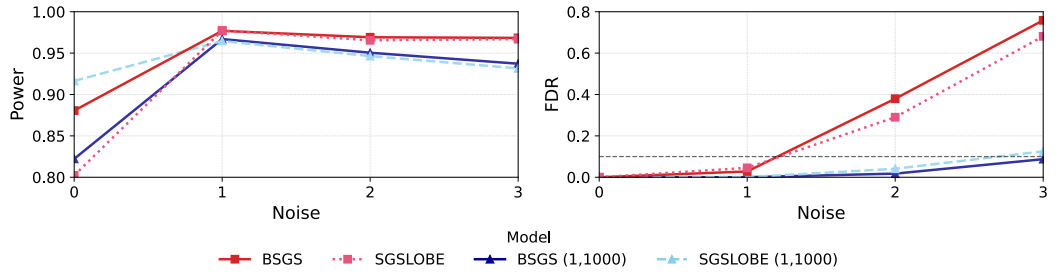



Figure 7.13: Power and FDR of BSGS under the Scheme 1 priors (BSGS) and $\theta_g, \theta_v \sim \text{Beta}(1, 1000)$ priors (BSGS (1,1000)). The comparison is also shown for SGSLOBE.

Figure E.11 presents the FDR and power for the other methods: BSGS-$\alpha$ im-

proves upon the weakness of BSGS under high noise, maintaining FDR control. Of the other approaches considered, TS-SLOPE is the only one that controls the FDR in any of the settings. Figures E.12 and E.13 present the MSE($\boldsymbol{\beta}$) and OOS error plots for this case, respectively. The Bayesian methods and scaled regression are found to obtain the best values for both. Figure E.14 illustrates how runtime increases with dimensionality. Although all methods exhibit similar scaling trends, the Bayesian approaches, while computationally expensive in absolute terms, show slightly slower growth with increasing dimension. In contrast, TSO demonstrates the most favourable scaling behaviour, with runtime increasing more gradually than for the other methods.

### Correlation

Our Bayesian methods are found to control the FDR under all cases of across-group correlation, while maintaining the highest power among all methods (Figure 7.14). The Knockoffs approach shows some promise under this setting, controlling the FDR for all correlation values for SLOPE and for $\rho_a \geq 0.1$ for SGS. In contrast to Knockoffs, TSO controls the FDR for $\rho_a \leq 0.1$. However, neither of these methods obtains the high power levels of the Bayesian methods.

For within-group correlation, the Bayesian methods also deliver FDR control, aside from the case of $\rho_w = 0.9$ for BSGS. The group-only nature of the BGSLOPE penalty allows it to be relatively unaffected by within-group correlation, as the model picks full groups. Similarly, the ABSLOPE model does not consider grouping structure, so it is also unaffected.

As in other cases, the FDR level and power are generally very robust for various correlation values for the Bayesian methods. The other competitive approaches considered here are much more sensitive, mostly losing FDR control once the correlations become large.

Figure E.15 presents the FDR and power for the other methods: BSGS-$\alpha$ achieves FDR control under both correlation cases, maintaining high power. As for BSGS, an increase in the FDR level is also observed for $\rho_a = 0.9$, but it does not lead to violation of FDR control. TS-SLOPE achieves FDR control for SLOPE and gSLOPE, but not SGS. Figures E.16 and E.17 present the MSE($\boldsymbol{\beta}$) and OOS error plots for this case, respectively. The Bayesian methods are found to obtain the best values for both.

Examining the noise estimation error under varying across-group correlations

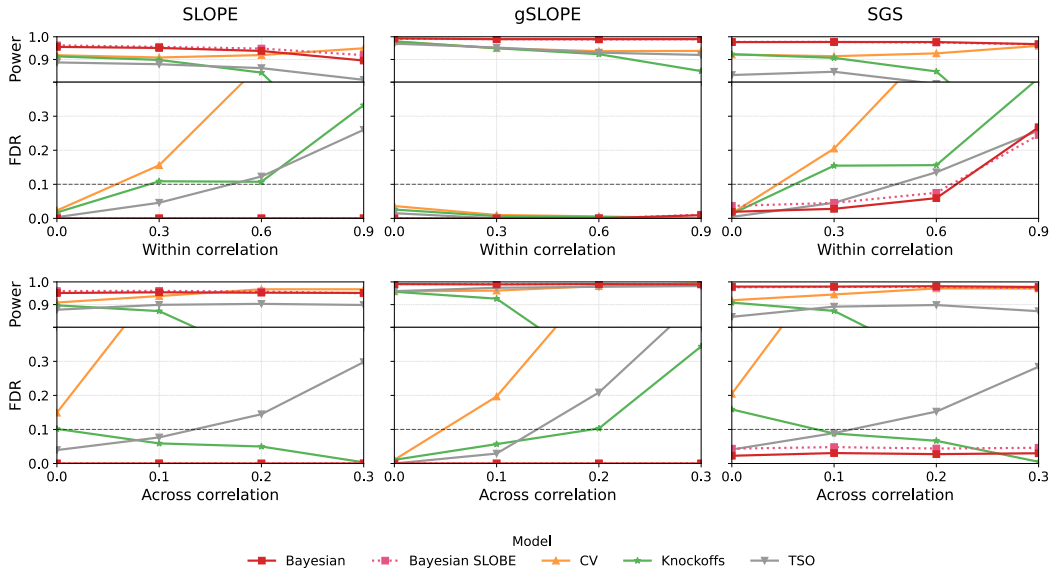



Figure 7.14: FDR (bottom plots) and power (top plots) for the best performing model selection approaches, as functions of within correlation (top row) and across correlation (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

provides particular insight (Figure E.9). The Bayesian models exhibit a consistent level of misestimation regardless of the correlation strength. In contrast, CV shows a roughly linear increase in error as correlation grows, while the scaled regression methods perform well up to $\rho_a \leq 0.1$, after which their errors increase dramatically. AS-SGS achieves highly accurate noise estimation across all correlation values.

### Summary of Results

Averaging across the six simulation scenarios presented in this section, our group-based Bayesian methods successfully control the FDR (Table 7.2) and achieve the highest power among all approaches that achieve FDR control (leading to the best $F_1$ scores). The Bayesian approaches also achieve the strongest predictive performance and obtain the least biased fitted values.

TSO was the only other method to control FDR across all model types, though with a slight loss of power likely due to type II errors from the lasso step (Section 7.6.2). The lasso enables TSO to achieve fast runtimes, with it also showing strong robustness and stability, but weak predictive performance. When combined with SLOPE, the Gaussian sequence restored predictive power without compromis-

ing FDR control (and even increasing power), highlighting the value of developing similar sequences for gSLOPE and SGS.

While Bayesian methods are generally reliable for estimating noise, they are not the most accurate in the cases of SLOPE and SGS. Specifically, for SLOPE, CV tends to yield more accurate noise estimates on average, whereas for SGS, AS-SGS provides exceptionally precise noise estimates. However, this accurate estimation of the noise for AS-SGS does not translate to strong performance in the other metrics, with it having extremely high FDR levels.

On the group metrics, BSGS did not control the FDR on average but achieved the highest power (Table E.5). It controlled variable FDR better than group FDR, while CV showed the opposite trend (Figure E.18). This stems from BSGS's hierarchical spike-and-slab structure, where groups are selected before variables, unlike CV's concurrent selection. Since variable-level control provides more informative associations, it is generally preferred.

Of all the SGS approaches, only BSGS-$\alpha$ and TSO were able to maintain bi-level FDR control, but both had a slight loss of power in comparison to BSGS. In fact, across all experiments, BSGS-$\alpha$ did not fail to control the FDR once. However, its estimation accuracy was lower, as BSGS provided more precise estimates of $\boldsymbol{\beta}$ and $\sigma$, resulting in superior predictive performance (Table 7.2).

Each method targets a specific metric but does not always manage to optimise it: CV aims to optimise predictive performance but underperforms in OOS; Knockoffs controls the FDR for SLOPE and gSLOPE but loses power; scaled models like AS-SGS recover noise accurately; and two-step procedures reduce bias, as seen with TSO for SLOPE. However, TS-SLOPE has been shown to be heavily biased [87], a finding supported by our results.

Based on Table 7.2, the Bayesian approaches provide the best model selection for SLOPE, offering FDR control, strong power and predictive performance, with competitive runtimes, and outperforming existing models on the metrics they are designed to optimise.

| Type | Model | FDR | Power | $F_1$ | MSE($\beta$) | MAE($\sigma$) | OOS | Time (s) |
|---|---|---|---|---|---|---|---|---|
| SLOPE | ABSLOPE | **$9 \times 10^{-6}$** $(9\times10^{-6})$ | 0.94 $(4\times10^{-3})$ | **0.97** $(2\times10^{-3})$ | **$2 \times 10^{-3}$** $(1\times10^{-4})$ | 0.81 $(2\times10^{-3})$ | **1.84** (0.06) | 3339.91 (98.38) |
| | SLOBE | **$3 \times 10^{-4}$** $(2\times10^{-4})$ | 0.94 (0.01) | 0.96 (0.01) | 0.02 (0.01) | 0.82 $(4\times10^{-3})$ | 23.66 (10.24) | 2700.49 (144.47) |
| | CV | 0.27 (0.01) | 0.92 (0.01) | 0.78 (0.01) | 0.18 (0.01) | **0.70** (0.02) | 81.62 (3.85) | 4.77 (0.20) |
| | Knockoffs | **0.10** (0.01) | 0.79 (0.01) | 0.80 (0.01) | 0.23 (0.02) | – | 362.54 (46.71) | 15.12 (0.89) |
| | Oracle | 0.54 (0.01) | 0.87 $(3\times10^{-3})$ | 0.58 (0.01) | 0.48 (0.01) | – | 629.36 (19.06) | 101.17 (3.24) |
| | Scaled | 0.47 (0.01) | **0.98** $(2\times10^{-3})$ | 0.67 (0.01) | 0.08 $(5\times10^{-3})$ | 4.30 (0.21) | 38.55 (2.55) | 194.14 (11.11) |
| | TS-SLOPE | **$4 \times 10^{-4}$** $(2\times10^{-4})$ | 0.61 (0.01) | 0.73 (0.01) | 0.96 (0.03) | – | 1510.52 (68.33) | 26.06 (1.24) |
| | TSO | **0.08** $(5\times10^{-3})$ | 0.88 (0.01) | 0.90 $(5\times10^{-3})$ | 0.02 $(2\times10^{-2})$ | – | 7.10 (0.64) | 1.56 (0.08) |
| | TSO-BH | **0.08** (0.01) | 0.83 (0.01) | 0.87 (0.01) | 1.99 (0.12) | – | 1713.95 (137.13) | **0.53** (0.01) |
| gSLOPE | BGSLOPE | **$9 \times 10^{-3}$** $(1\times10^{-3})$ | **0.99** $(4\times10^{-3})$ | **0.99** $(3\times10^{-3})$ | **$5 \times 10^{-3}$** $(3\times10^{-4})$ | **0.43** (0.01) | **2.01** (0.05) | 511.21 (22.99) |
| | SLOBE | **$8 \times 10^{-3}$** $(1\times10^{-3})$ | **0.99** $(4\times10^{-3})$ | **0.99** $(3\times10^{-3})$ | **$5 \times 10^{-3}$** $(3\times10^{-4})$ | **0.43** (0.01) | **2.01** (0.05) | 321.30 (13.56) |
| | CV | **0.08** $(5\times10^{-3})$ | 0.96 (0.01) | 0.92 (0.01) | 0.68 (0.03) | 1.12 (0.03) | 220.46 (9.51) | 750.76 (20.64) |
| | Knockoffs | **0.03** (0.01) | 0.90 (0.01) | 0.92 (0.01) | 0.18 (0.02) | – | 305.01 (31.66) | 1311.43 (52.47) |
| | Oracle | 0.58 (0.01) | 0.96 $(4\times10^{-3})$ | 0.55 (0.01) | 0.49 (0.01) | – | 617.95 (18.87) | 33.59 (1.73) |
| | Scaled | 0.56 (0.01) | **0.99** $(2\times10^{-3})$ | 0.59 (0.01) | 0.10 (0.01) | 3.45 (0.21) | 29.56 (2.12) | 77.60 (12.30) |
| | TS-gSLOPE | **$6 \times 10^{-3}$** $(1\times10^{-3})$ | 0.89 (0.01) | 0.94 (0.01) | 0.87 (0.03) | – | 589.73 (27.51) | 507.42 (18.33) |
| | TSO | **0.03** $(2\times10^{-3})$ | 0.96 (0.01) | 0.96 (0.01) | 1.95 (0.12) | – | 1968.43 (159.04) | **1.04** (0.04) |
| SGS | AS-SGS | 0.70 (0.01) | **0.99** $(1\times10^{-3})$ | 0.42 (0.01) | 0.10 $(3\times10^{-3})$ | **0.06** (0.01) | 17.29 (0.40) | 206.35 (25.84) |
| | BSGS | **0.08** $(4\times10^{-3})$ | 0.97 $(4\times10^{-3})$ | 0.93 $(3\times10^{-3})$ | $5 \times 10^{-3}$ $(1\times10^{-5})$ | 0.44 $(4\times10^{-3})$ | 3.96 (0.76) | 385.80 (14.76) |
| | SGSLOBE | **0.09** $(4\times10^{-3})$ | 0.97 $(3\times10^{-3})$ | 0.93 $(3\times10^{-3})$ | **$3 \times 10^{-3}$** $(2\times10^{-4})$ | 0.44 $(5\times10^{-3})$ | **2.40** (0.13) | 428.13 (16.44) |
| | BSGS-$\alpha$ | **$8 \times 10^{-3}$** $(1\times10^{-3})$ | 0.92 (0.01) | 0.94 (0.01) | 0.17 (0.05) | 0.61 (0.01) | 250.39 (68.00) | 537.29 (17.82) |
| | SGSLOBE-$\alpha$ | **0.02** $(2\times10^{-3})$ | 0.95 $(4\times10^{-3})$ | **0.96** $(3\times10^{-3})$ | $9 \times 10^{-3}$ $(1\times10^{-3})$ | 0.58 (0.01) | 7.57 (1.25) | 688.77 (22.19) |
| | CV | 0.30 (0.01) | 0.93 (0.01) | 0.75 (0.01) | 0.18 (0.01) | 0.66 (0.02) | 80.33 (3.83) | 1061.34 (37.58) |
| | Knockoffs | 0.14 (0.01) | 0.80 (0.01) | 0.78 (0.01) | 0.23 (0.02) | – | 357.97 (46.12) | 1607.69 (65.59) |
| | Oracle | 0.51 (0.01) | 0.87 $(3\times10^{-3})$ | 0.60 (0.01) | 0.49 (0.01) | – | 629.27 (19.07) | 26.21 (1.24) |
| | Scaled | 0.45 (0.01) | 0.98 $(2\times10^{-3})$ | 0.67 (0.01) | 0.10 (0.01) | 5.55 (0.17) | 54.43 (2.79) | 46.39 (1.78) |
| | TS-SGS | 0.17 (0.01) | 0.80 (0.01) | 0.84 (0.01) | 0.34 (0.01) | – | 425.05 (22.17) | 694.72 (29.89) |
| | TSO | **0.08** (0.01) | 0.83 (0.01) | 0.87 (0.01) | 1.97 (0.12) | – | 1671.26 (134.68) | **0.98** (0.04) |

Table 7.2: Key metrics averaged over six simulation cases shown by model type, with standard errors in grey. SLOPE and SGS report variable metrics; gSLOPE reports group metrics. Best-performing models in each metric per model type are in **bold**, except FDR, where all FDR $\leq 0.1$ are **bold**.

### Additional Insights

Here, we briefly discuss some additional insights into the group-based Bayesian methods.

**Impact of $\alpha$ on SGS models.** Figure 7.15 shows that BSGS is extremely robust to the choice of $\alpha$, achieving FDR control under all values of $\alpha \in [0.05, 0.95]$, which only TSO is also able to achieve. BSGS achieves this while maintaining higher power and lower bias than TSO.

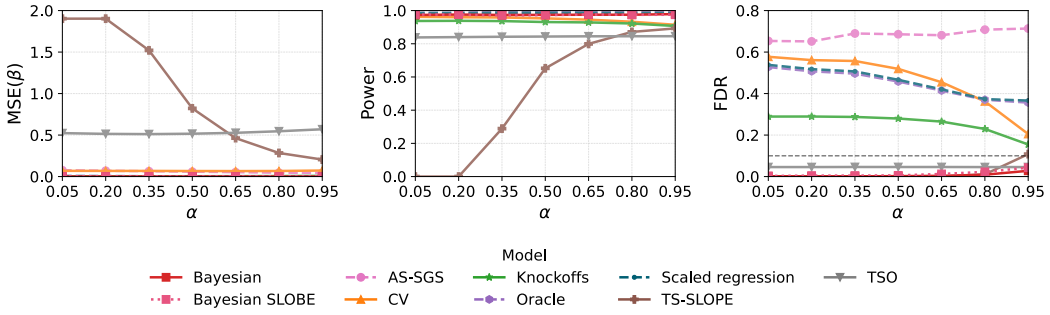



Figure 7.15: MSE($\boldsymbol{\beta}$), power, and FDR for all SGS model selection approaches as a function of $\alpha$ under the baseline setting (Table A.6).

The added adaptivity of BSGS over CV allows it to adapt to changing $\alpha$ to achieve strong performance for all values. On the other hand, for TSO, the use of the lasso in Step 1 allows it to filter out irrelevant variables (possibly in active groups), minimising the downside of a group-only fit towards $\alpha = 0.05$. This explains the consistent performance of TSO, regardless of $\alpha$. The performance of scaled regression and the oracle method is very similar in this case due to scaled regression estimating the noise very accurately (the average MAE($\sigma$) was 0.03).

Generally, most models are found to have decreasing FDR levels approaching $\alpha = 0.95$. This is to be expected, as for $\alpha = 0.05$ the models perform more closely to their group-only counterparts, in which achieving variable FDR is unlikely.

**Equal group size.** The approaches were also applied to an equal-sized grouping structure, varying the size of each group $p_j \in \{5, 10, 15, 20\}$ (Figure E.19). Under all group sizes, the Bayesian models (including BSGS-$\alpha$) and TSO are found to control the FDR, while the Bayesian models achieve higher power. No other methods were able to control the FDR for all model types.

## 7.8 Real Data Experiments

In this section, we apply the model selection approaches that showed the most promise in the synthetic study to real data: the Bayesian methods, CV, and TSO.

### 7.8.1 Datasets

The models are applied to seven real-world datasets previously used in this thesis: Cancer, Colitis, BRCA1, Carbotax, Rhee, Scheetz, and Trust-experts. The datasets have varied characteristics to test the models under different scenarios; five of the datasets are high-dimensional, four different grouping structure approaches are taken, and two of the datasets are semi-synthetic. All datasets have continuous responses, so are fit using linear models.

The Cancer and Colitis datasets (as used in Chapters 3 and 4) originally have binary responses. To enable evaluation of the FDR on real data and allow us to use this dataset with a linear model, a synthetic response was generated for each using the linear model: $\mathbf{y}_{\text{semi-syn}} = \mathbf{X}_{\text{real}}\boldsymbol{\beta}_{\text{semi-syn}} + \boldsymbol{\mathcal{N}}_n(\mathbf{0}, \mathbf{I}_n)$, with group sparsity proportion set at 0.9 and variable sparsity proportion in an active group at 0.85, and the signal strength at 5.

Each dataset was divided into five 80/20 train/test partitions, and the final results were averaged over these splits. To assess predictive performance, we use the Normalised MSE (NMSE) of the prediction, defined by

$$\text{NMSE} = \text{MSE}(\mathbf{y}_{\text{test}}, \mathbf{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{\text{train}})/\text{Var}(\mathbf{y}_{\text{test}}).$$

This gives a dimensionless metric: NMSE = 1 corresponds to predicting the mean of $\mathbf{y}$ and values below 1 indicate improvement.

Further details on the datasets are provided in Table 7.3 and Appendix A.3.3.

| **Dataset** | $p$ | $n$ | $m$ | **Group sizes** | **Grouping** | **Response** | **Downloaded on** |
|---|---|---|---|---|---|---|---|
| Cancer | 7057 | 60 | 1277 | $[1, 292]$ | Gene sets | Semi-synthetic continuous | 08/2024 |
| Colitis | 11999 | 127 | 1528 | $[1, 497]$ | Gene sets | Semi-synthetic continuous | 08/2024 |
| BRCA1 | 10000 | 536 | 189 | $[1, 3780]$ | SVD | Continuous | 05/2024 |
| Carbotax | 10000 | 101 | 200 | $[1, 126]$ | K-means | Continuous | 08/2024 |
| Rhee | 361 | 842 | 207 | $[1, 9]$ | SVD | Continuous | 10/2025 |
| Scheetz | 18975 | 120 | 379 | $[1, 165]$ | K-means | Continuous | 08/2024 |
| Trust-experts | 101 | 9759 | 7 | $[4, 51]$ | Factors | Continuous | 05/2024 |

Table 7.3: Dataset information for the seven datasets used in the real data analysis.

### 7.8.2 Model Hyperparameters

The models were applied in the same way as for the synthetic study, with the following minor changes. For the SGS models, two changes are made. First, $\alpha = 0.99$ is used, as in Chapter 3 it was observed that this yields better predictive performance on real data. Second, as we expect the real data signals to be weaker, the priors on the mixture proportions have been adjusted to reflect this: $\theta_g, \theta_v \sim \text{Beta}(5, 1)$. This allows for BSGS to pick denser models.

Detailed model and algorithmic hyperparameter information is provided in Tables A.4 and A.6.

### 7.8.3 Results

Table 7.4 reports the NMSE scores across all datasets. For SLOPE and SGS, the Bayesian approach attains the lowest NMSE on all datasets except the semi-synthetic Cancer dataset, where TSO performs best. For gSLOPE, BGSLOPE yields the best performance on all datasets except Colitis and BRCA1, where CV performs best. It is interesting to note the extremely poor predictive performance of TSO on the Colitis data for both gSLOPE and SGS, but not for SLOPE. These results mirror the findings from the synthetic study, where incorporating TSO with the Gaussian sequence led to substantial predictive improvements.

| **Dataset** | **SLOPE** | | | **gSLOPE** | | | **SGS** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Bayesian | CV | TSO | Bayesian | CV | TSO | Bayesian | CV | TSO |
| Cancer | 0.803 (0.111) | 0.479 (0.094) | **0.235** (0.053) | **0.258** (0.031) | 0.390 (0.088) | 0.361 (0.121) | 0.291 (0.039) | 0.437 (0.095) | **0.287** (0.057) |
| Colitis | **0.419** (0.052) | 0.891 (0.131) | 0.898 (0.139) | 0.759 (0.105) | **0.700** (0.095) | 458.813 (151.254) | **0.401** (0.052) | 0.640 (0.105) | 801.935 (450.085) |
| BRCA1 | **0.445** (0.043) | 0.486 (0.062) | 0.555 (0.048) | 0.884 (0.094) | **0.766** (0.067) | 1.028 (0.056) | **0.467** (0.043) | 0.498 (0.060) | 1.028 (0.056) |
| Carbotax | **0.838** (0.175) | 0.923 (0.209) | 1.000 (0.223) | **0.742** (0.156) | 0.961 (0.209) | 1.001 (0.223) | **0.878** (0.195) | 0.955 (0.216) | 0.997 (0.225) |
| Rhee | **0.138** (0.007) | 0.371 (0.022) | 0.450 (0.022) | **0.140** (0.008) | 0.348 (0.020) | 0.970 (0.037) | **0.138** (0.007) | 0.328 (0.023) | 0.970 (0.037) |
| Scheetz | **0.442** (0.102) | 0.478 (0.107) | 1.026 (0.388) | **0.479** (0.112) | 0.694 (0.229) | 1.026 (0.388) | **0.715** (0.236) | 0.744 (0.270) | 1.026 (0.388) |
| Trust-experts | **0.345** (0.009) | 0.420 (0.009) | 0.399 (0.009) | **0.345** (0.009) | 0.352 (0.009) | 0.847 (0.015) | **0.345** (0.009) | 0.352 (0.009) | 0.935 (0.015) |

Table 7.4: NMSE for the Bayesian methods, CV, and TSO for each model type (SLOPE, gSLOPE, SGS), with standard errors shown in grey. The best performing method for each data within the model type has been highlighted in **bold**. The best method overall for a dataset is underlined.

Overall, these results demonstrate that the Bayesian methods consistently deliver superior predictive performance on real data. This trend mirrors the findings from the synthetic experiments, confirming that Bayesian model selection is the most effective approach when prediction accuracy is the primary objective. While comparing SLOPE variants is not our primary objective, SLOPE achieved the best performance on four datasets, SGS on three, and gSLOPE on two.

Table E.6 shows the NMSE for the Bayesian models compared to the SLOBE variants. For SLOPE and gSLOPE, the simplified sampling scheme used in SLOBE substantially reduces predictive performance on most datasets. In contrast, for SGS, the SGSLOBE model performs better in most cases, suggesting that in the complex sampling regime of BSGS, the reduced variability introduced by SGSLOBE is beneficial. In fact, SGSLOBE is found to be the best-performing method across all variants for five of the datasets.

| **Dataset** | **BSGS** | | **BSGS** $\alpha = 0.95$ | | **BSGS default** | | **BSGS-**$\alpha$ | |
|---|---|---|---|---|---|---|---|---|
| | Base | SLOBE | Base | SLOBE | Base | SLOBE | Base | SLOBE |
| Cancer | 0.291 (0.039) | 0.291 (0.039) | 0.282 (0.035) | 0.281 (0.036) | 0.403 (0.045) | 0.290 (0.039) | <u>0.272</u> (0.036) | 0.275 (0.036) |
| Colitis | 0.401 (0.052) | <u>0.400</u> (0.052) | 5.187 (0.735) | 5.187 (0.735) | 0.783 (0.103) | 0.773 (0.103) | 0.740 (0.098) | 0.737 (0.098) |
| BRCA1 | 0.467 (0.043) | <u>0.445</u> (0.047) | 0.609 (0.075) | 0.459 (0.051) | 1.028 (0.056) | 0.873 (0.049) | 0.510 (0.056) | 0.500 (0.056) |
| Carbotax | 0.878 (0.195) | 0.844 (0.177) | 0.875 (0.197) | 0.828 (0.162) | 0.999 (0.224) | 0.823 (0.176) | 0.986 (0.228) | <u>0.789</u> (0.164) |
| Rhee | <u>0.138</u> (0.007) | <u>0.138</u> (0.007) | 0.139 (0.007) | <u>0.138</u> (0.007) | 0.832 (0.041) | 0.970 (0.037) | 0.139 (0.008) | <u>0.138</u> (0.007) |
| Scheetz | 0.715 (0.236) | <u>0.397</u> (0.079) | 0.808 (0.263) | 0.416 (0.088) | 1.026 (0.388) | 1.026 (0.388) | 1.026 (0.388) | 0.498 (0.132) |
| Trust-experts | <u>0.345</u> (0.009) | <u>0.345</u> (0.009) | <u>0.345</u> (0.009) | <u>0.345</u> (0.009) | 0.347 (0.009) | 0.347 (0.009) | <u>0.345</u> (0.009) | <u>0.345</u> (0.009) |

Table 7.5: NMSE for the BSGS models and the SLOBE variants, with standard errors shown in grey. BSGS (which uses $\alpha = 0.99$), BSGS $\alpha = 0.95$, and BSGS-$\alpha$ all use the $\theta_g, \theta_v \sim \text{Beta}(5, 1)$ priors. BSGS default uses the default prior scheme from the synthetic study (Scheme 1, see Section 7.5). The best method overall for a dataset is <u>underlined</u>.

Table 7.5 presents the NMSE for several additional BSGS variants, to assess which hyperparameter configurations yield the best predictive performance. The BSGS model with priors $\theta_g, \theta_v \sim \text{Beta}(5, 1)$ clearly outperforms the variant using the default priors from the synthetic study, indicating that accounting for the higher

noise level typically present in real data allows domain knowledge to enhance model performance. Moreover, the model with $\alpha = 0.99$ consistently outperforms that with $\alpha = 0.95$, providing further evidence that SGS models applied to real data should adopt $\alpha = 0.99$ as a default. Finally, BSGS-$\alpha$, which also employs the Beta$(5, 1)$ prior structure, performs comparably to BSGS using the same priors with $\alpha = 0.99$.

## 7.9 Discussion

SLOPE models have been shown to control the FDR when the design matrix is orthogonal. In this chapter, we address the problem of model selection for SLOPE models under more general settings, with the goal of restoring FDR control. To achieve this, we develop new methodologies and compare them to existing approaches to identify the optimal strategy for implementing SLOPE under these broader conditions.

Our main methodological development is the proposal of group-based Bayesian SLOPE models, BGSLOPE and BSGS, which use spike-and-slab priors that are equivalent to gSLOPE and SGS, respectively. The Bayesian framework allows the models to learn the noise alongside the regression parameters, while providing uncertainty quantification, feature importance, and adaptivity to the underlying sparsity structure. BSGS offers the first continuous spike-and-slab framework for sparse-group models, providing the aforementioned benefits at both the variable and group levels. We also propose a novel two-step approach, Two-step Orthogonal (TSO), which attempts to transform a general setting into an orthogonal one, allowing the SLOPE FDR properties to be used, and an adaptive scaled regression approach (AS-SGS).

We carry out a comprehensive comparison of different model selection approaches for SLOPE models using both synthetic and real data studies. We compare our group-based Bayesian models and TSO with CV, Knockoffs, scaled regression, and two-step SLOPE to identify the most effective approach for enabling SLOPE to maintain FDR control under general design settings, as well as achieving strong predictive performance.

The Bayesian models clearly outperformed other methods across our synthetic experiments, showing strong FDR control, predictive performance, and parameter recovery. Of the models that obtained FDR control on average across all cases, the Bayesian models retained the highest power. Aside from the Bayesian methods, TSO

was also found to perform well, achieving FDR control in extremely fast runtime. The procedure was found to perform especially well for SLOPE, due to the existence of the Gaussian sequence, which resulted in strong predictive performance. In the real data study, the Bayesian models achieved the best predictive performance, comprehensively beating other methods on the majority of datasets considered.

Our Bayesian methodology eliminates the need to tune $\lambda$ by learning it via the noise, but introduces additional hyperparameters compared to the frequentist counterparts. For BSGS, we need to set $(\alpha, q_v, q_g, d_1, d_2, e_1, e_2)$ and a $\boldsymbol{\beta}$ initialisation, though $\alpha$ can be learned using BSGS-$\alpha$. Section 7.5 shows that the models are generally insensitive to prior hyperparameters on synthetic data, yet they allow encoding of domain knowledge and provide an opportunity for improved model performance. For example, in the high-noise synthetic experiment (Figure 7.13), we set $d_1 = e_1 = 1$ and $d_2 = e_2 = 1000$ to strongly encourage sparsity. This adjustment enabled BSGS to control the FDR, which it failed to do under the default scheme. In the real data study, the default scheme struggled to capture the signals, resulting in poor predictive performance. On the other hand, setting $d_1 = e_1 = 5$ and $d_2 = e_2 = 1$ produced denser models, which allowed the signal to be picked up. This is in a similar vein to the findings of Section 3.6.5 in Chapter 3 for SGS.

We conclude that under general settings, the recommended application of SLOPE models should be a Bayesian framework, as developed in this chapter. This is supported by the strong empirical performance of the Bayesian methods on synthetic and real data. The Bayesian models effectively control the FDR, achieve high power, yield less biased estimates, and produce accurate predictions.

# CHAPTER 8

# Discussion

In this thesis, we developed new methodologies for sparse-group settings. The developments are primarily motivated by applications in genetics, where controlling the FDR is crucial, and genes naturally occur in groups. Accordingly, the proposed methods are designed to perform well in such contexts, while being formulated in a general framework to ensure broad applicability. Our research was guided by practical problems, ensuring that the developments are useful to practitioners when applying sparse-group models.

Here, we provide a summary of the contributions, discuss limitations of our methods, and propose future research directions.

## 8.1 Summary of Contributions

The thesis has made several significant contributions to the sparse-group and FDR literature.

In Chapter 3, we developed a sparse-group method, Sparse-group SLOPE (SGS), that incorporates SLOPE into a sparse-group framework, allowing for bi-level selection. SGS has a non-separable penalty norm, so gradient descent is not guaranteed to converge. Therefore, SGS is optimised using a proximal splitting algorithm, ATOS. ATOS was adapted to efficiently incorporate the existing SLOPE and gSLOPE proximal operators, using a derived transformation that enables the gSLOPE operator's direct application within ATOS. We also developed new interdependent penalty sequences that account for the two levels of sorting, providing bi-level FDR control under orthogonal settings, as verified through experimental evaluation. Our

theoretical results are the first in the literature to concurrently prove FDR control at both the variable and group levels. SGS was applied to synthetic and real data, in which it was found to outperform existing regression approaches in disease prediction tasks.

Sparse-group models, particularly SGS, are computationally expensive to fit, and the literature offers no strong screening methods for them. To address this, in Chapter 4, we introduced a sparse-group screening framework that eliminates inactive features before optimisation, yielding substantial computational savings. The framework exploits the sparse-group structure of the norms through two layers of screening: first, removing inactive groups, then discarding inactive variables within the remaining groups. Applied to gSLOPE and SGS, it provides the first screening procedures developed for either model. These rules, derived using subdifferentials to handle non-separability, significantly reduce fitting time without compromising solution optimality.

In Chapter 5, we extended this sparse-group screening framework to the Sparse-group Lasso (SGL) and the adaptive SGL (aSGL) to form the Dual Feature Reduction (DFR) method. This is achieved using the dual norm of SGL, given by the $\epsilon$-norm. We provide theoretical results that allow the aSGL dual norm to also be expressed as the $\epsilon$-norm. Our rules are the first strong screening rules for SGL and the first of any type for aSGL. DFR was shown to clearly outperform existing screening approaches for SGL.

Model selection is a central challenge in penalised regression, as it governs model sparsity and predictive performance. While all penalised methods face this issue, it is particularly pronounced in SLOPE, where it directly determines the FDR level. In Chapter 7, we address model selection for SLOPE under general settings. Although SLOPE's FDR guarantees hold for orthogonal designs, they often fail in general settings due to the lack of methods for tuning the regularisation parameter to maintain reliable FDR control. To achieve this aim, we developed Bayesian methodology for group-based SLOPE models, BGSLOPE and BSGS, using spike-and-slab priors in a hierarchical framework. The BSGS model is the first to introduce continuous spike-and-slab priors for sparse-group models, extending its applicability to a variety of norms, including SGL. Our Bayesian models are adaptive to the underlying sparsity structure, provide uncertainty quantification, and offer flexibility for incorporating practitioner knowledge through the choice of prior hyperparameters.

We also proposed a new two-step method, Two-step Orthogonal (TSO), which

seeks to transform a general setting into an orthogonal one, as well as an adaptive version of scaled regression for SGS, called Adaptively Scaled SGS (AS-SGS). Our experimental results compared our proposed new methods to other existing approaches for all SLOPE models (SLOPE, gSLOPE, SGS) to determine the best implementation solution for SLOPE under a variety of general synthetic and real data settings. The results demonstrated that our Bayesian methods control the FDR in most general settings while achieving higher power and more accurate predictions than other competitive approaches. These findings indicate that the optimal implementation of SLOPE in general settings is through the Bayesian methodology developed in this chapter.

In this thesis, our primary aim was to develop methodology for SGS, but we often found that no work existed on extending the methodologies of interest from SLOPE to gSLOPE. As such, in Chapters 4 and 7, we also develop methodology for gSLOPE en route to extending it to SGS.

We have highlighted the key contributions above, but note that the methodology developed in this thesis applies more broadly. SLOPE models generalise other norms, so any SLOPE-based method developed extends to them by adjusting the weights, which is particularly useful for the OWL model, as it remains understudied in the literature.

## 8.2 Methodological Limitations

While our developed methods have demonstrated strong performance in our experiments, often outperforming existing approaches, like any methods, they have certain limitations.

### 8.2.1 Orthogonal Design Matrix and Linear Models

Theoretical results show that SLOPE models control the FDR under orthogonal design matrices in linear models; Theorem 3.1 shows this for SGS. While for SLOPE the FDR control is exact, for gSLOPE and SGS, it is overly conservative (see Remark 2.3).

Orthogonal settings in practice are relatively limited. Under more general settings, the regularisation parameter $\lambda$ requires tuning. Traditionally, this is done using CV, but our results in Section 3.5 show that SLOPE models do not maintain

FDR control with general design matrices using CV. While some work has been done to extend the FDR control under orthogonal designs from linear models to generalised ones [106], there is a lack of guidance in the literature on best practices of applying SLOPE models when the input matrix is not orthogonal.

In Chapter 7, we aim to overcome this limitation by developing Bayesian methodology. The methods were found to control the FDR in experiments under the majority of general settings considered. However, additional theoretical development is needed to better understand the FDR properties of our Bayesian models, as well as extension beyond the linear model case.

### 8.2.2 Additional Complexity of Sparse-group Models

Sparse-group models, due to their two layers of penalisation, are naturally more complex than single-norm models. This makes them relatively expensive computationally, although the screening rules developed in this thesis help mitigate this. In addition to this, sparse-group models have several other limitations and challenges, though some also represent opportunities to unlock improved model performance.

#### Tuning and Hyperparameters

Sparse-group models have additional hyperparameters compared to their single-norm counterparts. This makes them relatively difficult to implement and optimise, as setting values manually requires domain knowledge. However, the additional hyperparameters provide flexibility, often improving model performance when chosen to reflect practitioner knowledge or carefully tuned.

SGS has several hyperparameters that have to be determined: $\alpha, \lambda, q_v, q_g$. Tuning over a grid of all four is computationally infeasible. In practice, typically only $\lambda$ is tuned, as empirical evidence indicates that it has the greatest impact on model performance across different problems. The remaining parameters can be set to effective default values, $\alpha = 0.95$, $q_v = 0.1$, and $q_g = 0.1$, which have been shown to yield robust performance in most settings. In the thesis, SGS was generally found to perform well using $\alpha = 0.99$ on real data and in Section 3.6.5 we demonstrate how tuning the values of $q_v, q_g$ can lead to more stable prediction along a path.

The Bayesian SGS model learns $\lambda$ via the noise but has additional hyperparameters compared to the frequentist SGS due to its Bayesian priors. Again, effective defaults are found in this thesis and Section 7.5 shows that BSGS is not sensitive to

the choice of the prior hyperparameters. Additionally, BSGS-$\alpha$ learns $\alpha$, removing a manual step.

BSGS provides an excellent illustration of how practitioner knowledge can be used to change the hyperparameters to improve model performance. Under high-noise synthetic data, setting $d_1 = e_1 = 1$, $d_2 = e_2 = 1000$ encouraged sparsity to regain FDR control (Figure 7.13). Under real data, using $d_1 = e_1 = 5$, $d_2 = e_2 = 1$ produced denser models, capturing more signal and producing better predictive performance (Section 7.8).

#### Unknown Penalty Sequence Terms

The max sequences derived to control the FDR for SGS in Theorem 3.1 rely on several simplifying assumptions to allow their calculation in practice. First, the quantity $a_j$, representing the number of active variables within an active group, is unknown and is estimated as $\hat{a}_j = \lfloor \alpha p_j \rfloor$, which was found to be a reasonable choice (Section 3.4.3). Additionally, computing the group sequences requires the term $\sum_{k \in G_j} v_k$. Since the exact mapping of variable penalties to variables is unknown, group size information is used to assign the largest penalties $\{v_1, \dots, v_{p_j}\}$ to the highest-ranking groups with the largest $\sqrt{p_j}\|\boldsymbol{\beta}^{(j)}\|_2$ values. Finally, because the sequences depend on each other, they must be initialised; in this thesis, we use the SLOPE BH sequence for that purpose.

### 8.2.3 Screening Rules Lipschitz Assumptions

Our screening rules derived in Chapters 4 and 5, as any strong rule, rely on assumptions. Lipschitz assumptions were used that are consistent with those used in the strong screening framework. While such assumptions are a limitation of any strong rule, the sparse-group rules introduce additional assumptions beyond standard strong rules to enable a second layer of screening.

Violations of the assumptions are checked for using the KKT conditions. In Chapter 4, the SGS KKT checks caused an increased number of violations, as the checks were overly conservative. However, the overall amount of KKT violations for both models was still relatively small, suggesting that these assumptions are not overly restrictive. For DFR in Chapter 5, across all simulations, only a single KKT violation was observed for SGL, and violations occurred only rarely for aSGL.

For SGS, an attempt was made to derive less conservative checks (Appendix

C.2.2), performed directly on the variables without an initial group check. These were found to be too lenient, incorrectly missing violations.

### 8.2.4 Non-overlapping Groups

In this thesis, we assume that the grouping structure contains only non-overlapping groups. This is a fairly common situation in practice, with the majority of group-based regression developments in the literature assuming the same. Extending our methodology to overlapping groups requires modifying the corresponding norm to account for this, as can be seen by the overlapping group lasso [150, 226]. As such, this was beyond the scope of the thesis.

A practical way to deal with overlapping groups is to duplicate variables across all groups they belong to [191]. As shown in Section 3.6, this approach allowed SGS to outperform the lasso and SLOPE in disease prediction tasks. However, it may cause contradictory inference if a variable is active in one group but not another. Percival [150] propose a more refined method, decomposing variables to share effects across groups.

## 8.3 Future Work

The methods developed in this thesis open several avenues for future research, potentially enhancing their capabilities and addressing current limitations. Since many of our developments represent early-stage approaches to specific problems, here we mention key future directions, though we anticipate numerous additional research avenues beyond those discussed. Some of the directions aim to tackle the limitations discussed in Section 8.2.

### 8.3.1 Screening Rules

As noted in the limitations, the SGS KKT checks were overly conservative, while less conservative variants (Appendix C.2.2) proved too lenient. Future work may explore alternative KKT checks that achieve a more appropriate balance.

An additional future direction includes the development of safe rules, which guarantee that only inactive variables are discarded. These could be incorporated into a hybrid scheme together with our strong rules [202, 231]. Deriving safe rules would facilitate a comparison between them and our proposed strong rules, offering

further insight into which type of screening is most effective for both non-separable and sparse-group norms.

### 8.3.2 New Optimisation Approach for SGS

SGS is currently fit using ATOS, a proximal algorithm that makes use of the proximal operators of SLOPE and gSLOPE. As observed in Larsson et al. [112], proximal algorithms can scale poorly in high dimensions for SLOPE problems. Accordingly, the authors developed a coordinate descent algorithm for SLOPE using the directional derivative and the SLOPE thresholding operator, with convergence guarantees that account for the problem's non-separability. As observed in Remark 4.1, we discovered a relation between SGS and SGL using the soft-thresholding operator. Building on the development of the coordinate descent algorithm for SLOPE and utilising this connection, a coordinate descent algorithm for SGS could be developed, potentially similar to the BCD algorithm used to solve SGL [141]. This, together with the screening rules developed in this thesis (which are compatible with any optimisation algorithm), would substantially improve the efficiency of SGS, thereby removing computational barriers to its application.

### 8.3.3 Sparse-group and SLOPE Knockoffs

Proving FDR control under general settings has remained elusive for SLOPE models. In contrast, the Knockoffs framework offers exciting theoretical guarantees on FDR control without requiring orthogonality. Knockoffs have been extended to group settings [44, 46], but there is no literature on sparse-group developments. In Dai and Barber [46], group Knockoffs is performed by reformulating the multi-task regression problem as a group lasso problem. This suggests the possibility of developing sparse-group Knockoffs within a multi-task framework, similar to Lozano and Swirszcz [126]. The literature on applying SLOPE with Knockoffs is extremely sparse.

In Chapter 7, we briefly explored and implemented Knockoffs with SLOPE models, including SGS. Our results showed that Knockoffs allowed SLOPE models to control the FDR under many settings, but often with a loss of power. Future work could explore the integration of the Knockoffs framework with SLOPE and sparse-group models. For both, a key open question concerns the design of appropriate knockoff variables. Although a vast body of literature exists on various knockoff constructions, there are currently no established approaches tailored specifically to

SLOPE or sparse-group models. In these cases, the design would need to account for SLOPE's clustering property and the dual layers of penalisation in sparse-group models. The coin flip property also warrants further investigation, particularly for sparse-group models, where knockoff variables can be placed either within the original groups or in separate ones.

### 8.3.4 Automatic Sparse-group Models

Applying sparse-group models requires additional domain and practitioner knowledge over a simpler model, such as the lasso. Two key components need to be specified: a value of $\alpha$ and a grouping structure. Future work should consider approaches for allowing sparse-group models to be applied in an automatic pipeline so that they can be applied as readily as the lasso.

To select the value of $\alpha$, a practitioner can use a grid search over $(\lambda, \alpha)$, although this is expensive. BSGS-$\alpha$ provides a useful alternative to this in learning the value of $\alpha$ alongside the noise and regression parameters and was shown to achieve very good selection performance. The BSGS-$\alpha$ model can be easily adapted for any sparse-group model by changing the weights. Alternative future directions include exploring bi-level [222, 235] or Bayesian optimisation [75, 145] frameworks with FDR control in mind, as well as drawing inspiration from the elastic net literature [11].

Most of the real data applications in this thesis assigned variables to groups using natural or provided grouping information. For example, for the genetics data, the genes were placed into pathways. However, we also used real datasets for which this was not the case and for which the variables had to be placed in groups before the group-based methods could be applied. In this thesis, this was done using K-means or Singular Value Decomposition, chosen for their simplicity (Appendix A.3.3). Future work should consider other approaches for assigning variables to groups in an algorithmic fashion. Potential approaches include learning the group structure as part of model training or performing feature grouping during exploratory data analysis using clustering-based methods [123, 208].

Resolving these two issues would make sparse-group models viable competitors to the lasso for any problem, not just those with natural grouping structures.

### 8.3.5 Model Selection Approaches

In Chapter 7, we found that the Bayesian SLOPE models and TSO achieved the best performance in controlling the FDR while maintaining high power compared to competing methods. Future work around model selection for SLOPE should focus on new advances for these two methods.

Another direction for future work on model selection with SLOPE models is to develop selection criteria or loss functions for CV that optimise FDR control rather than predictive performance, which requires FDR estimation (see Luo et al. [127], Stoica and Babu [182] for potential approaches). This, in essence, is what Knockoffs aims to achieve.

#### Bayesian Models

In our synthetic data study in Section 7.7, the Bayesian methods were found to control the FDR under most general settings, using the sequences designed for the frequentist approaches. Future work should consider the development of new sequences designed specifically for the Bayesian models that take the weighted input into account.

Another future direction for Bayesian models is to consider whether the FDR control mechanism should come through the thresholding of the inclusion probabilities, rather than through the underlying model and penalty sequences; this is explored in Zhang et al. [233], where the thresholding is based on classical Bayesian utility considerations. In this thesis, we have applied median thresholding instead.

#### TSO and Gaussian Sequences

TSO used the lasso in its first step to reduce the input dimensionality below $p$. The lasso was chosen for its computational simplicity and its property of selecting fewer than $n$ variables, which is desirable in high-dimensional settings. Future work could explore alternative methods that offer guarantees on retaining true signal variables.

In this thesis, SLOPE was applied using the BH sequence. While Bogdan et al. [24] introduced the Gaussian sequence for more general settings, it reduces to the lasso when $p > n$ and was therefore not used in the thesis until the development of TSO. For TSO-SLOPE, however, the Gaussian sequence replaced the BH sequence. This resulted in selection performance comparable to TSO-gSLOPE and TSO-SGS, but with substantially improved predictive accuracy, addressing the key limitation

observed for TSO in our studies. This strong performance highlights an important future direction for gSLOPE and SGS: developing analogous Gaussian sequences for these models.

### 8.3.6 Overlapping Group Norms

As noted in the limitations, handling overlapping groups would require developing new norms. Future work could extend the proposed methodology in this direction. This would be particularly promising for DFR, which could leverage existing overlapping group norms, such as the overlapping group lasso proposed in Yuan et al. [226], by deriving appropriate screening rules rather than new norms. Additionally, since DFR employs ATOS, which is highly flexible to the incorporation of new norms, such an implementation would be straightforward.

## 8.4 Concluding Remarks

This thesis advances the sparse-group literature by developing methods that simultaneously achieve sparsity at both the variable and group levels while controlling the false discovery rate. By introducing SGS, sparse-group screening approaches, and Bayesian methodology, this thesis provides statistically rigorous tools that are also computationally efficient and practically implementable. These contributions enable practitioners to model high-dimensional grouped data effectively, which is particularly valuable in genetics, where identifying disease-associated genes can improve diagnostics, guide treatment, and advance understanding of underlying biological mechanisms.

# Bibliography


[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1972.

[2] I. Ahmed, A.-L. Hartikainen, M.-R. Järvelin, and S. Richardson. False Discovery Rate Estimation for Stability Selection: Application to Genome-Wide Association Studies. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011.

[3] T. Ahmed and W. U. Bajwa. ExSIS: Extended sure independence screening for ultrahigh-dimensional linear models. *Signal Processing*, 159:33–48, 2019.

[4] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[5] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[6] A. Atamturk and A. Gomez. Safe screening rules for $\ell_0$-regression from perspective relaxations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 421–430. PMLR, 2020.

[7] F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

[8] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. In *Optimization for Machine Learning*. The MIT Press, 2011.

[9] R. Bai, V. Ročková, and E. I. George. *Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO*, pages 81–108. Chapman and Hall/CRC, 2021.

[10] R. Bai, G. E. Moran, J. L. Antonelli, Y. Chen, and M. R. Boland. Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models. *Journal of the American Statistical Association*, 117(537):184–197, 2022.

[11] M.-F. F. Balcan, M. Khodak, D. Sharma, and A. Talwalkar. Provably tuning the elastic net across instances. In *Advances in Neural Information Processing Systems*, volume 35, pages 27769–27782. Curran Associates Inc., 2022.

[12] R. Bao, B. Gu, and H. Huang. Fast OSCAR and OWL Regression via Safe Screening Rules. In *Proceedings of the 37th International Conference on Machine Learning*, pages 653–663. PMLR, 2020.

[13] R. Bao, Q. Lu, and Y. Zhang. Safe Screening Rules for Group SLOPE. In *Machine Learning and Knowledge Discovery in Databases, Research Track*, pages 3–19. Springer–Verlag, 2025.

[14] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5), 2015.

[15] R. F. Barber and E. J. Candès. A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47(5), 2019.

[16] R. F. Barber and A. Ramdas. The $p$-filter: Multilayer False Discovery Rate Control for Grouped Hypotheses. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1247–1268, 2017.

[17] R. F. Barber, E. Candes, L. Janson, E. Patterson, and M. Sesia. *knockoff: The Knockoff Filter for Controlled Variable Selection*, 2022. URL `CRAN.R-project.org/package=knockoff`. R package version 0.3.6.

[18] S. Bates, E. Candès, L. Janson, and W. Wang. Metropolized Knockoff Sampling. *Journal of the American Statistical Association*, 116(535):1413–1427, 2021.

[19] S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, 119(546):1434–1445, 2024.

[20] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2017.

[21] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[22] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995.

[23] A. Blain, A. R. Lobo, J. Linhart, B. Thirion, and P. Neuvial. When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs. *arXiv preprint arXiv:2407.06892*, 2025.

[24] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. SLOPE—Adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3), 2015.

[25] H. D. Bondell and B. J. Reich. Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR. *Biometrics*, 64(1):115–123, 2008.

[26] A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval. Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso. *IEEE Transactions on Signal Processing*, 63(19):5121–5132, 2015.

[27] S. Boyd. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2010.

[28] D. A. Brannan. *A First Course in Mathematical Analysis*. Cambridge University Press, 2006.

[29] R. P. Brent. *Algorithms for minimization without derivatives*. Dover, 2013.

[30] D. Brzyski, A. Gossmann, W. Su, and M. Bogdan. Group SLOPE – Adaptive Selection of Groups of Predictors. *Journal of the American Statistical Association*, 114(525):419–433, 2019.

[31] Z. Bu, J. Klusowski, C. Rush, and W. Su. Algorithmic analysis and statistical estimation of slope via approximate message passing. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates Inc., 2019.

[32] Z. Bu, J. M. Klusowski, C. Rush, and W. J. Su. Characterizing the SLOPE trade-off: A variational perspective and the Donoho-Tanner limit. *The Annals of Statistics*, 51(1):33–61, 2023.

[33] G. Buch, A. Schulz, I. Schmidtmann, K. Strauch, and P. S. Wild. Sparse Group Penalties for bi-level variable selection. *Biometrical Journal*, 66(4), 2024.

[34] M. E. Burczynski, R. L. Peterson, N. C. Twine, K. A. Zuberek, B. J. Brodeur, L. Casciotti, V. Maganti, P. S. Reddy, A. Strahs, F. Immermann, W. Spinelli, U. Schwertschlag, A. M. Slager, M. M. Cotreau, and A. J. Dorner. Molecular Classification of Crohn's Disease and Ulcerative Colitis Patients Using Transcriptional Profiles in Peripheral Blood Mononuclear Cells. *The Journal of Molecular Diagnostics*, 8(1):51–61, 2006.

[35] O. 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.

[36] M. Cai, M. Dai, J. Ming, H. Peng, J. Liu, and C. Yang. BIVAS: A Scalable Bayesian Method for Bi-Level Variable Selection With Applications. *Journal of Computational and Graphical Statistics*, 29(1):40–52, 2020.

[37] T. T. Cai and Y. Wu. Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory*, 60:2217–2232, 2014.

[38] E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018.

[39] M. Capanu, M. Giurcanu, C. B. Begg, and M. Gönen. Two-stage subsampling variable selection for sparse high-dimensional generalized linear models. *Statistical Methods in Medical Research*, 34(7):1504–1521, 2025.

[40] G. Casella, M. Ghosh, J. Gill, and M. Kyung. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

[41] R.-B. Chen, C.-H. Chu, S. Yuan, and Y. N. Wu. Bayesian Sparse Group Selection. *Journal of Computational and Graphical Statistics*, 25(3):665–683, 2016.

[42] D. Chételat, J. Lederer, and J. Salmon. Optimal two-step prediction in regression. *Electronic Journal of Statistics*, 11(1):2519–2546, 2017.

[43] D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated Lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317, 2021.

[44] B. B. Chu, J. Gu, Z. Chen, T. Morrison, E. Candès, Z. He, and C. Sabatti. Second-order group knockoffs with applications to genome-wide association studies. *Bioinformatics*, 40(10):btae580, 2024.

[45] K. Ciesielski. On Stefan Banach and some of his results. *Banach Journal of Mathematical Analysis*, 1:1–10, 2007.

[46] R. Dai and R. F. Barber. The knockoff filter for FDR control in group-sparse and multitask regression. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48, pages 1851–1859. JMLR, 2016.

[47] G. M. D'Angelo, D. Rao, and C. C. 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.

[48] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

[49] D. Davis and W. Yin. A three-operator splitting scheme and its optimization applications. *Set-Valued and Variational Analysis*, 25:829–858, 2017.

[50] B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, 27(1), 1999.

[51] L. H. Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284, 2014.

[52] X. Dupuis and P. Tardivel. The solution path of SLOPE. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 775–783. PMLR, 2024.

[53] R. Edgar, M. Domrachev, and A. E. Lash. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30:207–210, 2002.

[54] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[55] L. El Ghaoui, V. Viallon, and T. Rabbani. Safe feature elimination in sparse supervised learning. Technical Report UCB/EECS-2010-126, EECS Department, University of California, Berkeley, 2010.

[56] K. A. Ellsworth, B. W. Eckloff, L. Li, I. Moon, B. L. Fridley, G. D. Jenkins, E. Carlson, A. Brisbin, R. Abo, W. Bamlet, G. Petersen, E. D. Wieben, and L. Wang. Contribution of FKBP5 Genetic Variation to Gemcitabine Treatment and Survival in Pancreatic Adenocarcinoma. *PLoS ONE*, 8(8): e70216, 2013.

[57] C. Elvira and C. Herzet. Safe rules for the identification of zeros in the solutions of the SLOPE problem. *SIAM Journal on Mathematics of Data Science*, 5(1):147–173, 2021.

[58] M. Evangelou, D. J. Smyth, M. D. Fortune, O. S. Burren, N. M. Walker, H. Guo, S. Onengut-Gumuscu, W.-M. Chen, P. Concannon, S. S. Rich, J. A. Todd, and C. Wallace. A Method for Gene-Based Pathway Analysis Using Genomewide Association Study Summary Statistics Reveals Nine New Type 1 Diabetes Associations. *Genetic Epidemiology*, 38(8):661–670, 2014.

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

[60] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(5):849–911, 2008.

[61] J. Fan, S. Guo, and N. Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(1):37–65, 2012.

[62] K. Fang, X. Wang, S. Zhang, J. Zhu, and S. Ma. Bi-level variable selection via adaptive sparse group Lasso. *Journal of Statistical Computation and Simulation*, 85(13):2750–2760, 2015.

[63] O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the Lasso. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 333–342. JMLR, 2015.

[64] F. Feser. *sgs: Sparse-group SLOPE: adaptive bi-level selection with FDR control*, 2023. URL `CRAN.R-project.org/package=sgs`. R package version 0.3.9.

[65] F. Feser. *dfr: Dual feature reduction for SGL*, 2024. URL `CRAN.R-project.org/package=dfr`. R package version 0.1.6.

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

[67] F. Feser and M. Evangelou. Dual Feature Reduction for the Sparse-Group Lasso and its Adaptive Variant. In *NeurIPS 2024 Workshop on Optimization for Machine Learning (OPT)*, 2024. NeurIPS Workshop Paper, not peer-reviewed.

[68] F. Feser and M. Evangelou. Dual Feature Reduction for the Sparse-group Lasso and its Adaptive Variant. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 17068–17099. PMLR, 2025.

[69] F. Feser and M. Evangelou. Strong Screening Rules for Group-based SLOPE Models. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 352–360. PMLR, 2025.

[70] M. Figueiredo and R. Nowak. Ordered weighted $\ell_1$ regularized regression with strongly correlated covariates: Theoretical aspects. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 930–938. PMLR, 2016.

[71] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[72] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 2010.

[73] F. Frommlet, P. Szulc, F. König, and M. Bogdan. Selecting predictive biomarkers from genomic data. *PLOS ONE*, 17(6):1–21, 2022.

[74] P. Gablenz and C. Sabatti. Catch me if you can: signal localization with knockoff e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae042, 2024.

[75] R. Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.

[76] A. Gelman. Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, 48(3):432–435, 2006.

[77] E. I. George and R. E. McCulloch. Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, 88(423):881, 1993.

[78] P. Ghosh, X. Tang, M. Ghosh, and A. Chakrabarti. Asymptotic Properties of Bayes Risk of a General Class of Shrinkage Priors in Multiple Hypothesis Testing Under Sparsity. *Bayesian Analysis*, 11(3), 2016.

[79] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

[80] A. Gossmann, S. Cao, and Y.-P. Wang. Identification of Significant Genetic Variants via SLOPE, and its Extension to Group SLOPE. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 232–240. Association for Computing Machinery, 2015.

[81] P. Graczyk, U. Schneider, T. Skalski, and P. Tardivel. A Unified Framework for Pattern Recovery in Penalized and Thresholded Estimation and its Geometry. *arXiv preprint arXiv:2307.10158*, 2023.

[82] J. Gu, Z. Chen, and Z. He. Pinpointing Important Genetic Variants via A Feature-level Filter and Group Knockoffs. *arXiv preprint arXiv:2408.12618*, 2024.

[83] Y. Guo, C. Wu, M. Guo, Q. Zou, X. Liu, and A. Keinan. Combining Sparse Group Lasso and Linear Mixed Model Improves Power to Detect Genetic Variants Underlying Quantitative Traits. *Frontiers in Genetics*, 10, 2019.

[84] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition.* Springer New York, 2009.

[85] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning With Sparsity.* Chapman and Hall/CRC, 2015.

[86] G. A. Heap, G. Trynka, R. C. Jansen, M. Bruinenberg, M. A. Swertz, L. C. Dinesen, K. A. Hunt, C. Wijmenga, D. A. vanHeel, and L. Franke. Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Medical Genomics*, 2(1):1, 2009.

[87] I. Hejný, J. Wallin, M. Bogdan, and M. Kos. Unveiling low-dimensional patterns induced by convex non-differentiable regularizers. *Annals of the Institute of Statistical Mathematics*, 2025.

[88] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[89] D. Homrighausen and D. J. McDonald. A study on tuning parameter selection for the high-dimensional lasso. *Journal of Statistical Computation and Simulation*, 88(15):2865–2892, 2018.

[90] X. Hong, Y. Guo, S. Chen, and J. Gao. Sparse model construction using coordinate descent optimization. In *Proceedings: 18th International Conference on Digital Signal Processing (DSP2013)*, 2013.

[91] J. X. Hu, H. Zhao, and H. H. Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105:1215–1227, 2010.

[92] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

[93] M. Humayoo and X. Cheng. Model-free Knockoffs for SLOPE–Adaptive Variable Selection with Controlled False Discovery Rate. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 302–307. IEEE, 2018.

[94] D. Iacopetta, R. Lappano, A. R. Cappello, M. Madeo, E. M. D. Francesco, A. Santoro, et al. SLC37A1 gene expression is up-regulated by epidermal growth factor in breast cancer cells. *Breast Cancer Research and Treatment*, 122:755–764, 2010.

[95] Y. Ida, Y. Fujiwara, and H. Kashima. Fast Sparse Group Lasso. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates Inc., 2019.

[96] N. Ignatiadis, R. Wang, and A. Ramdas. Asymptotic and compound e-values: multiple testing and empirical Bayes. *arXiv preprint arXiv:2409.19812*, 2025.

[97] A. Javanmard and H. Javadi. False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1):1212–1253, 2019.

[98] J. Jia and B. Yu. On model selection consistency of the elastic net when $p \gg n$. *Statistica Sinica*, 20(2):595–611, 2010.

[99] W. Jiang, M. Bogdan, J. Josse, S. Majewski, B. Miasojedow, and V. Ročková. Adaptive Bayesian SLOPE: Model Selection With Incomplete Data. *Journal of Computational and Graphical Statistics*, 31(1):113–137, 2022.

[100] T. Johnson and C. Guestrin. Blitz: A principled meta-algorithm for scaling sparse optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1171–1179. PMLR, 2015.

[101] V. E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

[102] I. M. Johnstone and D. M. Titterington. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367:4237–4253, 2009.

[103] E. Katsevich and C. Sabatti. Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, 13(1), 2019.

[104] Y. Kim, W. Wang, P. Carbonetto, and M. Stephens. A flexible empirical Bayes approach to multiple linear regression, and connections with penalized regression. *J. Mach. Learn. Res.*, 25(1), 2024.

[105] S. Koner, N. De Sarkar, and N. Laha. False discovery rate control: Moving beyond the Benjamini–Hochberg method. *bioRxiv preprint bioRxiv:2024.01.13.575531*, 2024.

[106] M. Kos and M. Bogdan. On the asymptotic properties of SLOPE. *Sankhya A*, 82:499–532, 2020.

[107] A. Koussounadis, S. P. Langdon, D. J. Harrison, and V. A. Smith. Chemotherapy-induced dynamic gene expression changes in vivo are prognostic in ovarian cancer. *British Journal of Cancer*, 110:2975–2984, 2014.

[108] P. J. Kremer, S. Lee, M. Bogdan, and S. Paterlini. Sparse portfolio selection via the sorted $\ell_1$-norm. *Journal of Banking & Finance*, 110, 2020.

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

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

[111] J. Larsson, M. Bogdan, and J. Wallin. The strong screening rule for SLOPE. In *Advances in Neural Information Processing Systems*, volume 33, pages 14592–14603. Curran Associates Inc., 2020.

[112] J. Larsson, Q. Klopfenstein, M. Massias, and J. Wallin. Coordinate Descent for SLOPE. *Proceedings of Machine Learning Research*, 206:4802–4821, 2022.

[113] J. Larsson, J. Wallin, M. Bogdan, E. van den Berg, C. Sabatti, E. Candes, E. Patterson, W. Su, J. Kała, K. Grzesiak, and M. Burdukiewicz. *SLOPE: Sorted $\ell_1$ Penalized Estimation*, 2022. URL `CRAN.R-project.org/package=SLOPE`. R package version 0.5.0.

[114] M. Lavielle. *Mixed Effects Models for the Population Approach*. Chapman and Hall/CRC, 2014.

[115] J. Lederer and C. Müller. Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

[116] E. R. Lee, J. Cho, and K. Yu. A systematic review on model selection in high-dimensional regression. *Journal of the Korean Statistical Society*, 48(1): 1–12, 2019.

[117] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

[118] C. Leng, Y. Lin, and G. Wahba. A note on the lasso and the related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284, 2006.

[119] L. Li, J.-W. Zhang, G. Jenkins, F. Xie, E. E. Carlson, B. L. Fridley, W. R. Bamlet, G. M. Petersen, R. R. McWilliams, and L. Wang. Genetic variations associated with gemcitabine treatment outcome in pancreatic cancer. *Pharmacogenetics and Genomics*, 26(12):527–537, 2016.

[120] R. Li, C. Chang, J. M. Justesen, Y. Tanigawa, J. Qian, T. Hastie, M. A. Rivas, and R. Tibshirani. Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. *Biostatistics*, 23(2): 522–540, 2022.

[121] X. Liang, A. Cohen, A. Sólon Heinsfeld, F. Pestilli, and D. J. McDonald. sparsegl: An R package for estimating sparse group lasso. *Journal of Statistical Software*, 110(6):1–23, 2024.

[122] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[123] H. Liu, X. Wu, and S. Zhang. Feature selection using hierarchical feature clustering. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 979–984. Association for Computing Machinery, 2011.

[124] W. Liu, Y. Ke, J. Liu, and R. Li. Model-Free Feature Screening and FDR Control with Knockoff Features. *Journal of the American Statistical Association*, 117(537):428–443, 2022.

[125] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[126] A. C. Lozano and G. Swirszcz. Multi-level lasso for sparse multi-task regression. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 595–602. Omnipress, 2012.

[127] Y. Luo, W. Fithian, and L. Lei. Estimating the FDR of variable selection. *arXiv preprint arXiv:2408.07231*, 2024.

[128] X.-J. Ma, Z. Wang, P. D. Ryan, S. J. Isakoff, A. Barmettler, A. Fuller, B. Muir, G. Mohapatra, R. Salunga, J. Tuggle, Y. Tran, D. Tran, A. Tassin, P. Amon, W. Wang, W. Wang, E. Enright, K. Stecker, E. Estepa-Sabal, B. Smith, J. Younger, U. Balis, J. Michaelson, A. Bhan, K. Habin, T. M. Baer, J. Brugge, D. A. Haber, M. G. Erlander, and D. C. Sgroi. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell*, 5(6):607–616, 2004.

[129] L. Meier. *grplasso: Fitting User-Specified Models with Group Lasso Penalty*, 2020. URL `CRAN.R-project.org/package=grplasso`. R package version 0.4.7.

[130] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[131] N. Meinshausen and P. Bühlmann. Stability Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473, 2010.

[132] A. Mendez-Civieta, M. C. Aguilera-Morillo, and R. E. Lillo. Adaptive sparse group LASSO in quantile regression. *Advances in Data Analysis and Classification*, 15(3):547–573, 2021.

[133] T. J. Mitchell and J. J. Beauchamp. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, 83(404):1023, 1988.

[134] A. Moscovich and S. Rosset. On the cross-validation bias due to unsupervised preprocessing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1474–1502, 2022.

[135] National Cancer Institute. The Cancer Genome Atlas Program, 1988. URL `cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`.

[136] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*, volume 1, pages 811–819. MIT Press, 2015.

[137] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. GAP Safe Screening Rules for Sparse-Group Lasso. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates Inc., 2016.

[138] E. Ndiaye, O. Fercoq, A. Gramfort, and J. Salmon. GAP safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18, 2016.

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

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

[141] J. Nutini, I. Laradji, and M. Schmidt. Let's make block coordinate descent converge faster: faster greedy rules, message-passing, active-set complexity, and superlinear convergence. In *Journal of Machine Learning Research*, volume 23. JMLR, 2022.

[142] K. Ogawa, Y. Suzuki, and I. Takeuchi. Safe screening of non-support vectors in pathwise SVM computation. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1382–1390. PMLR, 2013.

[143] U. Oswal, C. Cox, M. A. L. Ralph, T. Rogers, and R. Nowak. Representational similarity learning with application to brain networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, volume 48, pages 1041–1049. JMLR, 2016.

[144] Y. Pan. Feature screening and FDR control with knockoff features for ultrahigh-dimensional right-censored data. *Computational Statistics & Data Analysis*, 173:107504, 2022.

[145] L. Papenmeier, M. Poloczek, and L. Nardi. Understanding high-dimensional Bayesian optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 47902–47923. PMLR, 2025.

[146] N. Parikh. Proximal Algorithms. *Foundations and Trends in Optimization*, 1 (3):127–239, 2014.

[147] F. Pedregosa and G. 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.

[148] H. Pei, L. Li, B. L. Fridley, G. D. Jenkins, K. R. Kalari, W. Lingle, G. Petersen, Z. Lou, and L. Wang. FKBP51 Affects Cancer Cell Response to Chemotherapy by Negatively Regulating Akt. *Cancer Cell*, 16(3):259–266, 2009.

[149] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack, and P. 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.

[150] D. Percival. Theoretical properties of the overlapping groups lasso. *Electronic Journal of Statistics*, 6:269–288, 2012.

[151] B. Poignard. Asymptotic theory of the adaptive Sparse Group Lasso. *Annals of the Institute of Statistical Mathematics*, 72(1):297–328, 2020.

[152] Z. Qin, K. Scheinberg, and D. Goldfarb. Efficient block-coordinate descent algorithms for the Group Lasso. *Mathematical Programming Computation*, 5: 143–169, 2013.

[153] D. G. Rasines and G. A. Young. Splitting strategies for post-selection inference. *Biometrika*, 110(3):597–614, 2022.

[154] S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, 26(1):35–67, 2016.

[155] Z. Ren and R. F. Barber. Derandomised knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):122–154, 2024.

[156] Z. Ren, Y. Wei, and E. Candès. Derandomizing Knockoffs. *Journal of the American Statistical Association*, 118(542):948–958, 2023.

[157] S.-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, 2006.

[158] R. Riccobello, G. Bonaccolto, P. J. Kremer, S. Paterlini, and M. Bogdan. Sparse graphical modelling for minimum variance portfolios. *SSRN Electronic Journal*, 2023.

[159] S. Roberts and G. Nowak. Stabilizing the lasso against cross-validation variability. *Computational Statistics & Data Analysis*, 70:198–211, 2014.

[160] V. Ročková and E. I. George. EMVS: The EM Approach to Bayesian Variable Selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014.

[161] V. Ročková and E. I. George. Bayesian penalty mixing: The case of a non-separable penalty. In *Statistical Analysis for High-Dimensional Data*, pages 233–254. Springer International Publishing, 2016.

[162] V. Ročková and E. I. George. The Spike-and-Slab LASSO. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

[163] S. Roman. *Advanced Linear Algebra*. Springer New York, 2008.

[164] Y. Romano, M. Sesia, and E. Candès. Deep Knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872, 2020.

[165] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1986.

[166] J. Sabates-Bellver, L. G. Van der Flier, M. de Palo, E. Cattaneo, C. Maake, H. Rehrauer, E. Laczko, M. A. Kurowski, J. M. Bujnicki, M. Menigatti, J. Luz, T. V. Ranalli, V. Gomes, A. Pastorelli, R. Faggiani, M. Anti, J. Jiricny, H. Clevers, and G. Marra. Transcriptome Profile of Human Colorectal Adenomas. *Molecular Cancer Research*, 5(12):1263–1275, 2007.

[167] S. Sadik, M. Et-Tolba, and B. Nsiri. A modified adaptive sparse-group lasso regularization for optimal portfolio selection. *IEEE Access*, 12:107337–107352, 2024. doi: 10.1109/ACCESS.2024.3438125.

[168] J. A. Salomon, A. Reinhart, A. Bilinski, E. J. Chua, W. La Motte-Kerr, M. M. Rönn, M. B. Reitsma, K. A. Morris, S. LaRocca, T. H. Farag, F. Kreuter, R. Rosenfeld, and R. 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.

[169] S. K. Sarkar. Two-stage stepup procedures controlling FDR. *Journal of Statistical Planning and Inference*, 138(4):1072–1084, 2008.

[170] T. E. Scheetz, K.-Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A. Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, T. L. Casavant, V. C. Sheffield, and E. 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.

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

[172] G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

[173] A. Sepehri. The Bayesian SLOPE. *arXiv preprint arXiv:1608.08968*, 2016.

[174] M. Sesia, C. Sabatti, and E. J. Candès. Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1):1–18, 2019.

[175] R. D. Shah and R. J. Samworth. Variable selection with error control: Another look at stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(1):55–80, 2013.

[176] A. Shibagaki, M. Karasuyama, K. Hatano, and I. Takeuchi. Simultaneous safe screening of features and samples in doubly sparse modeling. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1577–1586. PMLR, 2016.

[177] N. Simon, J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(i05), 2011.

[178] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[179] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. *SGL: Fit a GLM (or Cox Model) With a Combination of Lasso and Group Lasso Regularization*, 2019. URL `CRAN.R-project.org/package=SGL`. R package version 1.3.

[180] A. D. Skol, M. M. Sasaki, and K. Onel. The genetics of breast cancer risk in the post-genome era: Thoughts on study design to move past BRCA and towards clinical relevance. *Breast Cancer Research*, 18:99, 2016.

[181] E. Steiger and M. Vingron. Sparse-Group Bayesian Feature Selection Using Expectation Propagation for Signal Recovery and Network Reconstruction. *arXiv preprint arXiv:1809.09367*, 2018.

[182] P. Stoica and P. Babu. False Discovery Rate (FDR) and Familywise Error Rate (FER) Rules for Model Selection in Signal Processing Applications. *IEEE Open Journal of Signal Processing*, 3:403–416, 2022.

[183] P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.

[184] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 36 (2):111–133, 2018.

[185] J. D. Struik. *A Source Book in Mathematics, 1200-1800*. Princeton University Press, 1986.

[186] W. Su and E. Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

[187] W. Su, M. Bogdan, and E. Candès. False discoveries occur early on the Lasso path. *The Annals of Statistics*, 45(5):2133–2150, 2017.

[188] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[189] T. Sun and C.-H. Zhang. Scaled Sparse Linear Regression. *Biometrika*, 99(4): 879–898, 2012.

[190] W. Sun, J. Wang, and Y. Fang. Consistent selection of tuning parameters via variable selection stability. *Journal of Machine Learning Research*, 14: 3419–3440, 2013.

[191] Z. Tang, Y. Shen, Y. Li, X. Zhang, J. Wen, C. Qian, W. Zhuang, X. Shi, and N. Yi. Group spike-and-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information. *Bioinformatics*, 34:901–910, 2018.

[192] R. Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[193] R. Tibshirani, J. Bien, J. Friedman, T. Hastie, N. Simon, J. Taylor, and R. 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.

[194] R. J. Tibshirani. The Lasso Problem and Uniqueness. *arXiv preprint arXiv:1206.0313*, 2012.

[195] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[196] 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.

[197] A. Virouleau, A. Guilloux, S. Gaïffas, and M. Bogdan. High-dimensional robust regression and outliers detection with SLOPE. *arXiv preprint arXiv:1712.02640*, 2017.

[198] V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

[199] K. Šehić, A. Gramfort, J. Salmon, and L. Nardi. LassoBench: A High-Dimensional Hyperparameter Optimization Benchmark Suite for Lasso. In *Proceedings of the First International Conference on Automated Machine Learning*, volume 188 of *Proceedings of Machine Learning Research*, pages 1–24. PMLR, 2022.

[200] H. Wagner and C. Duller. Bayesian model selection for logistic regression models with random intercept. *Computational Statistics & Data Analysis*, 56 (5):1256–1274, 2012.

[201] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[202] C. Wang and P. Breheny. Adaptive hybrid screening for efficient lasso optimization. *Journal of Statistical Computation and Simulation*, 92(11):2233–2256, 2022.

[203] H. Wang, Y. Yang, Z. Bu, and W. Su. The complete lasso tradeoff diagram. In *Advances in Neural Information Processing Systems*, volume 33, pages 20051–20060. Curran Associates Inc., 2020.

[204] J. Wang, J. Zhou, P. Wonka, and J. Ye. Lasso screening rules via dual Polytope Projection. In *Proceedings of the 26th International Conference on Neural In-*

*formation Processing Systems*, volume 1, pages 1070–1078. Curran Associates Inc., 2013.

[205] J. Wang, J. Zhou, J. Liu, P. Wonka, and J. Ye. A Safe Screening Rule for Sparse Logistic Regression. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates Inc., 2014.

[206] J. Wang, Z. Zhang, and J. Ye. Two-layer feature reduction for sparse-group lasso via decomposition of convex sets. *Journal of Machine Learning Research*, 20(163):1–42, 2019.

[207] L. Wang, B. Peng, J. Bradic, R. Li, and Y. Wu. A tuning-free robust and efficient approach to high-dimensional regression. *Journal of the American Statistical Association*, 115(532):1700–1714, 2020.

[208] P. Wang, B. Xue, J. Liang, and M. Zhang. Feature clustering-assisted feature selection with differential evolution. *Pattern Recognition*, 140:109523, 2023.

[209] R. Wang and A. Ramdas. False Discovery Rate Control with E-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 2022.

[210] L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000.

[211] L. Wasserman and K. Roeder. High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009.

[212] A. Weinstein, W. J. Su, M. Bogdan, R. F. Barber, and E. J. Candès. A power analysis for model-X knockoffs with $\ell_p$-regularized statistics. *The Annals of Statistics*, 51(3):1005–1029, 2023.

[213] Y. Wu and L. Wang. A survey of tuning parameter selection for high-dimensional regression. *Annual Review of Statistics and Its Application*, 7: 209–226, 2020.

[214] Z. Xiang, H. Xu, and P. J. Ramadge. Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates Inc., 2011.

[215] Z. J. Xiang and P. J. Ramadge. Fast lasso screening tests based on correlations. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2137–2140, 2012.

[216] G. Xu, X. Yan, J. Chen, X. Guo, X. Guo, Y. Tang, and Z. Shi. Bioinformatics analysis of key candidate genes and pathways in ulcerative colitis. *Biological and Pharmaceutical Bulletin*, 43(11):1760–1766, 2020.

[217] X. Xu and M. Ghosh. Bayesian Variable Selection and Estimation for Group Lasso. *Bayesian Analysis*, 10(4):909–936, 2015.

[218] C. Yang, X. Wan, Q. Yang, H. Xue, and W. Yu. Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. *BMC Bioinformatics*, 11(S1):S18, 2010.

[219] S. Yang, J. Wen, X. Zhan, and D. Kifer. ET-Lasso: A New Efficient Tuning of Lasso-type Regularization for High-Dimensional Data. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 607–616, 2019.

[220] S. Yang, S. Chen, P. Wang, A. Chen, and T. Tian. TSPLASSO: A two-stage prior lasso algorithm for gene selection using omics data. *IEEE Journal of Biomedical and Health Informatics*, 28(1):526–537, 2024.

[221] Y. Yang. Can the Strengths of AIC and BIC be Shared? A Conflict Between Model Indentification and Regression Estimation. *Biometrika*, 92:937–950, 2005.

[222] Y. Yang, P. Xiao, S. Ma, and K. Ji. First-order federated bilevel learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22029–22037, 2025.

[223] D. Yogatama and N. A. Smith. Linguistic Structured Sparsity in Text Categorization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 786–796. Association for Computational Linguistics, 2014.

[224] Y. Yu and Y. Feng. Modified cross-validation for penalized high-dimensional linear regression models. *Journal of Computational and Graphical Statistics*, 23(4):1009–1027, 2014.

[225] Z. Yu, E. Jiang, X. Wang, Y. Shi, A. J. Shangguan, L. Zhang, and J. Li. Sushi domain-containing protein 3: A potential target for breast cancer. *Cell Biochemistry and Biophysics*, 72:321–324, 2015.

[226] L. Yuan, J. Liu, and J. Ye. Efficient methods for overlapping group lasso. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates Inc., 2011.

[227] M. Yuan and Y. 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.

[228] N. Zemlianskaia, W. J. Gauderman, and J. P. Lewinger. A Scalable Hierarchical Lasso for Gene-Environment Interactions. *Journal of Computational and Graphical Statistics*, 31(4):1091–1103, 2022.

[229] X. Zeng and M. A. T. Figueiredo. The atomic norm formulation of OSCAR regularization with application to the Frank-Wolfe algorithm. In *2014 22nd European Signal Processing Conference*, pages 780–784, 2014.

[230] X. Zeng and M. A. T. Figueiredo. The Ordered Weighted $\ell_1$ Norm: Atomic Formulation, Projections, and Algorithms. *arXiv preprint arXiv:1409.4271*, 2015.

[231] Y. Zeng, T. Yang, and P. Breheny. Hybrid safe-strong rules for efficient optimization in lasso-type problems. *Computational Statistics & Data Analysis*, 153, 2021.

[232] J. Zhang, X. J. Jeng, and H. Liu. Some two-step procedures for variable selection in high-dimensional linear regression. *arXiv preprint arXiv:0810.1644*, 2008.

[233] L. Zhang, V. Baladandayuthapani, B. K. Mallick, G. C. Manyam, P. A. Thompson, M. L. Bondy, and K.-A. Do. Bayesian Hierarchical Structured Variable Selection Methods with Application to Molecular Inversion Probe Studies in Breast Cancer. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(4):595–620, 2014.

[234] W. Zhang, B. Hong, W. Liu, J. Ye, D. Cai, X. He, and J. Wang. Scaling Up Sparse Support Vector Machines by Simultaneous Feature and Sample

Reduction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4016–4025. PMLR, 2017.

[235] Y. Zhang, P. Khanduri, I. Tsaknakis, Y. Yao, M. Hong, and S. Liu. An introduction to bilevel optimization: Foundations and applications in signal processing and machine learning. *IEEE Signal Processing Magazine*, 41(1): 38–59, 2024.

[236] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(90):2541–2563, 2006.

[237] Y. Zhao and X. Huo. A survey of numerical algorithms that can solve the lasso problems. *WIREs Computational Statistics*, 15(4), 2023.

[238] L. W. Zhong and J. T. Kwok. Efficient sparse modeling with automatic feature grouping. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 9–16. Omnipress, 2011.

[239] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

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

# APPENDIX A

# General

## A.1 Notation

| Symbol | Definition |
|---|---|
| **Data** | |
| $p, n$ | Number of factors/observations |
| $[p]$ | Set of indices up to $p$: $\{1, \dots, p\}$ |
| $\mathbf{X}, \mathbf{y}$ | $p \times n$ input data matrix and $n$-length response vector |
| $\boldsymbol{\beta}$ | Model coefficients |
| $\epsilon$ | Residuals |
| $\mathcal{G}$ | Grouping m-partition of the input space |
| $G_j$ | Variable indexes for group $j$ |
| $p_j$ | Group size of group $j$ |
| $S_v, S_g$ | Signal variable/group set |
| **Model** | |
| $\lambda$ | Regularisation parameter |
| $\alpha$ | Convex value in $[0, 1]$ defining the balance between variable and group penalisation |
| $f$ | A general loss function |
| $\nabla$, $\partial$ | Gradient and subdifferential |
| $\mathbf{v}, \mathbf{w}$ | Vector of variable/group penalty weights |
| $q_v, q_g$ | Variable/group FDR control parameter |
| **Other** | |
| $\Phi$ | CDF of a standard Gaussian |
| $\|\cdot\|_p$, $J(\cdot)$ | p-norm function and norm defining a model |
| $\oslash$ | Hadamard division |
| $x_{[t]}$ | Value of variable $x$ at iteration $t$ in an algorithm |
| $\text{prox}_f$ | Proximal mapping for function $f$ |
| $r_v, r_g$ | Variable/group rank functions which give the rank of the input in decreasing order |
| $b_{(1)} \geq \cdots \geq b_{(p)}$ | Order statistics of the absolute coefficients, used for the SLOPE norm |

Table A.1: General notation used in the thesis.

| Symbol | Definition |
|---|---|
| **Chapter 3** | |
| $a_i$ | Number of active variables in an active group $i$ |
| $F_{\text{FN}}$, $F_{\chi_k}$ | CDFs of a folded Gaussian and a $\chi$ distribution with $k$ degrees of freedom |
| $\gamma$ | Step size in ATOS |
| $\eta$ | Backtracking hyperparameter in ATOS |
| **Chapters 4 and 5** | |
| $\mathcal{S}_v, \mathcal{S}_g$ | Screened variable/group set |
| $\mathcal{E}_v, \mathcal{E}_g$ | Optimisation variable/group set |
| $\mathcal{A}_v, \mathcal{A}_g$ | Active variable/group set |
| $(\cdot)_\downarrow$ and $(\cdot)_{\lvert\downarrow\rvert}$ | Sort a vector into decreasing and decreasing absolute form |
| $\preceq$ | Element-wise inequality |
| cumsum$(\cdot)$ | Cumulative sum operator |
| card$(\cdot)$ | Cardinality of a vector |
| $(\cdot)_+ = \max(x, 0)$ | Positive part function |
| **Chapter 7** | |
| $A$, $\tilde{A}$, $\hat{A}$ | Weight matrices for SLOPE, gSLOPE, and SGS with entries $a$, $\tilde{a}$, and $\hat{a}$ |
| $\boldsymbol{\delta}$, $\boldsymbol{\gamma}$ | Variable/group inclusion parameter |
| $\theta_v$, $\theta_g$ | Variable/group mixture parameter |
| $c_v$, $c_g$ | Variable/group signal strength ratio |
| $\mathbf{x}_{-j}$ | Vector $\mathbf{x}$ with the $j$th entry removed |
| $\pi$ | Probability distribution |
| $\mathbb{P}$ | Probability measure |

Table A.2: Chapter-specific notation.

| Symbol | Definition |
|---|---|
| **General** | |
| ATOS | Adaptive Three Operator Splitting |
| BCD | Blockwise Coordinate Descent |
| BH | Benjamini-Hochberg |
| CV | Cross-Validation |
| FDR | False Discovery Rate |
| gLasso/gOSCAR/gSLOPE | Group Lasso/OSCAR/SLOPE |
| MAE | Mean Absolute Error |
| MSE | Mean Squared Error |
| OSCAR | Octagonal Shrinkage and Clustering Algorithm for Regression |
| OWL | Ordered Weighted $\ell_1$ |
| SGL/SGO/SGS | Sparse-group Lasso/OSCAR/SLOPE |
| SLOPE | Sorted $\ell_1$ Penalised Estimation |
| **Chapters 4 and 5** | |
| aSGL | Adaptive Sparse-group Lasso |
| DFR | Dual Feature Reduction |
| KKT | Karush–Kuhn–Tucker |
| SAFE | Safe Feature Elimination |
| **Chapter 7** | |
| ABSLOPE | Adaptive Bayesian SLOPE |
| AS-SGS | Adaptively Scaled SGS |
| BGSLOPE | Bayesian Group SLOPE |
| BSGS | Bayesian Sparse-group SLOPE |
| SAEM | Stochastic Approximation Expectation-maximisation |
| TSO | Two-step Orthogonal |
| TS-SLOPE | Two-step SLOPE |

Table A.3: Common acronyms used in the thesis.

## A.2 Definitions

**Definition A.1** (Power)**.** Let $\hat{S}_v \subseteq [p]$ be the estimated support. Then, the *Power* is defined as

$$\text{Power} = \mathbb{E}\left[\frac{|\hat{S}_v \cap S_v|}{\max(1, |S_v|)}\right].$$

The numerator counts the number of true discoveries (*i.e.*, correctly selected variables), so that the power quantifies the expected proportion of true signals that are recovered. The power takes values in the range $[0, 1]$.

**Definition A.2** ($\text{F}_1$ Score)**.** Let $\hat{S}_v \subseteq [p]$ be the estimated support. Then, the $F_1$ *Score* is defined as the harmonic mean of precision and recall:

$$\text{F}_1 = \mathbb{E}\left[\frac{2|\hat{S}_v \cap S_v|}{|\hat{S}_v| + |S_v|}\right].$$

This metric balances false positives and false negatives, providing a single measure of selection accuracy. It takes values in the range $[0, 1]$.

## A.3 Computational Information

### A.3.1 Computing Resources

The synthetic data experiments were executed on a high-performance computing cluster (x86-64 Linux GNU) and the real data experiments were conducted on a Apple Macbook Air (M1, 8GB).

### A.3.2 Hyperparameters

| Parameter | Synthetic Data | Real Data |
|---|---|---|
| **ATOS** | | |
| Max iterations | 5000 | 10000 (5000 in Chapter 7) |
| Backtracking | 0.7 | 0.7 |
| Max backtracking iterations | 100 | 100 |
| Convergence tolerance | $10^{-5}$ | $10^{-5}$ |
| Convergence criteria | $\|x - z\|_2$ | $\|x - z\|_2$ |
| Standardisation | $\ell_2$ | $\ell_2$ |
| Warm starts | Yes | Yes |
| Intercept | Yes for linear only | Yes for linear only |
| **SAEM** | | |
| Maximum iterations | 500 | 500 |
| Convergence tolerance | $10^{-5}$ | $10^{-5}$ |
| Standardisation | $\ell_2$ | $\ell_2$ |

Table A.4: Algorithmic hyperparameters used for running ATOS in the synthetic and real data experiments in Chapters 3, 4, 5, and 7, and for running SAEM in Chapter 7. The BCD algorithm in Chapter 5 uses the same hyperparameters as ATOS.

| Parameter | Chapter | | | |
|---|---|---|---|---|
| | 3 | 4 | 5 | 7 |
| $p$ | 800 | Varied | 1000 | 500 |
| $n$ | 200 | 400 | 200 | 400 |
| $m$ (uneven cases) | 160 | Varied | 22 | 37 |
| $m$ (even cases) | – | – | 50 | Varied |
| Group sizes (uneven cases) | $[3, 7]$ | $[3, 25]$ | $[3, 100]$ | $[5, 25]$ |
| Group sizes (even cases) | – | – | 20 | Varied |
| Signal $(\beta_i, i \in S_v)$ | 5 | $\mathcal{N}(0, 5)$ | $\mathcal{N}(0, 4)$ | $\mathcal{N}(5, 10)$ |
| Variable sparsity | Varied | 0.7 | 0.8 | 0.7 |
| Group sparsity | Varied | 0.85 | 0.8 | 0.8 |
| Correlation $(\rho)$ | Varied | Varied | 0.3 | $\rho_w = 0.3, \rho_a = 0$ |
| Residuals $(\epsilon_i, i \in [p])$ | Adaptive | $\mathcal{N}(0, 1)$ | $\mathcal{N}(0, 1)$ | $\mathcal{N}(0, 1)$ |

Table A.5: Base case data hyperparameters used in the main synthetic experiments.

| Parameter | Synthetic Data | Real Data |
|---|---|---|
| **Chapter 3** | | |
| $\alpha$ | 0.95 | 0.99 |
| $q_v$ | 0.1 | $10^{-4}$ (Colitis), $10^{-5}$ (Cancer) |
| $q_g$ | 0.1 | $10^{-10}$ |
| Path length ($l$) | 20 | 100 |
| Path termination ($\lambda_l$) | $0.15\lambda_1$ | $0.1\lambda_1$ (Colitis), $0.01\lambda_1$ (Cancer) |
| **Chapter 4** | | |
| $\alpha$ | 0.95 | 0.99 |
| $q_v$ | 0.05 | 0.01 |
| $q_g$ | 0.05 | 0.01 |
| Path length ($l$) | 50 | 100 |
| Path termination ($\lambda_l$) | $0.05\lambda_1$ | $0.01\lambda_1$ |
| **Chapter 5** | | |
| $\alpha$ | 0.95 | 0.99 |
| $b_1 = b_2$ | 0.1 | 0.1 |
| Path length ($l$) | 50 | 100 |
| Path termination ($\lambda_l$) | $0.1\lambda_1$ | $0.2\lambda_1$ |
| **Chapter 7** | | |
| $\alpha$ (SGS only) | 0.95 | 0.99 |
| $q_v$ | 0.1 | 0.1 |
| $q_g$ | 0.1 | 0.1 |
| $\boldsymbol{\beta}$ initialisation (Bayesian only) | Lasso | Lasso |
| $\theta$ Beta prior (ABSLOPE) | $d_1 = d_2 = 0.01n$ | $d_1 = d_2 = 0.01n$ |
| $\theta$ Beta prior (BGSLOPE) | $d_1 = d_2 = 0.01n$ | $d_1 = d_2 = 0.01n$ |
| $\theta_g$ Beta prior (BSGS) | $d_1 = 0.003n, d_2 = 0.015n$ | $d_1 = 5, d_2 = 1$ |
| $\theta_v$ Beta prior (BSGS) | $e_1 = 0.003n, e_2 = 0.015n$ | $e_1 = 5, e_2 = 1$ |
| Path length ($l$) | 20 | 20 |
| Path termination ($\lambda_l$) | $0.1\lambda_1$ | $0.1\lambda_1$ |
| **All Chapters** | | |
| SLOPE weights | BH | BH |
| gSLOPE weights | Mean | Mean |
| SGS weights | vMean, gSLOPE mean | vMean, gSLOPE mean |
| Path shape | Log-linear | Log-linear |

Table A.6: Model hyperparameters used in the synthetic and real data experiments.

### A.3.3 Datasets

The datasets used in this thesis maintain consistent general characteristics across chapters, though their dimensionality may vary. These differences arise because the datasets were downloaded at different times and are regularly updated at the source. Further variation in dimensionality results from pathway matching, also updated at the source, which influences the set of genes retained.

- Adenoma: Transcriptome profile data to identify formation of colorectal Adenomas [166].
  - Response (binary): Binary labels for whether the sample came from Adenoma or normal mucosa.
  - Data matrix: Transcriptome profile measurements.
  - Grouping structure: Genes were assigned to pathways using the regulatory target gene sets, downloaded from `https://gsea-msigdb.org/gsea/msigdb/human/collections.jsp`. In Chapter 4 only the C3 set was used and in Chapter 5 all nine gene sets were used.
  - Downloaded from `https://ncbi.nlm.nih.gov/` under dataset ID `GDS2947`.
  - Used in Chapters 4 and 5.
- BRCA1: Gene expression data for breast cancer tissue samples [135].
  - Response (continuous): Gene expression measurements for the BRCA1 gene.
  - Data matrix: Gene expression measurements for the other genes.
  - Grouping structure: Variables were grouped via singular value decomposition.
  - Downloaded from `https://iowabiostat.github.io/data-sets/` under name `brca1`.
  - Used in Chapters 5 and 7.
- Cancer: Breast cancer patients treated with tamoxifen for 5 years [128].
  - Response (binary): Binary labels classifying whether the cancer had recurred. In Chapter 7 a synthetic (continuous) response was generated.
  - Data matrix: Gene expression measurements.

  - Grouping structure: Genes were assigned to pathways using the regulatory target gene sets, downloaded from `https://gsea-msigdb.org/gsea/msigdb/human/collections.jsp`. In Chapter 3 all nine gene sets were used and in Chapters 4 and 7 only the C3 set was used.
  - Downloaded from `https://ncbi.nlm.nih.gov/` under dataset ID `GDS807`.
  - Used in Chapters 3, 4, and 7.

- Carbotax: Carbotax study of ovarian tumour growth [107].
  - Response (continuous): Relative tumour volume ($\log_2$ scale).
  - Data matrix: Gene expression measurements. 10000 factors were randomly sampled from a collection of 34964.
  - Grouping structure: Variables were grouped using K-means clustering [125].
  - Downloaded from `https://iowabiostat.github.io/data-sets/` under name `Koussounadis2014`.
  - Used in Chapters 4 and 7.

- Celiac: Gene expression data of primary leucocytes to classify celiac disease [86].
  - Response (binary): Binary labels as to whether a patient has celiac disease.
  - Data matrix: Gene expression measurements from the primary leucocytes.
  - Grouping structure: Genes were assigned to pathways using the regulatory target gene sets, downloaded from `https://gsea-msigdb.org/gsea/msigdb/human/collections.jsp`. In Chapter 4 only the C3 set was used and in Chapter 5 all nine gene sets were used.
  - Downloaded from `https://ncbi.nlm.nih.gov/` under dataset ID `GDS3646`.
  - Used in Chapters 4 and 5.

- Colitis: Blood cells data for classifying whether a patient has colitis [34].
  - Response (binary): Binary labels classifying whether a patient has colitis. In Chapter 7 a synthetic (continuous) response was generated.

  - Data matrix: Gene expression measurements.
  - Grouping structure: Genes were assigned to pathways using the regulatory target gene sets, downloaded from `https://gsea-msigdb.org/gsea/msigdb/human/collections.jsp`. In Chapter 3 all nine gene sets were used and in Chapters 4, and 7 only the C3 set was used.
  - Downloaded from `https://ncbi.nlm.nih.gov/` under dataset ID `GDS1615`.
  - Used in Chapters 3, 4 and 7.

- Rhee: HIV protease gene mutations and their impact on resistance to the drug Nelfinavir [157].
  - Response (continuous): Results of the drug susceptibility assay. Higher values correspond to increased drug resistance.
  - Data matrix: Binary indicators of mutations at specific positions in the HIV protease gene.
  - Grouping structure: Variables were grouped via singular value decomposition.
  - Used in Chapter 7.
  - Downloaded from `https://iowabiostat.github.io/data-sets/` under name `Rhee2006`.

- Scheetz: Gene expression data in the mammalian eye [170].
  - Response (continuous): Gene expression measurements for the TRIM32 gene.
  - Data matrix: Gene expression measurements for other genes.
  - Grouping structure: In Chapters 4 and 7, variables were grouped using K-means clustering [125] and in Chapter 5 they were grouped via singular value decomposition.
  - Used in Chapters 4, 5, and 7.
  - Downloaded from `https://iowabiostat.github.io/data-sets/` under name `Scheetz2006`.

- Trust-experts: Survey response data as to how much participants trust experts (*e.g.*, doctors, nurses, scientists) to provide COVID-19 news and information [168].

  - 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.
  - Used in Chapters 5 and 7.
  - Downloaded from `https://github.com/dajmcdon/sparsegl`.

- Tumour: Gene expression data of pancreatic Cancer samples to identify tumorous tissue [56, 119, 148].
  - Response (binary): Binary labels indicating if a sample is from tumour tissue.
  - Data matrix: Gene expression measurements.
  - Grouping structure: Genes were assigned to pathways using the regulatory target gene sets, downloaded from `https://gsea-msigdb.org/gsea/msigdb/human/collections.jsp`. In Chapter 4 only the C3 set was used and in Chapter 5 all nine gene sets were used.
  - Downloaded from `https://ncbi.nlm.nih.gov/` under dataset ID `GDS4102`.
  - Used in Chapters 4 and 5.

### A.3.4 Model Software

The following packages were used to implement the methods in this thesis:

- General
  - `sgs` R package [64] (our package): Implementation of gSLOPE, gOSCAR, SGS, and SGO models using ATOS; uses screening rules.
  - `glmnet` R package for the lasso [72]. The package uses cyclical coordinate descent.
  - `SLOPE` R package for SLOPE and OSCAR [113]. The package uses the FISTA (proximal) algorithm.
  - `grplasso` R package for gLasso [129]. The package uses coordinate descent.
- Chapter 3

  - `SGL` R package for SGL [179]. The package uses blockwise coordinate descent. Used for SGL only in Chapter 3.

- Chapter 5
  - `dfr` R package [65] (our package): Implementation of DFR for SGL and aSGL using ATOS.
  - The GAP Safe rules were implemented in Python using `https://github.com/EugeneNdiaye/Gap_Safe_Rules`.

- Chapter 7
  - `knockoff` R package for Knockoffs [17]. The package uses second-order knockoff generation.
  - AS-SGS, scaled regression, TSO, and TS-SLOPE, and the group-based Bayesian models were implemented from scratch in R. Code provided in `https://github.com/ff1201/thesis-chapter-7`.
  - ABSLOPE was implemented using modified R code from `https://github.com/wjiang94/ABSLOPE/tree/master`.

# APPENDIX B

# Additional Material for Chapter 3

## B.1 Sparse-group SLOPE

### B.1.1 Path Start

*Proof of Proposition 3.1.* The aim is to find the value of $\lambda$ at which the first variable enters the model. The SGS KKT conditions are given by

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

When all features are zero, the KKT conditions are

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

By the reverse triangle inequality and the ordering of the group weights

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

Now, noting that $\text{cumsum}(\mathbf{y}) \succeq \text{cumsum}(\mathbf{x}) \iff \mathbf{y} \succeq \mathbf{x}$ [111], for $\mathbf{x} \in J_{\text{gslope}}(\mathbf{0}; \mathbf{w})$ it holds that

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

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

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

□

### B.1.2 Binomial Loss Function

To apply ATOS to a binary response, the loss function needs to be convex and $L_f$-smooth. The loss function for logistic regression satisfies these constraints, given by $f(\mathbf{b}) = -\frac{1}{n}\log(\mathcal{L}(\mathbf{b}; \mathbf{y}, \mathbf{X}))$, where $\mathcal{L}$ is the log-likelihood of a binomial distribution, given by

$$\mathcal{L}(\mathbf{b}; \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{n} \left\{ y_i \mathbf{b}^\top \mathbf{x}_i - \log(1 + \exp(\mathbf{b}^\top \mathbf{x}_i)) \right\}. \tag{B.1}$$

The negative of the log-likelihood is used, as this is equivalent to maximising the likelihood.

### B.1.3 Fitting Algorithm

**Theorem B.1.** *The SGS penalty is convex.*

*Proof.* The SLOPE penalty is convex [24]. Similarly, the group SLOPE penalty is also convex [30]. Finally, the sum of convex functions is convex. Hence, the penalty function for SGS is convex. □

### Proximal Operators

**Algorithm 7** Proximal mapping for SLOPE (Algorithm 3 in Bogdan et al. [24])

**Input:** Non-negative and non-increasing sequences $\mathbf{y}, \mathbf{v} \in \mathbb{R}^p$

**while** $\mathbf{y} - \mathbf{v}$ is not non-increasing **do**

Identify non-decreasing and non-constant segments $i : j$ such that

$$y_i - v_i \leq y_{i+1} - v_{i+1} \leq \cdots \leq y_j - v_j \text{ and } y_i - v_i < y_j - v_j$$

Replace values of $\mathbf{y}$ and $\mathbf{v}$ over such segments by their averages:

$$y_k \leftarrow \frac{1}{j-i+1}\sum_{k=i}^{j} y_k, \quad v_k \leftarrow \frac{1}{j-i+1}\sum_{k=i}^{j} v_k, \quad \forall k \in \{i, \ldots, j\}$$

**end while**

**Output:** Proximal SLOPE solution $\mathbf{x} = (\mathbf{y} - \mathbf{v})_+ \in \mathbb{R}^p$

**Algorithm 8** Proximal mapping for gSLOPE (Algorithm 2 in Gossmann et al. [80])

**Input:** Grouped vectors $\mathbf{x} = (\mathbf{x}^{(1)\top}, \ldots, \mathbf{x}^{(m)\top})$ and $\mathbf{y} = (\mathbf{y}^{(1)\top}, \ldots, \mathbf{y}^{(m)\top})$, non-negative and non-increasing penalty sequence $\mathbf{w} \in \mathbb{R}^m$

Compute group norms: $\tilde{\mathbf{y}} = (\|\mathbf{y}^{(1)}\|_2, \ldots, \|\mathbf{y}^{(m)}\|_2)^\top$

Apply SLOPE proximal operator (Algorithm 7) to group norms: $\tilde{\mathbf{x}} = \text{prox}_{\text{slope}}(\tilde{\mathbf{y}}; \mathbf{w})$

**for** $j = 1, \ldots, m$ **do**

Rescale each group: $\mathbf{x}^{(j)} \leftarrow \frac{\tilde{x}_j}{\tilde{y}_j}\mathbf{y}^{(j)}$

**end for**

**Output:** Proximal gSLOPE solution $\mathbf{x} \in \mathbb{R}^m$

### Initial Step Size

Step sizes are often set to 1 by default, although Pedregosa and Gidel [147] recommend the following scheme for ATOS:

1. Set $\epsilon = 10^{-3}, \tilde{\mathbf{z}} = \mathbf{z}_{[0]} - \epsilon\nabla f(\mathbf{z}_{[0]})$.
2. Calculate $\epsilon = 0.1\epsilon$ until $f(\tilde{\mathbf{z}}) \leq f(\mathbf{z}_{[0]})$.
3. Calculate $\gamma_{[0]} = 4\left(f(\mathbf{z}_{[0]}) - f(\tilde{\mathbf{z}}_{[0]})\right)\left\|\nabla f(\mathbf{z}_{[0]})\right\|_2^{-2}$.

## B.2 FDR Control

### B.2.1 FDR Control Proof

The proof of Theorem 3.1 uses several lemmas, which are presented in Appendix B.2.2, with proofs given in Appendix B.2.3. Lemmas B.1 and B.3 were used to derive the penalty sequences (described in Appendix B.2.4).

*Proof of Theorem 3.1.* For the Normal Means model, the SGS problem is

$$\hat{\boldsymbol{\beta}} = \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2}\|\mathbf{y}-\mathbf{b}\|_2^2 + \alpha \sum_{i=1}^{p} v_i |b|_{(i)} + (1-\alpha)\sum_{j=1}^{m} w_j \sqrt{p_j} \|\mathbf{b}^{(j)}\|_2 \right\}. \tag{B.2}$$

It suffices to consider the scenario where $p = n$. For the hypotheses, we have that

$$\begin{aligned}
&\mathbb{P}\left(H_i^v \text{ rejected}\right) = \mathbb{P}\left(\hat{\beta}_i \neq 0 : \|\hat{\boldsymbol{\beta}}^{(j)}\|_2 \neq 0, i \in G_j\right), \\
&\mathbb{P}\left(H_j^g \text{ rejected}\right) = \mathbb{P}\left(\|\hat{\boldsymbol{\beta}}^{(j)}\|_2 \neq 0 : \exists i \in G_j \text{ s.t. } \hat{\beta}_i \neq 0\right).
\end{aligned}$$

Without loss of generality, we set the problem up so that the first $p_0$ variable hypotheses are null, *i.e.*, $\beta_i = 0$ for $i \leq p_0$, and that $m_0$ group hypotheses are null, *i.e.*, $\|\boldsymbol{\beta}^{(j)}\|_2 = 0$ for $j \in \mathcal{Z} \subset [m]$, where $\mathcal{Z}$ denotes the indices of the zero groups. We do not assume these are the first $m_0$ hypothesis, as is done for the variables, so these situations can co-occur. Further, again without loss of generality, we assume the groups are indexed so that $\sum_{i\in G_1}|\beta_i| \geq \ldots \geq \sum_{i\in G_m}|\beta_i|$.

Hence, the variable and group FDR are given by

$$\begin{aligned}
\text{vFDR} &= \mathbb{E}\left[\frac{V^v}{\max(R^v,1)}\right] = \sum_{i=1}^{p}\mathbb{E}\left[\frac{V^v}{r}\mathbf{1}_{\{R^v=r\}}\right] \qquad \text{(B.3)} \\
&= \sum_{r=1}^{p}\frac{1}{r}\sum_{i=1}^{p_0}\mathbb{P}(H_i^v \text{ rejected and } R^v = r), \\
\text{gFDR} &= \sum_{r=1}^{m}\frac{1}{r}\sum_{j=1}^{m_0}\mathbb{P}(H_j^g \text{ rejected and } R^g = r). \qquad \text{(B.4)}
\end{aligned}$$

We first bound the variable FDR, for which we need to bound the key quantity, $\mathbb{P}(H_i^v \text{ rejected and } R^v = r)$. This is achieved using Lemmas B.1 and B.2, which are

used to calculate, for $i \in G_j$,

$$\begin{aligned}\mathbb{P}(H_i^v \text{ rejected and } R^v = r) &\leq \mathbb{P}\left(|y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j \text{ and } \tilde{R}^v = r-1\right) \\ &= \mathbb{P}\left(|y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j\right)\mathbb{P}\left(\tilde{R}^v = r-1\right),\end{aligned}$$

where the second step is by independence of $\mathbf{y}$ and $\tilde{\mathbf{y}}$ (from Lemma B.2). Now, from the definition of $v_r$, we have

$$\begin{aligned}v_r &\geq \frac{\Phi^{-1}\left(1-\frac{q_v r}{2p}\right) - \frac{1}{3}(1-\alpha)a_j w_j}{\alpha} \\ \implies 1 - \Phi\left(\alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j\right) &\leq \frac{q_v r}{2p}. \end{aligned} \tag{B.5}$$

Hence,

$$\begin{aligned}&\mathbb{P}\left(|y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j\right) \\ &= \mathbb{P}\left(y_i > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j\right) + \mathbb{P}\left(y_i < -\alpha v_r - \frac{1}{3}(1-\alpha)a_j w_j\right) \\ &= 1 - \Phi\left(\alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j\right) + \Phi\left(-\alpha v_r - \frac{1}{3}(1-\alpha)a_j w_j\right) \\ &\leq \frac{q_v r}{2p} + \frac{q_v r}{2p} = \frac{q_v r}{p},\end{aligned}$$

where $\Phi\left(-\alpha v_r - \frac{1}{3}(1-\alpha)a_j w_j\right) = 1 - \Phi(\alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j) \leq \frac{q_v r}{2p}$, by the symmetry of $y_i$, as $y_i \sim \mathcal{N}(0,1)$ for $i \leq p_0$. Hence,

$$\mathbb{P}(H_i^v \text{ rejected and } R^v = r) \leq \frac{q_v r}{p}\mathbb{P}(\tilde{R}^v = r-1).$$

Plugging this into Equation B.3 gives

$$\begin{aligned}\text{vFDR} &= \sum_{r=1}^{p}\frac{1}{r}\sum_{i=1}^{p_0}\mathbb{P}(H_i^v \text{ rejected and } R^v = r) \leq \sum_{r=1}^{p}\frac{1}{r}\sum_{i=1}^{p_0}\frac{q_v r}{p}\mathbb{P}(\tilde{R}^v = r-1) \\ &= \sum_{r=1}^{p}\sum_{i=1}^{p_0}\frac{q_v}{p}\mathbb{P}(\tilde{R}^v = r-1) = \sum_{r=1}^{p}\frac{q_v p_0}{p}\mathbb{P}(\tilde{R}^v = r-1) \\ &= \sum_{r\geq 1}\frac{q_v p_0}{p}\mathbb{P}(\tilde{R}^v = r-1) = \frac{q_v p_0}{p}, \text{ by Lemma B.1 assumption,}\end{aligned}$$

which completes the variable section of the proof.

To bound the group FDR, a similar strategy is used, in which we require a bound for $\mathbb{P}(H_j^g \text{ rejected and } R^g = r)$. This bound uses Lemmas B.3 and B.4, so that

$$\begin{aligned}
&\mathbb{P}(H_j^g \text{ rejected and } R^g = r) \\
&\leq \mathbb{P}\left(\sum_{i \in G_j} |y_i| > \alpha \sum_{i \in G_r} v_i + (1-\alpha) w_r p_r \text{ and } \tilde{R}^g = r - 1\right) \\
&= \mathbb{P}\left(\sum_{i \in G_j} |y_i| > \alpha \sum_{i \in G_r} v_i + (1-\alpha) w_r p_r\right) \mathbb{P}\left(\tilde{R}^g = r - 1\right),
\end{aligned}$$

where the second step again follows from independence of $\mathbf{y}$ and $\tilde{\mathbf{y}}$ (Lemma B.4). Now, from the definition of the group weights

$$w_j \geq \frac{F_{\text{FN}}^{-1}(1 - \frac{q_g r}{m}) - \alpha \sum_{i \in G_r} v_i}{(1-\alpha) p_r} \implies 1 - F_{\text{FN}}\left(\alpha \sum_{i \in G_r} v_i + (1-\alpha) p_r w_r\right) \leq \frac{q_g r}{m}.$$

Hence,

$$\mathbb{P}\left(\sum_{i \in G_j} |y_i| > \alpha \sum_{i \in G_r} v_i + (1-\alpha) p_r w_r\right) \leq \frac{q_g r}{m}.$$

Therefore,

$$\mathbb{P}(H_j^g \text{ rejected and } R^g = r) \leq \frac{q_g r}{m} \mathbb{P}(\tilde{R}^g = r - 1).$$

Plugging this into Equation B.4 gives

$$\text{gFDR} = \sum_{r=1}^{m} \frac{1}{r} \sum_{j=1}^{m_0} \mathbb{P}(H_j^g \text{ rejected and } R^g = r) \leq \frac{q_g m_0}{m}.$$

Hence, SGS controls both the variable and group FDR, proving bi-level FDR control under orthogonality. □

*Proof of Corollary 3.1.* Under orthogonal $\mathbf{X}$, such that $\mathbf{X}^\top \mathbf{X} = \mathbf{I}$, the general linear model reduces to that of the Normal Means model $\mathring{\mathbf{y}} := \mathbf{X}^\top \mathbf{y} = \boldsymbol{\beta} + \mathring{\boldsymbol{\epsilon}}$, where $\mathring{\mathbf{y}} \sim \boldsymbol{\mathcal{N}}_p(\boldsymbol{\beta}, \mathbf{I}_p)$ and $\mathring{\boldsymbol{\epsilon}} \sim \boldsymbol{\mathcal{N}}_p(\mathbf{0}, \mathbf{I}_p)$, which satisfy the conditions of Theorem 3.1, providing bi-level FDR control for SGS. □

*Proof of Lemma 3.1.* First, we rewrite the penalty sequence generation process as

a composite function. Without loss of generality, we consider the clamped version of the sequences (see Remark B.1). Let $f : [0, M_v]^p \to [0, M_v]^p$ and $g : [0, M_g]^m \to [0, M_g]^m$ be the functions that generate the variable and group sequences, given by

$$f_i(\mathbf{w}) = \max\left\{0, \max_{j\in[m]}\left(\frac{1}{\alpha}\Phi^{-1}\left(1-\frac{q_v i}{2p}\right) - \frac{1}{3\alpha}(1-\alpha)a_j w_j\right)\right\},\ i \in [p],$$

$$g_i(\mathbf{v}) = \max\left\{0, \max_{j\in[m]}\left(\frac{F_{\text{FN}}^{-1}\left(1-\frac{q_g i}{m}\right) - \alpha\sum_{k\in G_j} v_k}{(1-\alpha)p_j}\right)\right\},\ i \in [m],$$

where $M_v$ and $M_g$ represent the maximum value a variable and group weight can take. These are given by $M_v = \frac{1}{\alpha}\Phi^{-1}(1-q_v/2p)$, under $i = 1$ and $\mathbf{w} \equiv \mathbf{0}$, and $M_g = F_{\text{FN}}^{-1}\left(1-\frac{q_g}{m}\right)/(1-\alpha)\min_{j\in[m]} p_j$, under $i = 1$ and $\mathbf{v} \equiv \mathbf{0}$. We focus solely on the dependence of the functions on the penalty sequences; dependence on other parameters is suppressed for notational simplicity.

Now, consider the composite function which defines the penalty generation process, given by $h : \mathcal{M}_v \to \mathcal{M}_v$ such that $h(\mathbf{v}) = f(g(\mathbf{v})) = (f \circ g)(\mathbf{v})$ and $\mathcal{M}_v = [0, M_v]^p$. Note that one could alternatively define the composite function as $h(\mathbf{w}) = (g \circ f)(\mathbf{w})$. The proof would then proceed analogously and prove the same result.

*(i) Existence:* Proving the sequences always exist is equivalent to showing that the function $h$ has a fixed point. By Brouwer's fixed-point theorem [165], if $\mathcal{M}_v$ is a nonempty, compact, and convex subset of $\mathbb{R}^p$, and $h$ is continuous and maps the domain into itself, then $h$ has a fixed point.

We first prove the various conditions on $\mathcal{M}_v$. The set is a subset of $\mathbb{R}^p$ by definition and is nonempty since $\mathbf{0} \in \mathcal{M}_v$. A set is compact if it is closed and bounded. To prove the set $\mathcal{M}_v$ is closed, consider the sequence $(\mathbf{v}^{(k)})_{k=1}^{\infty}$ of vectors in $\mathcal{M}_v$ that converges to $\mathbf{v} \in \mathbb{R}^p$. For each $i \in [p]$ and all $k$, $0 \le v_i^{(k)} \le M_v$. Since $v_i^{(k)} \to v_i$ and inequalities are preserved under limits, it follows that $0 \le v_i \le M_v$, so that $\mathcal{M}_v$ contains all its limit points.

Further, to prove boundedness, there must exist a point $v' \in \mathbb{R}^p$ and $L \in \mathbb{R}$ such that $\|\mathbf{v} - \mathbf{v}'\| < L$, for all $\mathbf{v} \in \mathcal{M}_v$ and for some norm $\|\cdot\|$. Consider the point $\mathbf{v}' \equiv \mathbf{0}$, $L = M_v + 1$, and the norm $\|\cdot\|_\infty$. Then, $\|\mathbf{v} - \mathbf{0}\|_\infty = \|\mathbf{v}\|_\infty = \max_{i\in[p]} v_i \le M_v < M_v + 1 = L$. Hence, $\mathcal{M}_v$ is compact. Finally, it is convex if for any points $\mathbf{v}, \mathbf{v'} \in \mathcal{M}_v$, and $L \in [0, 1]$, $L\mathbf{v} + (1-L)\mathbf{v'} \in \mathcal{M}_v$. Consider that for all $i \in [p]$, $0 \le v_i \le M_v, 0 \le v_i' \le M_v$, so that $0 \le Lv_i + (1-L)v_i' \le LM_v + (1-L)M_v = M_v$. Therefore $L\mathbf{v} + (1-L)\mathbf{v'} \in \mathcal{M}_v$, so that $\mathcal{M}_v$ is convex.

The conditions required on $h$ are proved next. To show the function maps the domain to itself, take a general point $\mathbf{v} \in \mathcal{M}_v$ and consider that $\mathbf{w} = g(\mathbf{v}) \geq 0$, so that $f(\mathbf{w}) \geq 0$, by definition. Similarly, we have $f(\mathbf{w}) \leq M_v$, showing that $h(\mathbf{v}) \in \mathcal{M}_v$. Finally, the function $h$ is continuous as it is the composition of two continuous functions, $f$ and $g$. These are both continuous as affine functions are continuous and the maximum of continuous functions is continuous. Therefore, by Brouwer's fixed-point theorem, the penalty sequences $(\mathbf{v}, \mathbf{w})$ always exist.

*(ii) Uniqueness:* Next, we aim to prove that the fixed point is unique under $\max_{j\in[m]} a_j < 3$. To do this, the Banach contraction mapping theorem [45] is used, which states that $h$ admits a unique fixed point if $h$ is a contraction mapping, *i.e.*, there exists a constant $L < 1$ for all $\mathbf{v}, \mathbf{v}' \in \mathcal{M}_v$ such that

$$\|h(\mathbf{v}) - h(\mathbf{v}')\|_\infty \leq L\|\mathbf{v} - \mathbf{v}'\|_\infty.$$

This is equivalent to finding a Lipschitz constant. If $f$ and $g$ are Lipschitz functions with $L_f$ and $L_g$, then $h = f \circ g$ is a Lipschitz function with constant $L_f L_g$. This can be seen clearly by

$$\|h(\mathbf{v}) - h(\mathbf{v}')\|_\infty = \|f(g(\mathbf{v})) - f(g(\mathbf{v}'))\|_\infty \leq L_f\|g(\mathbf{v}) - g(\mathbf{v})'\|_\infty \leq L_f L_g\|\mathbf{v} - \mathbf{v}'\|_\infty.$$

To find $L_g$, consider

$$\begin{aligned}|g(v_i) - g(v_i')| &\leq \max_{j\in[m]} \left( \frac{\alpha|\sum_{k\in G_j}(v_k - v_k')|}{(1-\alpha)p_j} \right) \leq \max_{j\in[m]} \left( \frac{\alpha \sum_{k\in G_j}|v_k - v_k'|}{(1-\alpha)p_j} \right) \\ &\leq \max_{j\in[m]} \left( \frac{\alpha p_j \|\mathbf{v} - \mathbf{v}'\|_\infty}{(1-\alpha)p_j} \right) = \max_{j\in[m]} \left( \frac{\alpha\|\mathbf{v} - \mathbf{v}'\|_\infty}{1-\alpha} \right),\end{aligned}$$

where the first inequality follows by Lipschitz continuity of the maximum operator, the second by the triangle inequality, and the final by definition of the infinity norm. Therefore, for all $j \in [m]$,

$$\begin{aligned}&|g(v_i) - g(v_i')| \leq \frac{\alpha}{1-\alpha}\|v_k - v_k'\|_\infty \\ \implies &\|g(\mathbf{v}) - g(\mathbf{v}')\|_\infty \leq \max_{j\in[m]} |g(v_i) - g(v_i')| \leq \frac{\alpha}{1-\alpha}\|\mathbf{v} - \mathbf{v}'\|_\infty,\end{aligned}$$

so that $L_g = \alpha/(1-\alpha)$. By similar calculations, $L_f = (1-\alpha)\max_{j\in[m]} a_j/3\alpha$. So, the Lipschitz constant of $h$ is $L_f L_g = \max_{j\in[m]} a_j/3$, proving that $h$ is a contrac-

tion mapping if $\max_{j\in[m]} a_j < 3$. Therefore, by the Banach contraction mapping theorem, a unique fixed point exists if $\max_{j\in[m]} a_j < 3$ with geometric convergence

$$\|\mathbf{v}^{(n)} - \mathbf{v}^*\|_\infty \leq \left(\frac{\max_{j\in[m]} a_j}{3}\right)^n \|\mathbf{v}^{(0)} - \mathbf{v}^*\|_\infty.$$

□

*Remark* B.1. Theorem 3.1 remains valid under the clamped versions of the sequences; however, for notational simplicity, the result is stated without the clamping.

### B.2.2 Lemma Statements

Lemmas B.1–B.4 are used directly in the proof of Theorem 3.1. Lemmas B.5 and B.6 are used to prove Lemmas B.1 and B.3, respectively. Some of the lemmas also required new norm results (presented in Appendix B.2.5): Lemmas B.5 and B.6 use Lemma B.7, and Lemmas B.2 and B.5 use Lemma B.8. In the lemmas, we use the following notation

$$a_j := |\{i \in G_j : \beta_i \neq 0\}|, \tag{B.6}$$
$$\tilde{a}_j := |\{i \in G_j : \beta_i = 0\}|, \tag{B.7}$$
$$\mathbb{I}_k^v := \{j \in [m] : G_j \cap \{k, \ldots, r\} \neq \varnothing\}, \tag{B.8}$$
$$\mathbb{J}_k^v := \{j \in [m] : G_j \cap \{r+1, \ldots, k\} \neq \varnothing\}, \tag{B.9}$$
$$\mathbb{I}_k^g := G_k \cup \cdots \cup G_r, \tag{B.10}$$
$$\mathbb{J}_k^g := G_{r+1} \cup \cdots \cup G_k. \tag{B.11}$$

**Lemma B.1.** *Let $H_i^v$ be a null hypothesis, so that $i \leq p_0, i \in G_j$, and $r \geq 1$. Then,*

$$\{\mathbf{y} : H_i^v \text{ rejected and } R^v = r\} = \left\{\mathbf{y} : |y_i| > \alpha v_r + \frac{1}{3}(1-\alpha) a_j w_j \text{ and } R^v = r\right\}.$$

**Lemma B.2.** *Consider applying SGS to $\tilde{\mathbf{y}} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_p)$ with weights $\tilde{\mathbf{v}} = (v_2, \ldots, v_p)$ and $\tilde{\mathbf{w}}$ (which is $\mathbf{w}$ with the $i$th group penalty removed if variable $i$ is a singleton group and $\mathbf{w}$ if it is not), and let $\tilde{R}^v$ be the number of rejections*

*generated. Then, for* $r \geq 1$ *and* $i \in G_j$,

$$\begin{aligned}&\left\{\mathbf{y} : |y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j \text{ and } R^v = r\right\}\\ \subset&\left\{\mathbf{y} : |y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j \text{ and } \tilde{R}^v = r-1\right\}.\end{aligned}$$

**Lemma B.3.** *Let* $H_j^g$ *be a null hypothesis, so that* $j \in \mathcal{Z}$*, and let* $r \geq 1$*. Then,*

$$\left\{\mathbf{y} : H_j^g \text{ rejected and } R^g = r\right\} = \left\{\mathbf{y} : \sum_{i\in G_j}|y_i| > \alpha \sum_{i\in G_r} v_i + (1-\alpha)w_r p_r \text{ and } R^g = r\right\}.$$

**Lemma B.4.** *Consider applying SGS to* $\tilde{\mathbf{y}}$*, which is* $\mathbf{y}$ *with the observations from group* $j \in [m]$ *removed, with weights* $\tilde{\mathbf{w}} = (w_2, \ldots, w_p)$ *and* $\tilde{\mathbf{v}} = \mathbf{v}\backslash\{v_i : i \in G_j\}$*, and let* $\tilde{R}^g$ *be the number of rejections generated. Then, for* $r \geq 1$

$$\begin{aligned}&\left\{\mathbf{y} : \sum_{i\in G_r}|y_i| > \alpha \sum_{i\in G_r} v_i + (1-\alpha)w_r p_r \text{ and } R^g = r\right\}\\ \subset&\left\{\mathbf{y} : \sum_{i\in G_r}|y_i| > \alpha \sum_{i\in G_r} v_i + (1-\alpha)w_r p_r \text{ and } \tilde{R}^g = r-1\right\}.\end{aligned}$$

**Lemma B.5.** *Consider non-increasing and non-negative sequences* $y_1 \geq \cdots \geq y_p \geq 0, v_1 \geq \cdots \geq v_p \geq 0$*, and* $w_1 \geq \cdots \geq w_m \geq 0$*, and let* $\hat{\mathbf{b}}$ *be the solution to*

$$\min_{\mathbf{b}} \underbrace{\left[\frac{1}{2}\|\mathbf{y}-\mathbf{b}\|_2^2 + \alpha\sum_{i=1}^p v_i b_i + (1-\alpha)\sum_{j=1}^m w_j\sqrt{p_j}\|\mathbf{b}^{(j)}\|_2\right]}_{=:f(\mathbf{b})}$$

*subject to* $b_1 \geq \cdots \geq b_p \geq 0$ *and* $\sqrt{p_1}\|\mathbf{b}^{(1)}\|_2 \geq \cdots \geq \sqrt{p_m}\|\mathbf{b}^{(m)}\|_2 \geq 0$.

*Then, if the first* $r$ $\hat{b}_i$ *are positive for* $i \in [p]$*, then for every* $k \leq r$*:*

$$\sum_{i=k}^r y_i > \alpha\sum_{i=k}^r v_i + \frac{1}{3}(1-\alpha)\sum_{j\in\mathbb{I}_k^v} a_j w_j, \tag{B.12}$$

*where* $a_j$ *and* $\mathbb{I}_k^v$ *are defined in Equations B.6 and B.8, respectively, and* $0 < h <$

$\min_{i\in\{k,\dots,r\}} b_i$. *Also, for every* $k \geq r+1$*:*

$$\sum_{i=r+1}^{k} y_i \leq \alpha \sum_{i=r+1}^{k} v_i + (1-\alpha)\sum_{j\in\mathbb{J}_k^v} w_j\sqrt{p_j}\sqrt{\tilde{a}_j}, \tag{B.13}$$

*where* $\tilde{a}_j$ *and* $\mathbb{J}_k^v$ *are defined in Equations B.7 and B.9, respectively.*

**Lemma B.6.** *Consider non-increasing and non-negative sequences* $\sum_{i\in G_1} y_i \geq \cdots \geq \sum_{i\in G_m} y_i \geq 0, v_1 \geq \cdots \geq v_p \geq 0$*, and* $w_1 \geq \cdots \geq w_m \geq 0$*, and let* $\hat{\mathbf{b}}$ *be the solution to*

$$\min_{\mathbf{b}} \underbrace{\left[\frac{1}{2}\|\mathbf{y}-\mathbf{b}\|_2^2 + \alpha\sum_{i=1}^{p} v_i b_i + (1-\alpha)\sum_{j=1}^{m} w_j\sqrt{p_j}\|\mathbf{b}^{(j)}\|_2\right]}_{=:f(\mathbf{b})}$$

*subject to* $b_1 \geq \cdots \geq b_p \geq 0$ *and* $\sqrt{p_1}\|\mathbf{b}^{(1)}\|_2 \geq \cdots \geq \sqrt{p_m}\|\mathbf{b}^{(m)}\|_2 \geq 0$.

*Then, if there are exactly* $r$ *non-zero* $\|\hat{\mathbf{b}}^{(j)}\|_2$ *values for* $j \in [m]$*, then for every* $k \leq r$*:*

$$\sum_{i\in\mathbb{I}_k^g} y_i > \alpha\sum_{i\in\mathbb{I}_k^g} v_i + (1-\alpha)\sum_{j=k}^{r} w_j p_j, \tag{B.14}$$

*where* $\mathbb{I}_k^g$ *is defined in Equation B.10 and for every* $k \geq r+1$*:*

$$\sum_{i\in\mathbb{J}_k^g} y_i \leq \alpha\sum_{i\in\mathbb{J}_k^g} v_i + (1-\alpha)\sum_{j=r+1}^{k} w_j p_j, \tag{B.15}$$

*where* $\mathbb{J}_k^g$ *is defined in Equation B.11.*

### B.2.3 Lemma Proofs

#### Lemma B.1

*Proof of Lemma B.1.* We use Lemma B.5 to prove Lemma B.1. Taking $k = r$ and $R = r$ in Equation B.12 and $k = r+1$ and $R = r$ in Equation B.13, we obtain the following two expressions

$$|y|_{(r)} > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j \quad \text{and} \quad |y|_{(r+1)} \leq \alpha v_{r+1} + (1-\alpha)w_{j'}\sqrt{p_{j'}}\sqrt{\tilde{a}_{j'}},$$

where $\mathbb{I}_r^v = j$ for $r \in G_j$ and $\mathbb{I}_{r+1}^v = j'$ for $r+1 \in G_{j'}$. To prove Lemma B.1, we first want to show

$$\begin{aligned}\{\mathbf{y} : H_i^v \text{ rejected and } R = r\} &= \{\mathbf{y} : \hat{b}_i \neq 0 \text{ and } R = r\} \\ &\subset \left\{\mathbf{y} : |y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j \text{ and } R = r\right\}.\end{aligned}$$

The first equality is by definition, so we are only proving the subset. If we fix $i \in [p]$ and suppose $\hat{b}_i$ is non-zero (so that we reject $H_i^v$), then it must hold that $|y_i| \geq |y|_{(r)} > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j$, proving that

$$\{\mathbf{y} : H_i^v \text{ rejected and } R = r\} \subset \left\{\mathbf{y} : |y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j \text{ and } R = r\right\}.$$

Conversely, to show the other direction, assume that $|y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j$ and $R = r$. Then, we must reject $H_i^v$, since $|y_i| > \alpha v_{r+1} + \frac{1}{3}(1-\alpha)a_{j'} w_{j'} \geq |y|_{(r+1)}$. This shows that

$$\{\mathbf{y} : H_i^v \text{ rejected and } R = r\} \supset \left\{\mathbf{y} : |y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j \text{ and } R = r\right\}.$$

□

**Lemma B.2**

In the proof of Lemma B.5, we used Bounds (i) from Lemmas B.7 and B.8 to obtain Equations B.12 and B.13. If instead we use Bound (ii) from Lemma B.7, we obtain

$$\sum_{i=k}^{r} y_i > \alpha \sum_{i=k}^{r} v_i + (1-\alpha) \sum_{j \in \mathbb{I}_k^v} \frac{w_j \sqrt{p_j} a_j}{2\|\hat{\mathbf{b}}^{(j)}\|_2}, \quad \sum_{i=r+1}^{k} y_i \leq \alpha \sum_{i=r+1}^{j} v_i + (1-\alpha) \sum_{j \in \mathbb{J}_k^v} \frac{w_j \sqrt{p_j} \tilde{a}_j}{2\|\hat{\mathbf{b}}^{(j)}\|_2}. \tag{B.16}$$

*Proof of Lemma B.2.* We first assume without loss of generality that $\mathbf{y} \succeq \mathbf{0}$. The solution to Equation B.2 has $r$ strictly positive values. We need to prove that if $y_1$ is rejected, then the solution to

$$\min_{\tilde{b}} \underbrace{\left[\frac{1}{2}\sum_{i=1}^{p-1}(\tilde{y}_i - \tilde{b}_i)^2 + \alpha \sum_{i=1}^{p-1} \tilde{v}_i |\tilde{b}|_{(i)} + (1-\alpha)\sum_{j=1}^{\tilde{m}} \tilde{w}_j \sqrt{\tilde{p}_j} \|\tilde{\mathbf{b}}^{(j)}\|_2\right]}_{=:g(\tilde{b})}, \tag{B.17}$$

has exactly $r-1$ non-zero values. Here, the removed $y_1$ can correspond either to a singleton group, in which case $\tilde{m} = m-1$ and $\tilde{\mathbf{w}} \in \mathbb{R}^{m-1}$, or it is part of a larger group, such that $\tilde{m} = m$ and $\tilde{\mathbf{w}} = \mathbf{w}$. We need to prove that the optimal solution $\hat{\mathbf{b}}$ to Equation B.17 has both at least and at most $r-1$ non-zero entries. For both, we argue by contradiction by using new suboptimal solutions, as in Lemma B.5.

*At least $r-1$ non-zero entries:* Suppose by contradiction that $\hat{\mathbf{b}}$ has $k-1$ non-zero values, $k < r$. If we denote $I = \{i : \tilde{y}_i \geq \tilde{y}_k \text{ and } \tilde{y}_i \leq \tilde{y}_{r-1}\}$, then we can denote a new suboptimal solution as

$$c_i = \begin{cases} h & \text{if } i \in I \\ \hat{b}_i & \text{otherwise,} \end{cases}$$

where $0 < h < \min_{i \in I} \hat{b}_i$ and $h \leq 1$. By optimality, we require $g(\hat{\mathbf{b}}) - g(\mathbf{c}) \leq 0$. To prove the contradiction, we will show the contrary. Denoting $\mathbb{I}_k^v$ as by Equation B.8 and $\tilde{\mathbb{I}}_k^v = \{j \in [\tilde{m}] : G_j \cap I \neq \varnothing\}$, we have

$$\begin{aligned}
&g(\hat{\mathbf{b}}) - g(\mathbf{c}) \\
&= h \sum_{i=k}^{r-1} \tilde{y}_{(i)} - \frac{1}{2} \sum_{i=k}^{r-1} h^2 - \alpha h \sum_{i=k}^{r-1} \tilde{v}_i + (1-\alpha) \sum_{j \in \tilde{\mathbb{I}}_k^v} \tilde{w}_j \sqrt{\tilde{p}_j} (\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2) \\
&\geq h \sum_{i=k}^{r-1} \tilde{y}_{(i)} - \frac{1}{2} \sum_{i=k}^{r-1} h^2 - \alpha h \sum_{i=k}^{r-1} \tilde{v}_i - (1-\alpha) h \sum_{j \in \tilde{\mathbb{I}}_k^v} \frac{\tilde{w}_j \sqrt{\tilde{p}_j} \tilde{a}_j}{2 \|\hat{\mathbf{b}}^{(j)}\|_2}, \quad \text{by Lemma (B.8)} \\
&= h \sum_{i=k}^{r-1} \tilde{y}_{(i-1)} - \frac{1}{2} \sum_{i=k+1}^{r} h^2 - \alpha h \sum_{i=k+1}^{r} \tilde{v}_i - (1-\alpha) h \sum_{j \in \tilde{\mathbb{I}}_k^v} \frac{\tilde{w}_j \sqrt{\tilde{p}_j} \tilde{a}_j}{2 \|\hat{\mathbf{b}}^{(j)}\|_2}, \quad \text{as } \tilde{v}_i = v_{i-1}, \\
&\geq h \sum_{i=k}^{r-1} \tilde{y}_{(i)} - \frac{1}{2} \sum_{i=k+1}^{r} h^2 - \alpha h \sum_{i=k+1}^{r} \tilde{v}_i - (1-\alpha) h \sum_{j \in \tilde{\mathbb{I}}_k^v} \frac{\tilde{w}_j \sqrt{\tilde{p}_j} \tilde{a}_j}{2 \|\hat{\mathbf{b}}^{(j)}\|_2}, \quad \text{as } \tilde{y}_{(i-1)} \geq y_{(i)}, \\
&\geq h \sum_{i=k}^{r-1} \tilde{y}_{(i)} - \frac{1}{2} \sum_{i=k+1}^{r} h^2 - \alpha h \sum_{i=k+1}^{r} \tilde{v}_i - (1-\alpha) h \sum_{j \in \mathbb{I}_k^v} \frac{w_j \sqrt{p_j} a_j}{2 \|\hat{\mathbf{b}}^{(j)}\|_2},
\end{aligned}$$

where the final inequality is because under $\tilde{\mathbb{I}}_k^v$ we have either the same number of summations as under $\mathbb{I}_k^v$, or one less. Now, by selecting $h$ small enough, we obtain $g(\hat{\mathbf{b}}) - g(\mathbf{c}) > 0$ by Equation B.16, giving the desired contradiction.

*At most $r-1$ non-zero entries:* The proof here is similar. We again argue

by contradiction that $\hat{\mathbf{b}}$ has $k$ non-zero values, $k \geq r$. Denoting $I = \{i : \tilde{y}_i \geq \tilde{y}_r \text{ and } \tilde{y}_i \leq \tilde{y}_k\}$, we denote a new suboptimal solution as

$$c_i = \begin{cases} \hat{b}_i - h & \text{if } i \in I \\ \hat{b}_i & \text{otherwise,} \end{cases}$$

where $0 < h < \min_{i \in I} \hat{b}_i$. By optimality, we require $g(\hat{\mathbf{b}}) - g(\mathbf{c}) \leq 0$. To prove the contradiction, we will again show the contrary. So,

$$\begin{aligned} g(\hat{\mathbf{b}}) - g(\mathbf{c}) &= -\frac{1}{2}\sum_{i=r}^{k} h^2 - h\sum_{i=r}^{k}(\tilde{y}_{(i)} - \hat{b}_{(i)}) + \alpha h \sum_{i=r}^{k} \tilde{v}_i \\ &\quad + (1-\alpha)\sum_{j \in \mathbb{I}_k^v} \tilde{w}_j \sqrt{\tilde{p}_j}(\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2) \\ &\geq -\frac{1}{2}\sum_{i=r}^{k} h^2 - h\sum_{i=r}^{k}(\tilde{y}_{(i)} - \hat{b}_{(i)}) + \alpha h \sum_{i=r}^{k} \tilde{v}_i \\ &\quad + \frac{h}{3}(1-\alpha)\sum_{j \in \mathbb{I}_k^v} \tilde{w}_j \tilde{a}_j. \end{aligned}$$

Now, by selecting $h$ to be very small and using Equation B.12, we obtain $g(\hat{\mathbf{b}}) - g(\mathbf{c}) > 0$, giving a contradiction and finishing the proof. □

**Lemma B.3**

*Proof of Lemma B.3.* We use Lemma B.6 to prove Lemma B.3. Taking $j = r$ and $R = r$ in Equation B.14 and $j = r+1$ and $R = r$ in Equation B.15, we obtain the following two expressions

$$\sum_{i \in G_r} |y_i| > \alpha \sum_{i \in G_r} v_i + (1-\alpha) w_r p_r \quad \text{and} \quad \sum_{i \in G_{r+1}} |y_i| \leq \alpha \sum_{i \in G_{r+1}} v_i + (1-\alpha) w_{r+1} p_{r+1}.$$

We first want to show

$$\{\mathbf{y} : H_j^g \text{ rejected and } R = r\} \subset \left\{ \mathbf{y} : \sum_{i \in G_j} |y_i| > \alpha \sum_{i \in G_r} v_i + (1-\alpha) w_r p_r \text{ and } R = r \right\}.$$

If we fix a group $j \in [m]$ such that $\|\hat{\mathbf{b}}^{(j)}\|_2 \neq 0$, then

$$\sum_{i \in G_j} |y_i| \geq \sum_{i \in G_r} |y_i| > \alpha \sum_{i \in G_r} v_i + (1-\alpha) w_r p_r,$$

proving that

$$\{\mathbf{y} : H_j^g \text{ rejected and } R = r\} \subset \left\{ \mathbf{y} : \sum_{i \in G_j} |y_i| > \alpha \sum_{i \in G_r} v_i + (1-\alpha) w_r p_r \text{ and } R = r \right\}.$$

To show the other direction, assume that $\sum_{i \in G_j} |y_i| > \alpha \sum_{i \in G_r} v_i + (1-\alpha) w_r p_r$ and $R = r$ for group $j$ as before. Then, we must reject $H_j^g$ since $\sum_{i \in G_j} |y_i| > \sum_{i \in G_{r+1}} |y_i|$. This shows that

$$\{\mathbf{y} : H_j^g \text{ rejected and } R = r\} \supset \left\{ \mathbf{y} : \sum_{i \in G_j} |y_i| > \alpha \sum_{i \in G_r} v_i + (1-\alpha) w_r p_r \text{ and } R = r \right\}.$$

□

**Lemma B.4**

*Proof of Lemma B.4.* Assume without loss of generality that $\mathbf{y} \succeq \mathbf{0}$. The solution to Equation B.2 has $r$ non-zero groups. We aim to prove that if $\{y_i : i \in G_1\}$ is rejected, then the solution to

$$\min_{\tilde{\mathbf{b}}} \underbrace{\left[ \frac{1}{2} \sum_{i \in \mathcal{G} \setminus G_1} (\tilde{y}_i - \tilde{b}_i)^2 + \alpha \sum_{i \in \mathcal{G} \setminus G_j} \tilde{v}_i |\tilde{b}|_i + (1-\alpha) \sum_{j=1}^{m-1} \tilde{w}_j \sqrt{\tilde{p}_j} \|\tilde{\mathbf{b}}^{(j)}\|_2 \right]}_{=: g(\tilde{\mathbf{b}})},$$

has exactly $r-1$ non-zero groups. To prove this, we will prove it has at least and at most $r-1$ non-zero groups. We again use proof by contradiction, as in Lemma B.2.

*At least $r-1$ non-zero groups:* Suppose by contradiction that $\hat{\mathbf{b}}$ has $k-1$ non-zero groups, for $k < r$. Let $I = \{G : \sum_{i \in G} \tilde{y}_i \geq \sum_{i \in G_k} \tilde{y}_i \text{ and } \sum_{i \in G} \tilde{y}_i \leq \sum_{i \in G_{r-1}} \tilde{y}_i\}$.

Denoting a new suboptimal solution as

$$c_i = \begin{cases} h & \text{if } i \in I \\ \hat{b}_i & \text{otherwise,} \end{cases}$$

where $0 < h < \min_{i \in I} \hat{b}_i$ and $h \leq 1$. By optimality, we should have $g(\hat{\mathbf{b}}) - g(\mathbf{c}) \leq 0$. However,

$$g(\hat{\mathbf{b}}) - g(\mathbf{c}) \geq h \sum_{i \in I} \tilde{y}_i - \frac{1}{2} \sum_{i \in I} h^2 - \alpha h \sum_{i \in I} \tilde{v}_i - (1-\alpha) h \sum_{j=k}^{r-1} \tilde{p}_j \tilde{w}_j.$$

The proof here is very similar to that of Lemma B.2, so we will only observe the following three facts:

- $\sum_{i \in I} \tilde{y}_i \geq \sum_{i \in I'} y_i$, where $I' = \{G : \sum_{i \in G} y_i \geq \sum_{i \in G_{k+1}} y_i \text{ and } \sum_{i \in G} y_i \leq \sum_{i \in G_r} y_i\}$.
- $\tilde{w}_i = w_{i+1}$, by design.
- $\alpha \sum_{i \in I} \tilde{v}_i \leq \alpha \sum_{i \in I'} v_i$, as we are summing over more penalty terms in the latter.

Using these, and by setting $h$ very small, we can apply Equation B.14 to show that $g(\hat{\mathbf{b}}) - g(\mathbf{c}) > 0$, which is a contradiction, so that we must have at least $r-1$ non-zero groups.

*At most $r-1$ non-zero groups:* Suppose by contradiction that $\hat{\mathbf{b}}$ has $k$ non-zero groups, with $k \geq r$. We again define an indicator set $I = \{G : \sum_{i \in G} \tilde{y}_i \geq \sum_{i \in G_r} \tilde{y}_i \text{ and } \sum_{i \in G} \tilde{y}_i \leq \sum_{i \in G_k} \tilde{y}_i\}$, and a new suboptimal solution

$$c_i = \begin{cases} \hat{b}_i - h & \text{if } i \in I \\ \hat{b}_i & \text{otherwise,} \end{cases}$$

with $0 < h < \min_{i \in I} \hat{b}_i$. Now,

$$\begin{aligned} g(\hat{\mathbf{b}}) - g(\mathbf{c}) &> -\frac{1}{2}\sum_{i \in I} h^2 - h\sum_{i \in I} \tilde{y}_i + \alpha h \sum_{i \in I} \tilde{v}_i + (1-\alpha)h \sum_{j=r}^{k} \tilde{w}_j \tilde{p}_j \\ &\geq -\frac{1}{2}\sum_{i \in I} h^2 - h\sum_{i \in I'} y_i + (1-\alpha)h \sum_{j=r+1}^{k+1} w_j p_j, \quad \text{by definition,} \end{aligned}$$

where $I' = \{G : \sum_{i \in G} y_i \geq \sum_{i \in G_{r+1}} y_i \text{ and } \sum_{i \in G} y_i \leq \sum_{i \in G_{k+1}} y_i\}$. Now, by defining $\mathbb{I}_k^g$ by Equation B.10 and looking at the proof of Lemma B.6, we see that we can bound $\alpha \sum_{i \in \mathbb{I}_k^g} v_i \geq 0$, so that instead of obtaining Equation B.14, we have

$$\sum_{i \in \mathbb{I}_k^g} y_i - (1-\alpha) \sum_{j=r+1}^{k} w_j p_j \leq 0.$$

By picking $h$ to be very small, from this we see that we must have $g(\hat{\mathbf{b}}) - g(\mathbf{c}) > 0$, which is a contradiction, so that the solution has at most $r-1$ non-zero groups. □

**Lemma B.5**

*Proof of Lemma B.5.* For the first claim, consider a new feasible (but suboptimal) solution

$$c_i = \begin{cases} \hat{b}_i - h & \text{if } i \in \{k, \ldots, r\} \\ \hat{b}_i & \text{otherwise,} \end{cases}$$

where $0 < h < \min_{i \in \{k,\ldots,r\}} b_i$. By optimality of $\hat{\mathbf{b}}$, we have $f(\hat{\mathbf{b}}) - f(\mathbf{c}) \leq 0$. Hence,

$$\begin{aligned} f(\hat{\mathbf{b}}) - f(\mathbf{c}) &= \frac{1}{2}\|\mathbf{y} - \hat{\mathbf{b}}\|_2^2 - \frac{1}{2}\|\mathbf{y} - \mathbf{c}\|_2^2 + \alpha \sum_{i=1}^{p} v_i(\hat{b}_i - c_i) \\ &\quad + (1-\alpha)\sum_{j=1}^{m} w_j \sqrt{p_j}(\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2) \\ &= \frac{1}{2}\sum_{i=1}^{p}\left[(y_i - \hat{b}_i)^2 - (y_i - c_i)^2\right] + \alpha \sum_{i=1}^{p} v_i(\hat{b}_i - c_i) \\ &\quad + (1-\alpha)\sum_{j=1}^{m} w_j \sqrt{p_j}(\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2). \end{aligned}$$

By definition of $\mathbf{c}$, it follows that

$$f(\hat{\mathbf{b}}) - f(\mathbf{c}) = \underbrace{\frac{1}{2}\sum_{i=k}^{r}\Big((y_i - \hat{b}_i)^2 - (y_i - c_i)^2\Big)}_{:=T_1} + \underbrace{\alpha\sum_{i=k}^{r} v_i(\hat{b}_i - c_i)}_{:=T_2} + \underbrace{(1-\alpha)\sum_{j\in\mathbb{I}_k^v} w_j\sqrt{p_j}(\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2)}_{:=T_3}.$$

Working through each term separately:

$\underline{T_1}$:

$$\begin{aligned} T_1 &= \frac{1}{2}\sum_{i=k}^{r}\Big((y_i - \hat{b}_i)^2 - (y_i - c_i)^2\Big) = \frac{1}{2}\sum_{i=k}^{r}(y_i^2 + \hat{b}_i^2 - 2y_i\hat{b}_i - y_i^2 + 2y_ic_i - c_i^2) \\ &= \frac{1}{2}\sum_{i=k}^{r}(\hat{b}_i^2 - c_i^2) + \sum_{i=k}^{r} y_i(c_i - \hat{b}_i) = \frac{1}{2}\sum_{i=k}^{r}(\hat{b}_i^2 - (\hat{b}_i - h)^2) + \sum_{i=k}^{r} y_i(\hat{b}_i - h - \hat{b}_i) \\ &= \frac{1}{2}\sum_{i=k}^{r}(\hat{b}_i^2 - \hat{b}_i^2 - h^2 + 2\hat{b}_ih) - \sum_{i=k}^{r} y_ih = -\frac{1}{2}\sum_{i=k}^{r} h^2 - h\sum_{i=k}^{r}(y_i - \hat{b}_i). \end{aligned}$$

$\underline{T_2}$:

$$T_2 = \alpha\sum_{i=k}^{r} v_i(\hat{b}_i - c_i) = \alpha h\sum_{i=k}^{r} v_i.$$

$\underline{T_3}$: Here, we apply Bound (i) from Lemma B.7. So, for $j \in [m]$,

$$\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2 \geq \frac{ha_j}{3\sqrt{p_j}}. \tag{B.18}$$

Hence,

$$T_3 = (1-\alpha)\sum_{j\in\mathbb{I}_k^v} w_j\sqrt{p_j}(\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2) \geq \frac{1}{3}(1-\alpha)h\sum_{j\in\mathbb{I}_k^v} a_jw_j.$$

Combining the three terms back together, we have that

$$0 \geq f(\hat{\mathbf{b}}) - f(\mathbf{c}) \geq -\frac{1}{2}\sum_{i=k}^{r} h^2 - h\sum_{i=k}^{r}(y_i - \hat{b}_i) + \alpha h \sum_{i=k}^{r} v_i + \frac{1}{3}(1-\alpha)h\sum_{j\in\mathbb{I}_k^v} a_j w_j$$

$$\implies -h\sum_{i=k}^{r}(y_i - \hat{b}_i) + \alpha h \sum_{i=k}^{r} v_i + \frac{1}{3}(1-\alpha)h\sum_{j\in\mathbb{I}_k^v} a_j w_j \leq -\frac{1}{2}\sum_{i=k}^{r} h^2.$$

Removing the common factor $h$ and then taking the limit as $h \to 0$ gives

$$\sum_{i=k}^{r} y_i - \sum_{i=k}^{r} \hat{b}_i - \alpha \sum_{i=k}^{r} v_i - \frac{1}{3}(1-\alpha)\sum_{j\in\mathbb{I}_k^v} a_j w_j \geq 0.$$

Now, by assumption we have $\sum_{i=k}^{r} \hat{b}_i > 0$, so

$$\sum_{i=k}^{r} y_i > \alpha \sum_{i=k}^{r} v_i + \frac{1}{3}(1-\alpha)\sum_{j\in\mathbb{I}_k^v} a_j w_j,$$

which proves the first claim. The second case is similar, but we instead consider a solution

$$c_i = \begin{cases} h & \text{if } i \in \{r+1, \ldots, k\} \\ \hat{b}_i & \text{otherwise,} \end{cases}$$

where $0 < h < \min_{i\in\{r+1,\ldots,k\}} b_i$ and $h \leq 1$. Hence,

$$f(\hat{\mathbf{b}}) - f(\mathbf{c}) = \underbrace{\frac{1}{2}\sum_{i=r+1}^{k}\Big((y_i - \hat{b}_i)^2 - (y_i - c_i)^2\Big)}_{:=T_1} + \underbrace{\alpha \sum_{i=r+1}^{k} v_i(\hat{b}_i - c_i)}_{:=T_2}$$
$$\underbrace{+ (1-\alpha)\sum_{j\in\mathbb{I}_k^v} w_j\sqrt{p_j}(\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2)}_{:=T_3}.$$

As $\hat{b}_i = 0$ for $k \geq r+1$, the calculations are simpler:

$\underline{T_1}$:

$$T_1 = \frac{1}{2}\sum_{i=r+1}^{k}\Big((y_i - \hat{b}_i)^2 - (y_i - c_i)^2\Big) = h\sum_{i=r+1}^{k} y_i - \frac{1}{2}\sum_{i=r+1}^{k} h^2.$$

$\underline{T_2}$:

$$T_2 = \alpha \sum_{i=r+1}^{k} v_i(\hat{b}_i - c_i) = -\alpha h \sum_{i=r+1}^{k} v_i.$$

$\underline{T_3}$: Using the reverse triangle inequality, we obtain

$$\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2 \geq -h\sqrt{\tilde{a}_j}.$$

The key thing to note here is that $c_i = h$ and $b_i = 0$ for $i \in \{r+1, \ldots, k\}$, so the zeros have been replaced by a positive scalar. Therefore $\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2 \leq 0$, so that the sign has not changed across the inequality; therefore, zero could not have been used as a tighter bound. Hence,

$$T_3 = (1-\alpha) \sum_{j \in \mathbb{J}_k^v} w_j \sqrt{p_j}(\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2) \geq -(1-\alpha)h \sum_{j \in \mathbb{J}_k^v} w_j \sqrt{p_j}\sqrt{\tilde{a}_j}.$$

Therefore,

$$0 \geq f(\hat{\mathbf{b}}) - f(\mathbf{c}) \geq h \sum_{i=r+1}^{k} y_i - \frac{1}{2} \sum_{i=r+1}^{k} h^2 - \alpha h \sum_{i=r+1}^{k} v_i - (1-\alpha)h \sum_{j \in \mathbb{J}_k^v} w_j \sqrt{p_j}\sqrt{\tilde{a}_j}.$$

Dividing by $h$ and taking the limit as $h \to 0$ gives

$$\sum_{i=r+1}^{k} y_i \leq \alpha \sum_{i=r+1}^{k} v_i + (1-\alpha) \sum_{j \in \mathbb{J}_k^v} w_j \sqrt{p_j}\sqrt{\tilde{a}_j}.$$

□

**Lemma B.6**

*Proof of Lemma B.6.* Consider a new feasible (but suboptimal) solution

$$c_i = \begin{cases} \hat{b}_i - h & \text{if } i \in \mathbb{I}_k^g \\ \hat{b}_i & \text{otherwise,} \end{cases}$$

where $0 < h < \min_{i \in \mathbb{I}_k^g} b_i$. By optimality, we have $f(\hat{\mathbf{b}}) - f(\mathbf{c}) \leq 0$. Hence,

$$\begin{aligned} f(\hat{\mathbf{b}}) - f(\mathbf{c}) =& \frac{1}{2} \sum_{i=1}^{p} \left[ (y_i - \hat{b}_i)^2 - (y_i - c)^2 \right] + \alpha \sum_{i=1}^{p} v_i (\hat{b}_i - c_i) \\ &+ (1-\alpha) \sum_{j=1}^{m} w_j \sqrt{p_j} (\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2). \end{aligned}$$

By definition of $\mathbf{c}$, it follows

$$\begin{aligned} f(\hat{\mathbf{b}}) - f(\mathbf{c}) =& \frac{1}{2} \sum_{i \in \mathbb{I}_k^g} \left( (y_i - \hat{b}_i)^2 - (y_i - c)^2 \right) + \alpha \sum_{i \in \mathbb{I}_k^g} v_i (\hat{b}_i - c_i) \\ &+ (1-\alpha) \sum_{j=k}^{r} w_j \sqrt{p_j} (\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2). \end{aligned}$$

The first two terms are as in Lemma B.5, but with different summation indices. Hence, we provide calculations only for the final term. For this term, we make use of Bound (i) in Lemma B.7 with $m = p$, so that

$$\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2 \geq h\sqrt{p_j}.$$

Therefore,

$$T_3 = (1-\alpha) \sum_{j=k}^{r} w_j \sqrt{p_j} (\|\hat{\mathbf{b}}^{(j)}\|_2 - \|\mathbf{c}^{(j)}\|_2) \geq (1-\alpha) h \sum_{j=k}^{r} w_j p_j.$$

Combining the three terms back together, we have that

$$0 \geq f(\hat{\mathbf{b}}) - f(\mathbf{c}) \geq -\frac{1}{2} \sum_{i \in \mathbb{I}_k^g} h^2 - h \sum_{i \in \mathbb{I}_k^g} (y_i - \hat{b}_i) + \alpha h \sum_{i \in \mathbb{I}_k^g} v_i + (1-\alpha) h \sum_{j=k}^{r} w_j p_j.$$

We divide by $h$ and then take the limit as $h \to 0$ to obtain

$$\sum_{i \in \mathbb{I}_k^g} y_i - \sum_{i \in \mathbb{I}_k^g} \hat{b}_i - \alpha \sum_{i \in \mathbb{I}_k^g} v_i - (1-\alpha) \sum_{j=k}^{r} w_j p_j \geq 0.$$

Now, by assumption we have $\sum_{i\in\mathbb{I}_k^g}\hat{b}_i > 0$, so

$$\sum_{i\in\mathbb{I}_k^g} y_i - \alpha\sum_{i\in\mathbb{I}_k^g} v_i - (1-\alpha)\sum_{j=k}^{r} w_j p_j > 0 \implies \sum_{i\in\mathbb{I}_k^g} y_i > \alpha\sum_{i\in\mathbb{I}_k^g} v_i + (1-\alpha)\sum_{j=k}^{r} w_j p_j,$$

which proves the first claim. The second case is similar, but we instead consider a solution

$$c = \begin{cases} h & \text{if } i \in \mathbb{J}_k^g \\ \hat{b}_i & \text{otherwise,} \end{cases}$$

where $0 < h < \min_{i\in\mathbb{J}_k^g} b_i$ and $h \leq 1$. The calculation is the same as in Lemma B.5, but with different indices and $\sqrt{\tilde{a}_j}$ replaced by $\sqrt{p_j}$. Hence, we obtain

$$\sum_{i\in\mathbb{J}_k^g} y_i \leq \alpha\sum_{i\in\mathbb{J}_k^g} v_i + (1-\alpha)\sum_{j=r+1}^{k} w_j p_j.$$

□

### B.2.4 Derivation of Penalty Sequences

#### Variable Sequence

Lemma B.1 is used to define a penalty sequence for the variables. We aim to choose $v_r$ such that

$$\mathbb{P}(H_i^v \text{ rejected}) = \mathbb{P}\left(|y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j\right) \leq \frac{q_v r}{p}.$$

So,

$$\begin{aligned} &\mathbb{P}\left(|y_i| > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j\right) \leq \frac{q_v r}{p} \\ \implies &\mathbb{P}\left(y_i > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j\right) + \mathbb{P}\left(-y_i > \alpha v_r + (1-\alpha)w_j\right) \leq \frac{q_v r}{p}. \end{aligned}$$

As $y_i \sim \mathcal{N}(0,1)$, because $i \leq p_0$, we have by symmetry, for each $r \in [p]$,

$$\mathbb{P}\left(y_i > \alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j\right) \leq \frac{q_v r}{2p} \implies 1 - \Phi\left(\alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j\right) \leq \frac{q_v r}{2p},$$

where $\Phi$ is the standard Gaussian cumulative distribution function. Hence, we seek

$$\alpha v_r + \frac{1}{3}(1-\alpha)a_j w_j = \Phi^{-1}\left(1 - \frac{q_v r}{2p}\right)$$
$$\implies v_r = \frac{1}{\alpha}\Phi^{-1}\left(1 - \frac{q_v r}{2p}\right) - \frac{1}{3\alpha}(1-\alpha)a_j w_j,\ r \in [p].$$

However, as we have no knowledge of structure of group $j$ (see Remark 2.3), we take the sequence over the maximum possible, to ensure definite FDR control. This gives the vMax sequence as

$$v_i^{\max} = \max_{j=1,\dots,m}\left\{\frac{1}{\alpha}\Phi^{-1}\left(1 - \frac{q_v i}{2p}\right) - \frac{1}{3\alpha}(1-\alpha)a_j w_j\right\},\ i \in [p].$$

**Group Sequence**

Lemma B.3 defines a group penalty sequence, with $w_r$ chosen to satisfy

$$\mathbb{P}(H_j^g \text{ rejected}) = \mathbb{P}\left(\sum_{i\in G_j}|y_i| > \alpha\sum_{i\in G_r} v_i + (1-\alpha)w_r p_r\right) \leq \frac{q_g r}{m}.$$

For $r \in [m]$, this is given by

$$w_r = \frac{F_{\text{FN}}^{-1}\left(1 - \frac{q_g r}{m}\right) - \alpha\sum_{i\in G_r} v_i}{(1-\alpha)p_r}.$$

As with the variable sequence in Appendix B.2.4, taking the maximum gives the gMax sequence as

$$w_i = \max_{j=1,\dots,m}\left\{\frac{F_{\text{FN}}^{-1}\left(1 - \frac{q_g i}{m}\right) - \alpha\sum_{k\in G_i} v_k}{(1-\alpha)p_j}\right\},\ i \in [m].$$

### B.2.5 Norm Results

**Lemma B.7.** *For a vector $\mathbf{x} \in (\mathbb{R}^+)^p$, $p > 0$, suppose we construct another vector*

$$\mathbf{y} = \begin{cases} x_i - h & \text{if } i \in M \\ x_i & \text{otherwise,} \end{cases}$$

*where* $0 < h < \min_{i\in M} x_i, M \subset [p], |M| = m > 0$. *Then, the following two bounds hold*

$$(i)\ \ \|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 \geq h\sqrt{p} - h\sqrt{p-m} \geq \frac{hm}{3\sqrt{p}} \geq 0.$$

$$(ii)\ \ \|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 \geq \frac{h}{2}\frac{\|\mathbf{x}\|_{1,M}}{\|\mathbf{x}\|_2} \geq 0.$$

*For Bound (ii), we require a slightly stronger assumption on* $h$*:* $0 < h \leq \bar{x}/2$.

*Proof.* Bound (i): For $i \notin M$, we rewrite $y_i = x_i - h + h$ and denote vectors $\tilde{\mathbf{h}} \in \mathbb{R}^p$, where $\tilde{h}_i = h, \forall i \in M$, and $\boldsymbol{\eta} \in \mathbb{R}^p$ such that $\eta_i = h$ for $i \notin M$ and 0 otherwise. Then, we can rewrite $\mathbf{y}$ as $\mathbf{y} = \mathbf{x} - \tilde{\mathbf{h}} + \boldsymbol{\eta}$. Using the triangle inequality, we have

$$\|\mathbf{y}\|_2 = \|\mathbf{x} - \tilde{\mathbf{h}} + \boldsymbol{\eta}\|_2 \leq \|\mathbf{x} - \tilde{\mathbf{h}}\|_2 - \|\boldsymbol{\eta}\|_2.$$

Therefore,

$$\|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 \geq \|\mathbf{x}\|_2 - \|\mathbf{x} - \tilde{\mathbf{h}}\| - \|\boldsymbol{\eta}\|_2.$$

For any $i \in M$, it holds

$$x_i - h = x_i\left(1 - \frac{h}{x_i}\right) \leq x_i\left(1 - \frac{h}{\max_{i\in M} x_i}\right) = x_i\left(1 - \frac{h}{\|\mathbf{x}\|_\infty}\right),$$

so that $\|\mathbf{x} - \tilde{\mathbf{h}}\|_2 \leq \|\mathbf{x}(1 - h/\|\mathbf{x}\|_\infty)\|_2$. Therefore,

$$\|\mathbf{x}\|_2 - \|\mathbf{x} - \tilde{\mathbf{h}}\| + \|\boldsymbol{\eta}\|_2 \geq \|\mathbf{x}\|_2 - \left\|\mathbf{x}\left(1 - \frac{h}{\|\mathbf{x}\|_\infty}\right)\right\|_2 - \|\boldsymbol{\eta}\|_2.$$

Now, as $(1 - h/\|\mathbf{x}\|_\infty) \in \mathbb{R}$ is a real scalar, the reverse triangle inequality becomes an equality, so that

$$\|\mathbf{x}\|_2 - \left\|\mathbf{x}\left(1 - \frac{h}{\|\mathbf{x}\|_\infty}\right)\right\|_2 = \left\|\mathbf{x} - \mathbf{x}\left(1 - \frac{h}{\|\mathbf{x}\|_\infty}\right)\right\|_2 = \left\|\frac{h\mathbf{x}}{\|\mathbf{x}\|_\infty}\right\|_2 = \frac{h\|\mathbf{x}\|_2}{\|\mathbf{x}\|_\infty}.$$

Additionally, we have $\|\boldsymbol{\eta}\|_2 = \sqrt{\sum_{i\notin M} h^2} = \sqrt{(p-m)h^2} = h\sqrt{p-m}$. Therefore,

$$\|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 \geq \frac{h\|\mathbf{x}\|_2}{\|\mathbf{x}\|_\infty} - h\sqrt{p-m} = h\left(\frac{\|\mathbf{x}\|_2}{\|\mathbf{x}\|_\infty} - \sqrt{p-m}\right).$$

By Hölder's inequality [163], $\|\mathbf{x}\|_\infty \geq p^{-1/2}\|\mathbf{x}\|_2$, so it follows that

$$\|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 \geq h\left(\frac{\|\mathbf{x}\|_2}{p^{-1/2}\|\mathbf{x}\|_2} - \sqrt{p-m}\right) = h\sqrt{p} - h\sqrt{p-m}.$$

Rewriting $\sqrt{p} - \sqrt{p-m}$ as $\frac{m}{\sqrt{p}+\sqrt{p-m}}$, we obtain

$$\|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 \geq \frac{hm}{\sqrt{p}+\sqrt{p-m}} \geq \frac{hm}{2\sqrt{p}+\sqrt{m}} \geq \frac{hm}{3\sqrt{p}} \geq 0.$$

Bound (ii): $\|\mathbf{y}\|_2$ can be expressed as

$$\begin{aligned}
\|\mathbf{y}\|_2 &= \left(\sum_{i\notin M} x_i^2 + \sum_{i\in M}(x_i-h)^2\right)^{1/2} = \left(\sum_{i=1}^{p} x_i^2 - 2h\sum_{i\in M} x_i + mh^2\right)^{1/2} \\
&= \left(\|\mathbf{x}\|_2^2 + h\left(mh - 2\sum_{i\in M} x_i\right)\right)^{1/2} \\
&= \left(\|\mathbf{x}\|_2^2\left(1 + \frac{h}{\|\mathbf{x}\|_2^2}\left(mh - 2\sum_{i\in M} x_i\right)\right)\right)^{1/2} \\
&= \|\mathbf{x}\|_2\left(1 + \underbrace{\frac{h}{\|\mathbf{x}\|_2^2}\left(mh - 2\sum_{i\in M} x_i\right)}_{=:t}\right)^{1/2}.
\end{aligned}$$

Now, to apply Bernoulli's inequality [28], we require

$$t = \frac{h}{\|\mathbf{x}\|_2^2}\left(mh - 2\sum_{i\in M} x_i\right) \geq -1 \implies \sum_{i\in M} x_i^2 - 2h\sum_{i\in M} x_i + mh^2 \geq 0.$$

It is clear that $mh^2 \geq 0$, so we only require

$$\sum_{i\in M}(x_i^2 - 2hx_i) \geq 0 \implies \sum_{i\in M} x_i \geq 2hm \implies h \leq \frac{1}{2m}\sum_{i\in M} x_i = \bar{x}/2.$$

Hence, assuming $h \leq \bar{\mathbf{x}}/2$, Bernoulli's inequality is applied to obtain

$$\left(1 + \frac{h}{\|\mathbf{x}\|_2^2}\left(mh - 2\sum_{i\in M} x_i\right)\right)^{1/2} \leq 1 + \frac{h}{2\|\mathbf{x}\|_2^2}\left(mh - 2\sum_{i\in M} x_i\right).$$

Therefore,

$$\|\mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \frac{h}{2\|\mathbf{x}\|_2}\left(mh - 2\sum_{i\in M} x_i\right).$$

Hence,

$$\begin{aligned}\|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 &\geq \frac{h}{2\|\mathbf{x}\|_2}\left(2\sum_{i\in M} x_i - mh\right) \geq \frac{h}{2\|\mathbf{x}\|_2}\left(\sum_{i\in M} x_i\right), \quad \text{as} \sum_{i\in M} x_i - mh \geq 0 \\ &= \frac{h\|\mathbf{x}\|_{1,M}}{2\|\mathbf{x}\|_2}, \quad \text{where } \|\mathbf{x}\|_{1,M} = \sum_{i\in M} x_i.\end{aligned}$$

□

**Lemma B.8.** *Suppose we have a vector* $\mathbf{x} \in (\mathbb{R}^+)^p$, $p > 0$, *where* $M = \{i \in [p] : x_i = 0\}, |M| = m$, *and suppose further that we construct another vector*

$$\mathbf{y} = \begin{cases} h & \text{if } i \in M \\ x_i & \text{otherwise,} \end{cases}$$

*where* $0 < h < \min_{i\in M} x_i$, *and* $h \leq 1$. *Then, the following bound holds*

$$0 \geq \|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 \geq -\frac{hm}{2\|\mathbf{x}\|_2}.$$

*Proof.* The proof is similar to that of Lemma B.7. We again rewrite $\|\mathbf{y}\|_2$ as

$$\|\mathbf{y}\|_2 = \left(\sum_{i\notin M} x_i^2 + \sum_{i\in M} h^2\right)^{1/2} = \left(\|\mathbf{x}\|_2^2 + h^2 m\right)^{1/2} = \|\mathbf{x}\|_2\left(1 + \frac{h^2 m}{\|\mathbf{x}\|_2^2}\right)^{1/2}.$$

Now, as $f(x) = \sqrt{x}$ is a concave function, we can bound it from above using a linear approximation (obtained by use of Taylor's expansion [185]) to give

$$\left(1 + \frac{h^2 m}{\|\mathbf{x}\|_2^2}\right)^{1/2} \leq 1 + \frac{h^2 m}{2\|\mathbf{x}\|_2^2}.$$

Hence,

$$\|\mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \frac{h^2 m}{2\|\mathbf{x}\|_2} \implies \|\mathbf{x}\|_2 - \|\mathbf{y}\|_2 \geq \frac{-h^2 m}{2\|\mathbf{x}\|_2} \geq \frac{-hm}{2\|\mathbf{x}\|_2},$$

where the last inequality follows from $h \leq 1$, proving the result. □

### B.2.6 Numerical Experiments

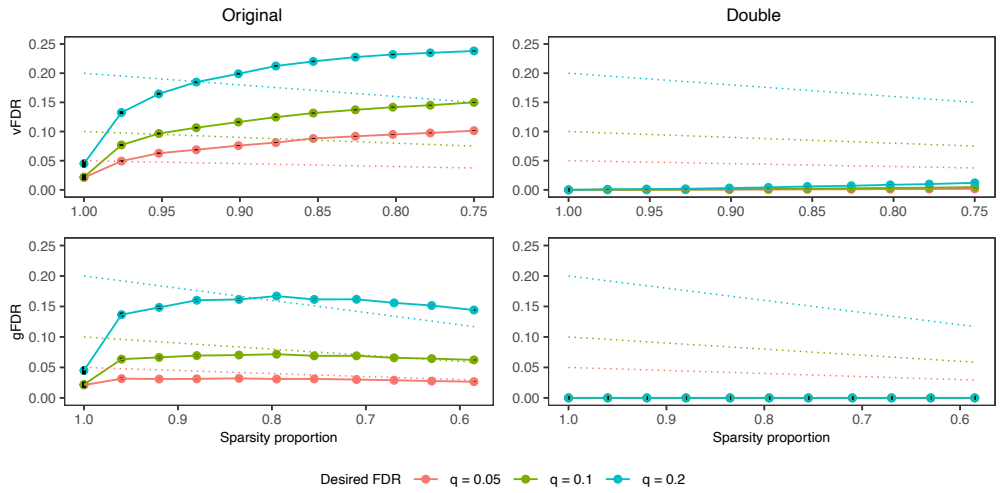



Figure B.1: vFDR and gFDR of SGS Original and SGS Double (using the max sequences) under orthogonal matrix $\mathbf{X} = \mathbf{I}_{1000}$, shown as a function of decreasing sparsity proportion. 200 equal groups of size five were generated, using signal effect $\beta_i = 5\delta\sqrt{2\log p}, i \in S_v, \delta \sim \mathcal{N}(0,1)$, with 1000 simulation repetitions per sparsity proportion. The dotted lines represent the theoretical upper bounds.

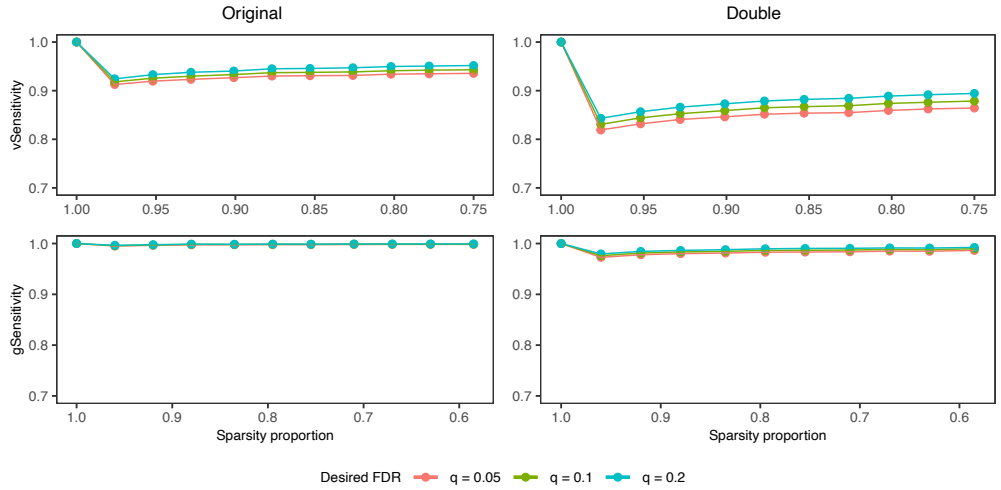



Figure B.2: vSensitivity and gSensitivity of SGS Original and SGS Double (using max sequences) under orthogonal matrix $\mathbf{X} = \mathbf{I}_{1000}$, shown as a function of decreasing sparsity proportion. 200 equal groups of size five were generated, using signal effect $\beta_i = 5\delta\sqrt{2\log p}, i \in S_v, \delta \sim \mathcal{N}(0,1)$, with 1000 simulation repetitions per sparsity proportion.

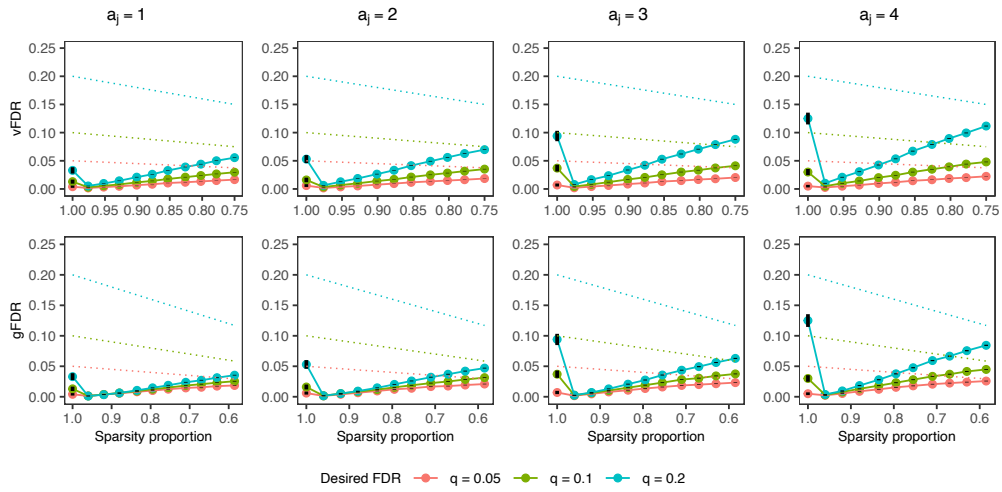



Figure B.3: vFDR and gFDR of SGS with vMax and gMax sequences under orthogonal matrix $\mathbf{X} = \mathbf{I}_{1000}$ with 200 equal groups of size 5. This is shown for different values of $a_j$, as a function of decreasing sparsity proportion. The signal effect was set at $\beta_i = 5\delta\sqrt{2\log p}, i \in S_v, \delta \sim \mathcal{N}(0,1)$, with 1000 simulation repetitions per sparsity proportion. The dotted lines represent the theoretical upper bounds.

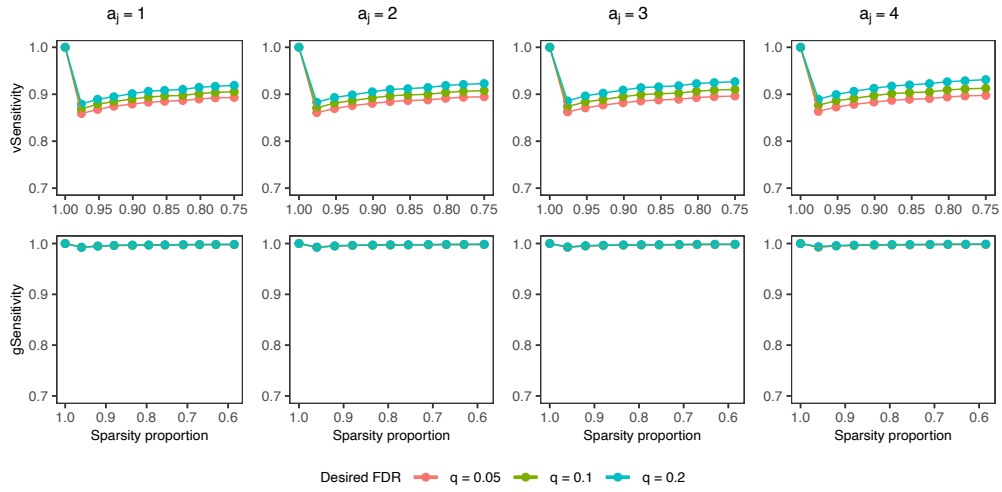



Figure B.4: vSensitivity and gSensitivity of SGS with vMax and gMax mean sequences under orthogonal matrix $\mathbf{X} = \mathbf{I}_{1000}$ with 200 equal groups of size 5. This is shown for different values of $a_j$, as a function of decreasing sparsity proportion. The signal effect was set at $\beta_i = 5\delta\sqrt{2\log p}, i \in S_v, \delta \sim \mathcal{N}(0,1)$, with 1000 simulation repetitions per sparsity proportion.

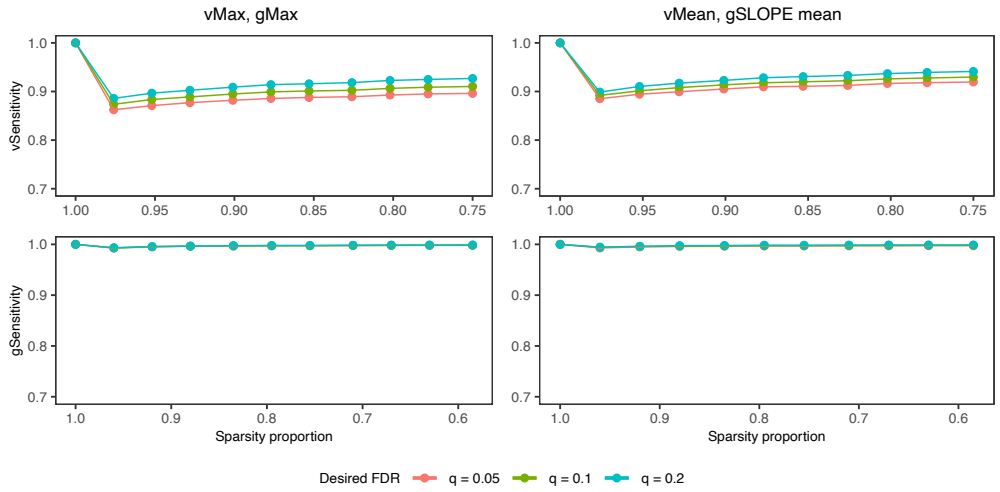



Figure B.5: vSensitivity and gSensitivity of SGS with the vMax, gMax, and the vMean, gSLOPE mean sequences under orthogonal matrix $\mathbf{X} = \mathbf{I}_{1000}$ with 200 equal groups of size five, as a function of decreasing sparsity proportion. The signal effect was set at $\beta_i = 5\delta\sqrt{2\log p}, i \in S_v$, $\delta \sim \mathcal{N}(0,1)$, with 1000 simulation repetitions per sparsity proportion.

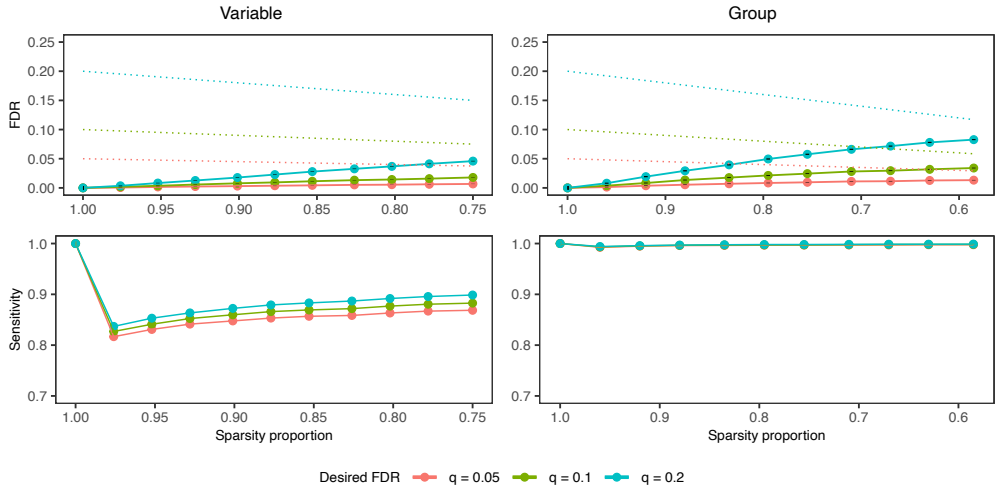



Figure B.6: Variable and group metrics of SGS with vMax and gMax sequences, using $\alpha = 0.95$, under orthogonal matrix $\mathbf{X} = \mathbf{I}_{1000}$ with 200 equal groups of size five. This is shown as a function of decreasing sparsity proportion. The signal effect was set at $\beta_i = 5\delta\sqrt{2\log p}, i \in S_v, \delta \sim \mathcal{N}(0,1)$, with 1000 simulation repetitions per sparsity proportion. The dotted lines represent the theoretical upper bounds.

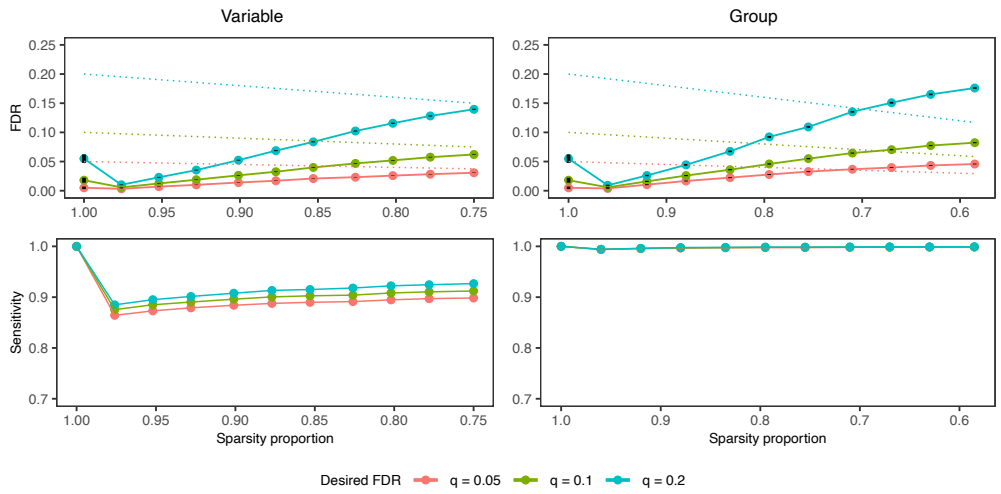



Figure B.7: Variable and group metrics of SGS with vMean and gMean sequences under orthogonal matrix $\mathbf{X} = \mathbf{I}_{1000}$ with 200 equal groups of size five. This is shown as a function of decreasing sparsity proportion. The signal effect was set at $\beta_i = 5\delta\sqrt{2\log p}, i \in S_v$, $\delta \sim \mathcal{N}(0,1)$, with 1000 simulation repetitions per sparsity proportion. The dotted lines represent the theoretical upper bounds.

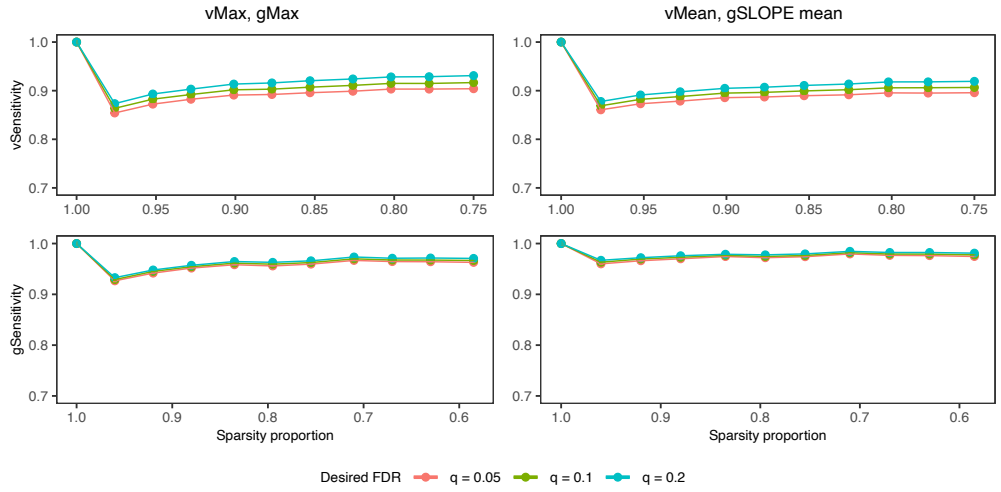



Figure B.8: vSensitivity and gSensitivity of SGS with the vMax, gMax and the vMean, gSLOPE mean sequences under orthogonal matrix $\mathbf{X} = \mathbf{I}_{1000}$ with 40 uneven groups of sizes $\{3, \ldots, 7\}$. This is shown as a function of decreasing sparsity proportion. The signal effect was set at $\beta_i = 5\delta\sqrt{2\log p}, i \in S_v$, $\delta \sim \mathcal{N}(0,1)$, with 1000 simulation repetitions per sparsity proportion.

## B.3 Synthetic Data Experiments

### B.3.1 Results

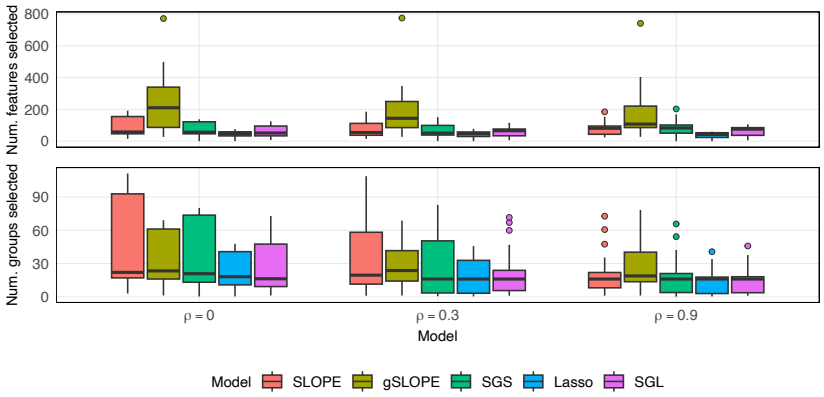



Figure B.9: The number of features (top) and groups (bottom) selected for SLOPE- and lasso-based models, for the three different correlation cases. Results were averaged over six sparsity cases with 600 simulations per correlation case.

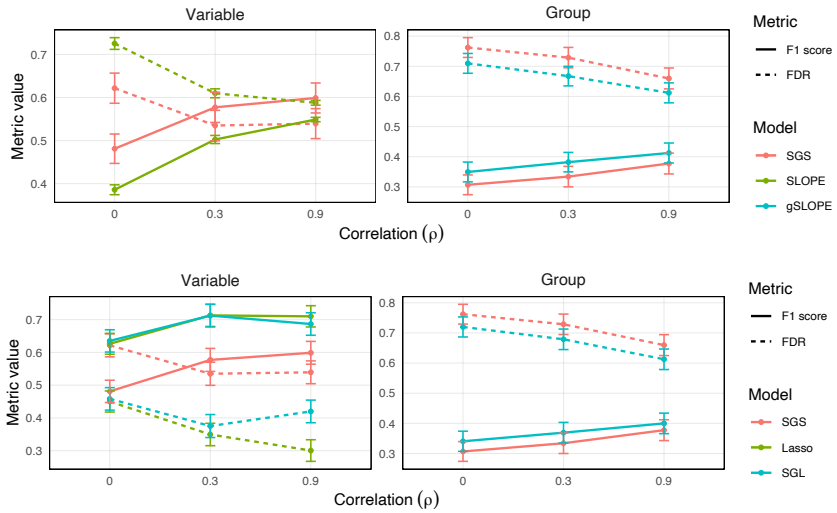



Figure B.10: $F_1$ score and FDR of the SLOPE-based (top) and lasso-based (bottom) models under $\beta_i = 5, i \in S_v$, for $\rho = 0, 0.3, 0.9$ within a group and $\rho = 0.05$ across groups, with standard errors shown. The results were averaged over the six sparsity cases, with the variable and group sparsity proportions at 0.91 and 0.92. 160 groups of sizes $\{3, \dots, 7\}$ were used and 600 simulation repetitions per correlation case.

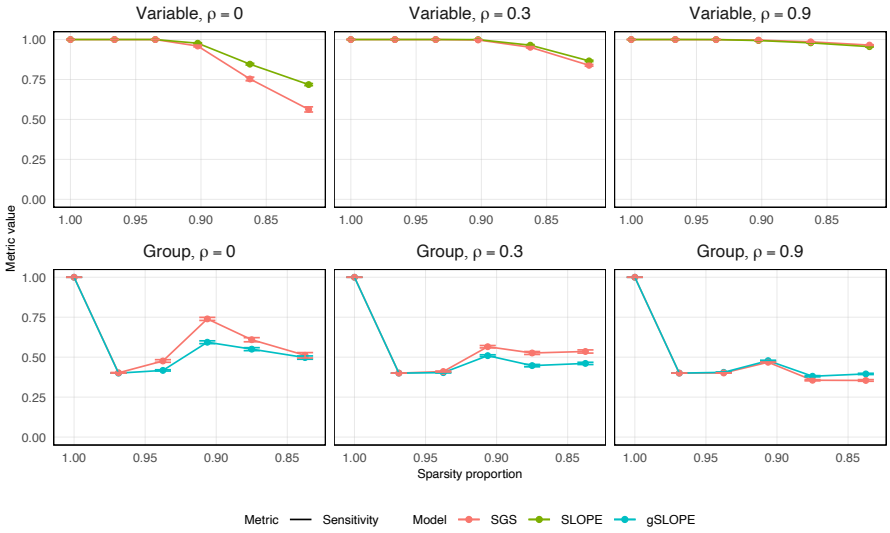



Figure B.11: Sensitivity as a function of decreasing sparsity proportion for the SLOPE-based models, under $\beta_i = 5, i \in S_v$, with 160 groups of sizes $\{3, \dots, 7\}$. Standard errors are shown; 100 simulations were run for each case.

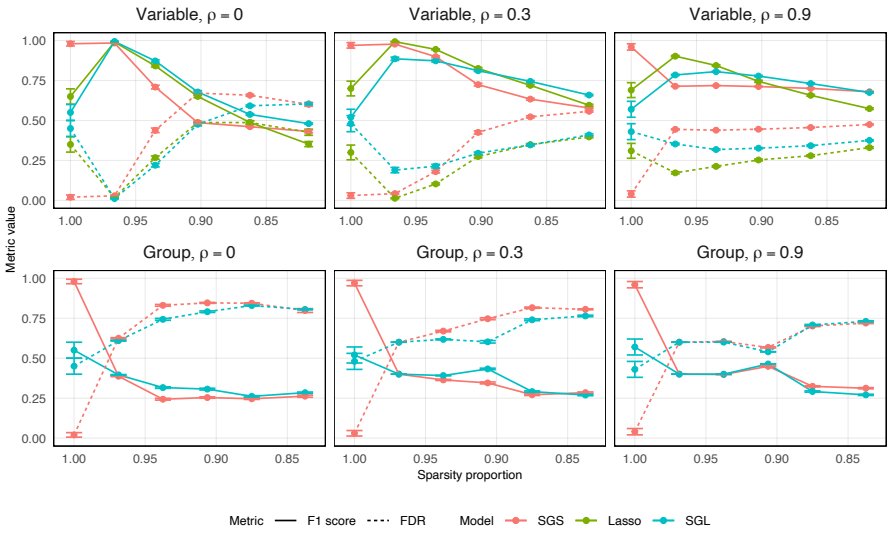



Figure B.12: $F_1$ score and FDR as functions of decreasing sparsity proportion for the lasso-based models and SGS, under $\beta_i = 5, i \in S_v$, with 160 groups of sizes $\{3, \dots, 7\}$. Standard errors are shown; 100 simulations were run for each case.

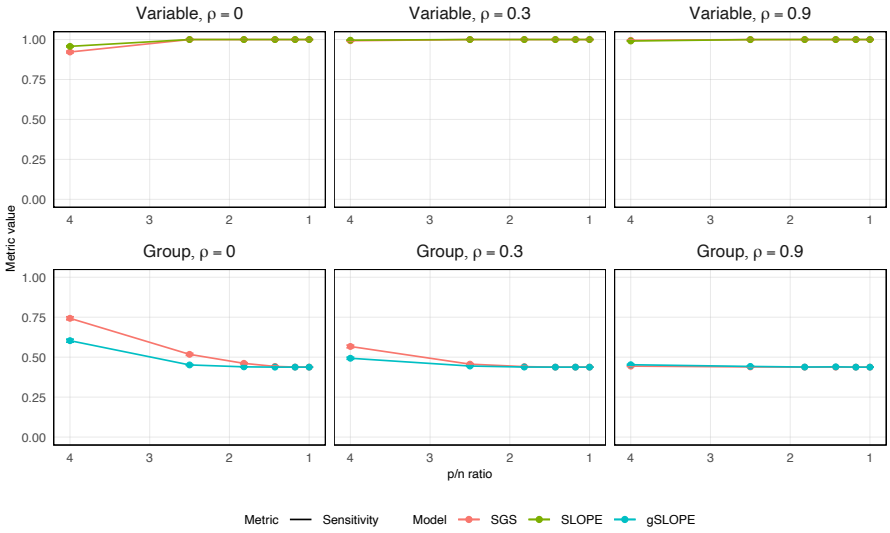



Figure B.13: Average sensitivity across 100 simulation iterations as a function of decreasing $p/n$ ratio, for the SLOPE-based models under $\beta_i = 5, i \in S_v$. A single sparsity case was considered: variable and group proportions were 0.82 and 0.84, with 160 groups of sizes $\{3, \ldots, 7\}$.

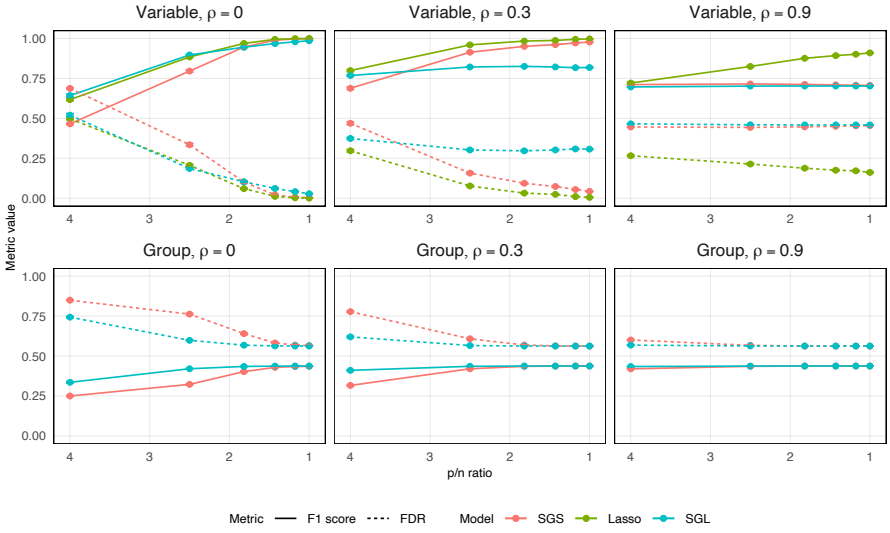



Figure B.14: Average $F_1$ score and FDR across 100 simulation iterations as functions of decreasing $p/n$ ratio, for the lasso-based models and SGS under $\beta_i = 5, i \in S_v$. A single sparsity case was considered: variable and group proportions were 0.82 and 0.84, with 160 groups of sizes $\{3, \ldots, 7\}$.

| Model | $\beta$ error | | Variable mean | | | Group mean | | |
|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | $F_1$ | FDR | Sens. | $F_1$ | FDR | Sens. |
| SGS | 0.48 $(10^{-3})$ | **0.16** $(10^{-4})$ | **0.74** $(10^{-3})$ | 0.36 $(10^{-3})$ | 0.94 $(10^{-3})$ | **0.44** $(10^{-3})$ | **0.60** $(10^{-3})$ | **0.56** $(10^{-3})$ |
| SLOPE | 0.43 $(10^{-3})$ | **0.16** $(10^{-4})$ | 0.57 $(10^{-3})$ | 0.53 $(10^{-3})$ | **0.96** $(10^{-4})$ | – | – | – |
| gSLOPE | **0.35** $(10^{-3})$ | **0.16** $(10^{-4})$ | – | – | – | 0.28 $(10^{-4})$ | 0.76 $(10^{-4})$ | 0.54 $(10^{-3})$ |
| Lasso | 0.65 $(10^{-3})$ | **0.16** $(10^{-4})$ | 0.73 $(10^{-3})$ | **0.28** $(10^{-3})$ | 0.83 $(10^{-3})$ | – | – | – |
| SGL | 90.78 (0.39) | 1.91 $(10^{-3})$ | 0.72 $(10^{-3})$ | 0.36 (0.01) | 0.91 $(10^{-3})$ | 0.38 $(10^{-3})$ | 0.65 $(10^{-3})$ | 0.52 $(10^{-3})$ |

Table B.1: MSE, MAE, $F_1$ score, FDR, and sensitivity under $\beta_i = 5, i \in S_v$, averaged over all correlation and sparsity cases, for SLOPE- and lasso-based models, with standard errors shown in grey. Variable and group sparsity proportions were 0.91 and 0.92, with 160 groups of sizes $\{3, \dots, 7\}$. In total, 1800 simulation repetitions were performed. **Bold** represents the best value for each metric.

| Model | $\beta$ error | | Variable mean | | | Group mean | | |
|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | $F_1$ | FDR | Sens. | $F_1$ | FDR | Sens. |
| SGL | 58.63 (0.86) | 1.40 (0.02) | 0.72 (0.01) | 0.36 (0.01) | **0.98** $(10^{-3})$ | **0.82** (0.01) | **0.18** (0.01) | **0.96** $(10^{-3})$ |
| Lasso | **0.43** (0.01) | **0.10** $(10^{-3})$ | **0.77** (0.01) | **0.25** (0.01) | 0.88 $(10^{-3})$ | 0.80 (0.01) | 0.20 (0.01) | **0.96** $(10^{-3})$ |

Table B.2: MSE, MAE, $F_1$ score, FDR, and sensitivity under $\beta_i = 5, i \in S_v$, averaged over all correlation and sparsity cases, for lasso-based models, with standard errors shown in grey. The variable and group sparsity proportions were 0.97 and 0.92, with 25 groups of sizes $\{5, \dots, 75\}$. In total, 1800 simulation repetitions were performed. **Bold** represents the best value obtained for each metric.

## B.4 Real Data Experiments

### B.4.1 Fitting information

The data was $\ell_2$-standardised for each model. For the SGS, gSLOPE, SGO, and gOSCAR models, no intercept was fit. The other models had an intercept fit. Applying an intercept centres the response. However, as the response is binary, applying an intercept is not as important for the optimisation as for a continuous response. All models were tested with and without an intercept, and the best-performing outcome was used in the final results.

### B.4.2 Results

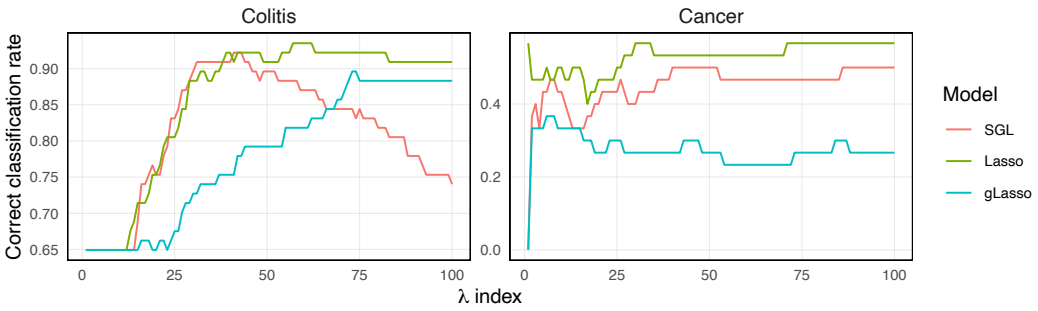



Figure B.15: Correct classification rate (%) of the lasso-based models applied to the Colitis and Cancer datasets, along a log-linear 100-$\lambda$ regularisation path.

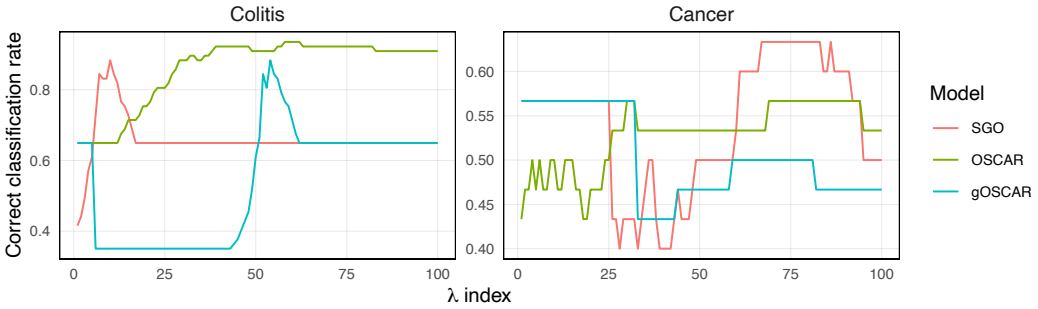



Figure B.16: Correct classification rate (%) of the OSCAR-based models applied to the Colitis and Cancer datasets, along a log-linear 100-$\lambda$ regularisation path.

| **Symbol** | **Gene name** | **Pathway** | $\hat{\beta}$ |
|---|---|---|---|
| COX6A1 | Cytochrome C Oxidase Subunit 6A1 | M40014 | −0.678 |
| SUSD3 | Sushi Domain Containing 3 | M40023 | −0.665 |
| TRIM46 | Tripartite Motif Containing 46 | M39067 | −0.656 |
| MMP10 | Matrix Metallopeptidase 10 | M41652 | −0.638 |
| CROCC | Ciliary Rootlet Coiled-Coil, Rootletin | M39136 | −0.360 |
| CD320 | CD320 Molecule | M39018 | 0.336 |
| RAP1GAP2 | RAP1 GTPase Activating Protein 2 | M40014 | −0.315 |
| SLC37A1 | Solute Carrier Family 37 Member 1 | M39064 | 0.307 |
| ACCS | 1-Aminocyclopropane-1-Carboxylate Synthase Homolog (Inactive) | M45728 | −0.279 |
| CABLES2 | Cdk5 And Abl Enzyme Substrate 2 | M39070 | 0.236 |

Table B.3: The top ten active genes as found by the optimal SGS solution for the Cancer (C8) dataset, given with their estimated coefficient value.

| **Model** | **Cancer** | | **Colitis** | |
|---|---|---|---|---|
| | Num. features | Num. groups | Num. features | Num. groups |
| SGS | 32 | 20 | 9 | 7 |
| SLOPE | 39 | 30 | 16 | 14 |
| gSLOPE | 0 | 0 | 1231 | 2 |
| SGL | 27 | 19 | 9 | 7 |
| Lasso | 0 | 0 | 11 | 10 |
| gLasso | 3 | 3 | 24 | 19 |
| SGO | 30 | 19 | 1 | 1 |
| OSCAR | 22 | 17 | 12 | 7 |
| gOSCAR | 0 | 0 | 1227 | 11 |

Table B.4: Number of selected features and groups in the optimal model for the Cancer and Colitis datasets.

| | SGS | Hyperparameters | | | Dataset information | | | |
|---|---|---|---|---|---|---|---|---|
| **Gene set** | **Peak classification (%)** | $\alpha$ | $q_v$ | $q_g$ | **# genes** | **# pathways** | **Pathway sizes** | **Avg. pathway size** |
| | | | | *Colitis dataset* | | | | |
| C1 | 93.5 | 0.95 | 0.15 | $1 \times 10^{-10}$ | 12321 | 292 | [1,470] | 42 |
| C2 | 94.8 | 0.95 | $1 \times 10^{-10}$ | $1 \times 10^{-10}$ | 12091 | 1193 | [1,888] | 10 |
| C3 | 97.4 | 0.99 | $1 \times 10^{-4}$ | $1 \times 10^{-10}$ | 12031 | 1408 | [1,723] | 9 |
| C4 | 94.8 | 0.99 | 0.05 | $1 \times 10^{-5}$ | 8482 | 613 | [1,287] | 14 |
| C5 | 93.5 | 0.99 | 0.05 | $1 \times 10^{-10}$ | 11555 | 614 | [1,1034] | 19 |
| C6 | 93.5 | 0.99 | $1 \times 10^{-5}$ | $1 \times 10^{-5}$ | 8749 | 185 | [1,169] | 47 |
| C7 | 96.1 | 0.99 | 0.15 | $1 \times 10^{-10}$ | 12084 | 936 | [1,172] | 13 |
| C8 | 94.8 | 0.95 | $1 \times 10^{-5}$ | $1 \times 10^{-10}$ | 11027 | 601 | [1,1007] | 18 |
| H | 97.4 | 0.90 | $1 \times 10^{-10}$ | $1 \times 10^{-10}$ | 3988 | 50 | [8,193] | 80 |
| | | | | *Cancer dataset* | | | | |
| C1 | 63.3 | 0.95 | 0.20 | $1 \times 10^{-10}$ | 7233 | 287 | [1,338] | 25 |
| C2 | 63.3 | 0.99 | $1 \times 10^{-10}$ | $1 \times 10^{-10}$ | 7145 | 1041 | [1,449] | 7 |
| C3 | 66.7 | 0.99 | 0.01 | $1 \times 10^{-10}$ | 7088 | 1132 | [1,449] | 6 |
| C4 | 56.6 | 0.01 | $1 \times 10^{-10}$ | 0.001 | 4106 | 475 | [1,140] | 9 |
| C5 | 60.0 | 0.95 | $1 \times 10^{-10}$ | $1 \times 10^{-10}$ | 6636 | 548 | [1,546] | 12 |
| C6 | 63.3 | 0.95 | 0.15 | $1 \times 10^{-10}$ | 4529 | 183 | [2,84] | 25 |
| C7 | 60.0 | 0.01 | $1 \times 10^{-10}$ | $1 \times 10^{-10}$ | 7163 | 896 | [1,83] | 8 |
| C8 | 66.7 | 0.99 | $1 \times 10^{-5}$ | $1 \times 10^{-10}$ | 6375 | 550 | [1,533] | 12 |
| H | 53.3 | 0.01 | $1 \times 10^{-10}$ | $1 \times 10^{-10}$ | 583 | 217 | [1,18] | 3 |

Table B.5: Peak correct classification rate (%) for SGS applied to the Colitis and Cancer datasets, shown for all gene sets with corresponding hyperparameters and dataset information.

# APPENDIX C

# Additional Material for Chapter 4

## C.1 Group SLOPE

### C.1.1 SLOPE Subdifferential

The subdifferential for SLOPE was derived in Larsson et al. [111] and, as it is a vital part of our arguments, it is reproduced in Definition C.1 for ease of reference. The primary use of the subdifferential is in checking the zero condition (Equation 4.6).

**Definition C.1** (SLOPE subdifferential [111])**.** Define the set $\mathcal{C}_i(\boldsymbol{\beta}) = \{j \in [p] : |\beta_i| = |\beta_j|\}$. Then, the subdifferential is given by

$$\partial J_{\text{slope}}(\boldsymbol{\beta}; \mathbf{v}) = \begin{cases} \left\{\mathbf{x} \in \mathbb{R}^{|\mathcal{C}_i|} : \operatorname{cumsum}(|\mathbf{x}|_\downarrow - \mathbf{v}_{r_v(\mathbf{x})_{\mathcal{C}_i}}) \preceq \mathbf{0}\right\} & \text{at zero,} \\ \Big\{\mathbf{x} \in \mathbb{R}^{|\mathcal{C}_i|} : \operatorname{cumsum}(|\mathbf{x}|_\downarrow - \mathbf{v}_{r_v(\mathbf{x})_{\mathcal{C}_i}}) \preceq \mathbf{0} \\ \quad \text{and } \textstyle\sum_{j \in \mathcal{C}_i}(|x_j| - \mathbf{v}_{r_v(s)_j}) = 0 \\ \quad \text{and } \operatorname{sign}(\boldsymbol{\beta}_{\mathcal{C}_i}) \equiv \operatorname{sign}(\mathbf{x})\Big\} & \text{otherwise.} \end{cases}$$

## C.2 Sparse-group SLOPE

### C.2.1 Theory

*Proof of Proposition 4.4.* The proof is similar to that of Proposition 4.1. Suppose we have $\mathcal{B} \neq \varnothing$ after running the algorithm. Then,

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

so that by the SGS subdifferential (Equation 4.8) all groups in $\mathcal{B}$ are inactive. Hence, $\mathcal{S}_g(\lambda_{k+1})$ will contain the active set $\mathcal{A}_g(\lambda_{k+1})$. □

*Proof of Proposition 4.5.* The proof is identical to that of Proposition 4.2, replacing $h(\cdot)$ with $\tilde{h}(\cdot)$ and $\lambda_{k+1}\mathbf{w}$ by $\lambda_{k+1}(1-\alpha)\mathbf{w}$. □

*Proof of Proposition 4.6.* Suppose we have $\mathcal{B} \neq \varnothing$ after running the algorithm. Then, we have

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

so that by the SGS subdifferential for non-zero groups (Equation 4.9) all variables in $\mathcal{B}$ are inactive. Hence, $\mathcal{S}_v(\lambda_{k+1})$ will contain the active set $\mathcal{A}_v(\lambda_{k+1})$. □

*Proof for Proposition 4.7.* The proof is identical to that of Proposition 4.2, replacing $h(\cdot)$ with $\bar{h}(\cdot)$, $\lambda_{k+1}\mathbf{v}$ with $\lambda_{k+1}\alpha\mathbf{v}$, and considering only variables in the groups contained in $\mathcal{A}_g(\lambda_{k+1})$. □

### C.2.2 Alternative KKT Checks

An alternative approach for SGS is to check the KKT conditions directly on the variables. The KKT conditions (Equation 4.7) can be rewritten as

$$-\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) - \lambda_{k+1}(1-\alpha)\partial J_{\text{gslope}}(\hat{\boldsymbol{\beta}}(\lambda_{k+1}); \mathbf{w}) \in \lambda_{k+1}\alpha\partial J_{\text{slope}}(\hat{\boldsymbol{\beta}}(\lambda_{k+1}); \mathbf{v}).$$

A KKT violation occurs when the zero subdifferential conditions are not satisfied:

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

The objective is to make the term inside the sorted absolute value operator as small as possible, given that the subdifferential term is unknown. To do this, a similar derivation as in Lemma 4.1 can be used to determine that the term must be the soft thresholding operator, so that a violation occurs if

$$
\operatorname{cumsum}\left(\left|S(\nabla f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})), \lambda_{k+1}(1-\alpha)\boldsymbol{\tau}\boldsymbol{\omega})\right|_{\downarrow} - \lambda_{k+1}\alpha\mathbf{v}\right) \succ \mathbf{0},
$$

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

## C.3 Screening Rule Framework

### C.3.1 Group SLOPE Algorithm

For $k = 1, \dots, l-1$:

1. Set $\mathcal{E}_g = \mathcal{S}_g(\lambda_{k+1}) \cup \mathcal{A}_g(\lambda_k)$, where $\mathcal{S}_g(\lambda_{k+1})$ is obtained using Proposition 4.2.
2. Compute $\hat{\boldsymbol{\beta}}(\lambda_{k+1})$ using Equation 2.1 with the gSLOPE norm (Equation 2.4) using only the groups in $\mathcal{E}_g$. For any groups not in $\mathcal{E}_g$, $\hat{\boldsymbol{\beta}}(\lambda_{k+1})$ is set to zero.
3. Check the KKT conditions (Equation 4.2) for all groups at this solution.
4. If there are no violations, we are done and keep $\hat{\boldsymbol{\beta}}(\lambda_{k+1})$. Otherwise, add the violating groups to $\mathcal{E}_g$ and return to Step 2.

### C.3.2 SGS Algorithm

For $k = 1, \dots, l-1$:

1. *Group screen step*: Calculate $\mathcal{S}_g(\lambda_{k+1})$ using Proposition 4.5.

2. *Variable screen step*: Set $\mathcal{E}_v = \mathcal{S}_v(\lambda_{k+1}) \cup \mathcal{A}_v(\lambda_k)$, where $\mathcal{S}_v(\lambda_{k+1})$ is obtained using Proposition 4.7 with only the groups in $\mathcal{S}_g(\lambda_{k+1})$.

3. Compute $\hat{\boldsymbol{\beta}}(\lambda_{k+1})$ by Equation 2.1 with the SGS norm (Equation 3.1) using only the variables in $\mathcal{E}_v$. For variables not in $\mathcal{E}_v$, $\hat{\boldsymbol{\beta}}(\lambda_{k+1})$ is set to zero.

4. Check the KKT conditions (Equation 4.7) for all variables at this solution.

5. If there are no violations, we are done and keep $\hat{\boldsymbol{\beta}}(\lambda_{k+1})$, otherwise add the violating variables to $\mathcal{E}_v$ and return to Step 3.

## C.4 Group-based OSCAR

This section provides supplementary materials for extending the proposed screening rules to group-based OSCAR models. SGO used $\alpha = 0.95$ and hyperparameters $a_1$ and $a_2$ were set as described in Section 3.2.1. gOSCAR used the group weights of SGO.

### C.4.1 Synthetic Data Results

Observations and conclusions made for the screening rules of gSLOPE and SGS are also found to be true for gOSCAR and SGO (Figures C.1–C.5).

Figure C.1 illustrates the impact of bi-level screening for SGO, similar to that observed for SGS. Figures C.2 and C.3 showcase the efficiency of the screening rules on the proportion of the selected groups/variables. The screening rules are found to be effective across different data characteristics, with the running time of the models significantly decreasing (Figure C.4). KKT violations for SGO are more common compared to gOSCAR (Figure C.5), due to the additional assumptions made at the second screening layer (as with SGS). Similar to Figure 4.6, the shape of the increasing number of KKT violations mirrors the log-linear shape of the regularisation path.

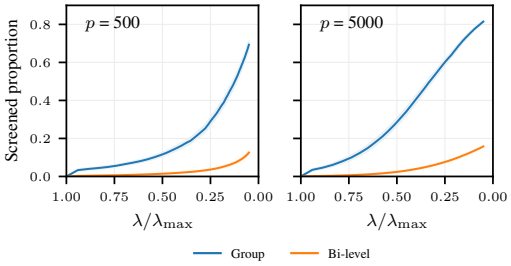



Figure C.1: The proportion of variables in $\mathcal{S}_v$ relative to $p$ for group-only and bi-level screening applied to SGO, for $\rho = 0.3$, plotted along the regularisation path with 95% confidence intervals. (the SGO equivalent of Figure 4.1).

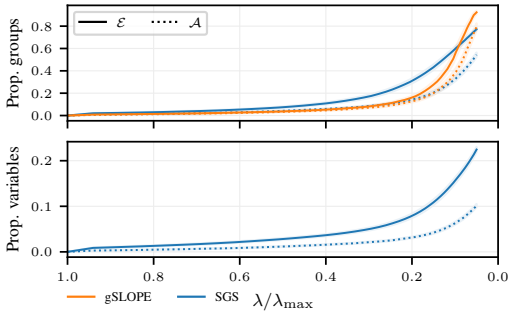



Figure C.2: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$ for both gOSCAR and SGO as a function of the regularisation path for the linear model with $p = 2750, \rho = 0.6, m = 197$. The results are averaged over 100 repetitions, with the shaded regions corresponding to 95% confidence intervals (the gOSCAR/SGO equivalent of Figure 4.2).

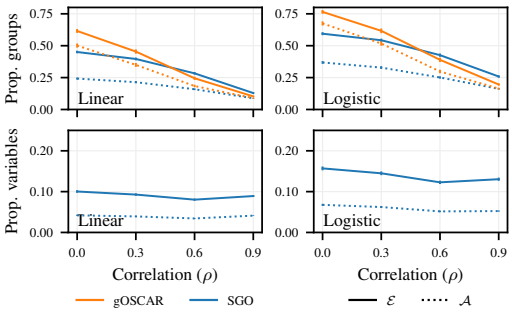



Figure C.3: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$, relative to the full input, shown for gOSCAR and SGO. This is shown as a function of the correlation ($\rho$), averaged over all cases of the input dimension ($p$), with 100 repetitions for each $p$, for both linear and logistic models, with standard errors shown (the gOSCAR/SGO equivalent of Figure 4.4).

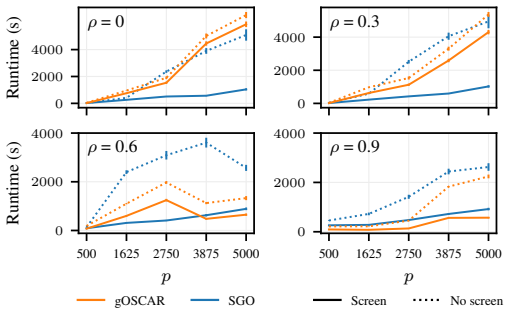



Figure C.4: Runtime (in seconds) for fitting 50 models along a path, shown for screening against no screening as a function of input dimensionality ($p$), broken down into different correlation cases, for the linear model. The results are averaged over 100 repetitions, with standard errors shown (the gOSCAR/SGO equivalent of Figure 4.5).

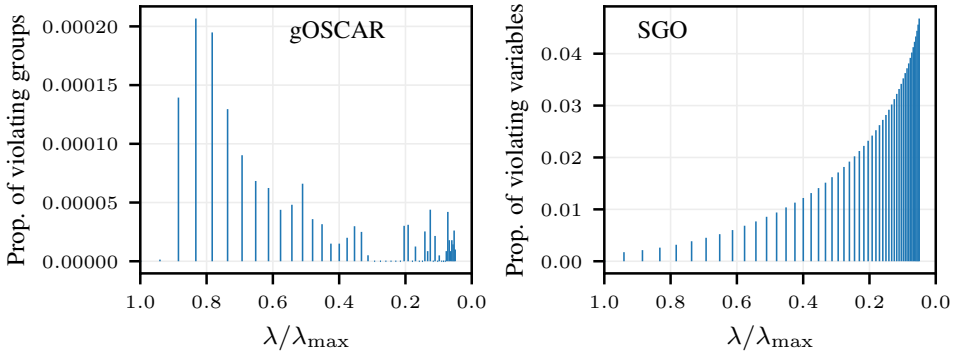



Figure C.5: The proportion of KKT violations relative to the full input for gOSCAR and SGO, under linear models, averaged over all synthetic data cases. This is shown as a function of the regularisation path (the gOSCAR/SGO equivalent of Figure 4.6).

### C.4.2 Real Data Results

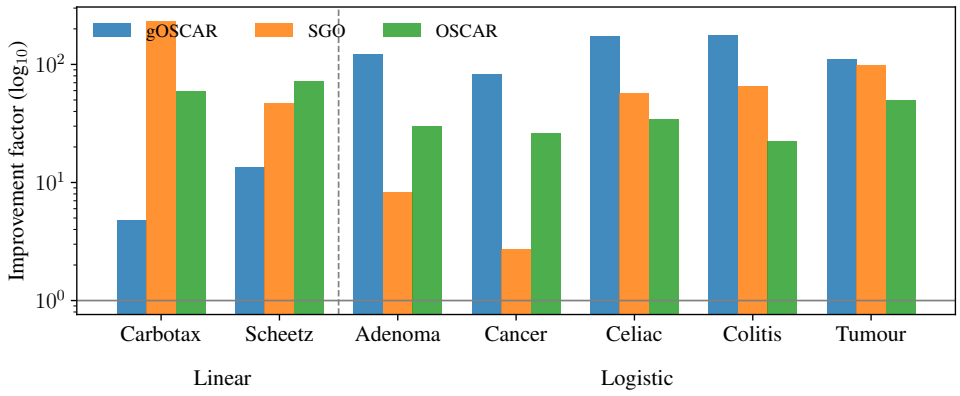



Figure C.6: The improvement factor ($\log_{10}$ scale) of OSCAR, gOSCAR, and SGO applied to the real datasets, for fitting 100 path models, split into response type.

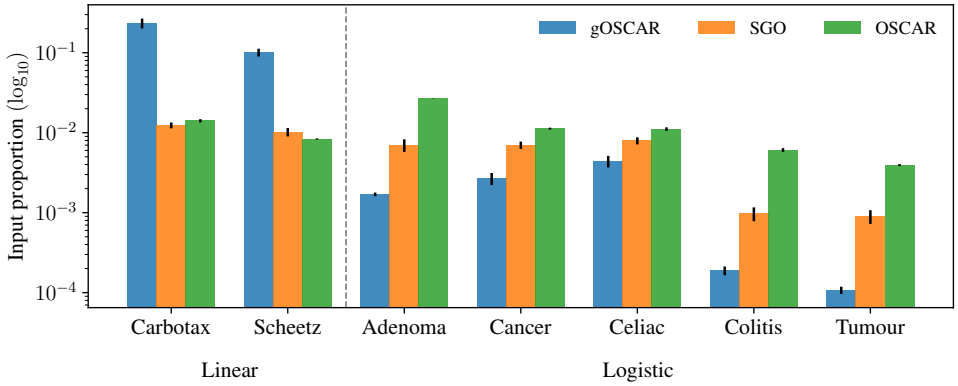



Figure C.7: The input proportion ($\log_{10}$ scale) of OSCAR, gOSCAR, and SGO applied to the real datasets, for fitting 100 path models, split into response type.

## C.5 Data Experiments

### C.5.1 Metrics

The following metrics are shown in the results tables in this section and Appendix D.3:

- $\mathcal{A}_v, \mathcal{A}_g$: the number of active variables/groups.
- $\mathcal{S}_v, \mathcal{S}_g$: the number of variables/groups in the screened sets.
- $\mathcal{E}_v, \mathcal{E}_g$: the number of variables/groups used in the optimisation process. As per Algorithms 2, $\mathcal{E}_v = \mathcal{S}_v \cup \mathcal{A}_v$. However, $\mathcal{E}_g$ is not produced as $\mathcal{E}_g = \mathcal{S}_g \cup \mathcal{A}_g$. Instead, $\mathcal{E}_g$ are the groups for which there are variables present in $\mathcal{E}_v$ to give a measure of the number groups used in the optimisation.
- $\mathcal{K}_v, \mathcal{K}_g$: the number of variable/group KKT violations.
- $\mathcal{E}_v/\mathcal{A}_v$ and $\mathcal{E}_g/\mathcal{A}_g$: the proportion of variables/groups used in the optimisation against the number active. Defines how efficient the rules are. A low value is best.
- $\mathcal{E}_v$ / $p$ and $\mathcal{E}_g$ / $m$: the variable/group input proportion. A low value is best.
- $\ell_2$ distance to no screen: $\ell_2$ distance from the fitted values obtained with screening to without.
- IF: the improvement factor.

### C.5.2 Solution Optimality

This section presents the accuracy of the models with and without screening, by comparing the $\ell_2$ distances observed between the screened and non-screened fitted values.

#### Synthetic Data

For the linear model, the maximum $\ell_2$ distances observed between the screened and non-screened fitted values were of order $10^{-6}$ for gSLOPE and $10^{-9}$ for SGS (Table C.2). Across the different cases, 98000 models were fit in total for each approach (excluding the models for $\lambda_1$, where no screening is applied). Of these model fits, there were no instances for gSLOPE where $\mathcal{E}_g$ was not a superset of $\mathcal{A}_g$. There was only one instance in which this occurred for SGS, where $\mathcal{E}_v$ was missing a single variable contained in $\mathcal{A}_v$ (which had a non-screen fitted value of $\hat{\beta} = -0.004$).

For the logistic model, the maximum $\ell_2$ distances observed between the screened and non-screened fitted values were of order $10^{-8}$ for gSLOPE and $10^{-9}$ for SGS. Across the different cases, there were no instances for gSLOPE or SGS where $\mathcal{E}$ was not a superset of $\mathcal{A}$.

#### Real Data

In the real data analysis, the estimated coefficients with and without screening were very close to each other for both SGS and gSLOPE (Table C.4). However, direct comparison is less meaningful here, as the models often failed to converge without screening, therefore not reaching the optimal solution.

### C.5.3 Additional Results from the Synthetic Data Experiments

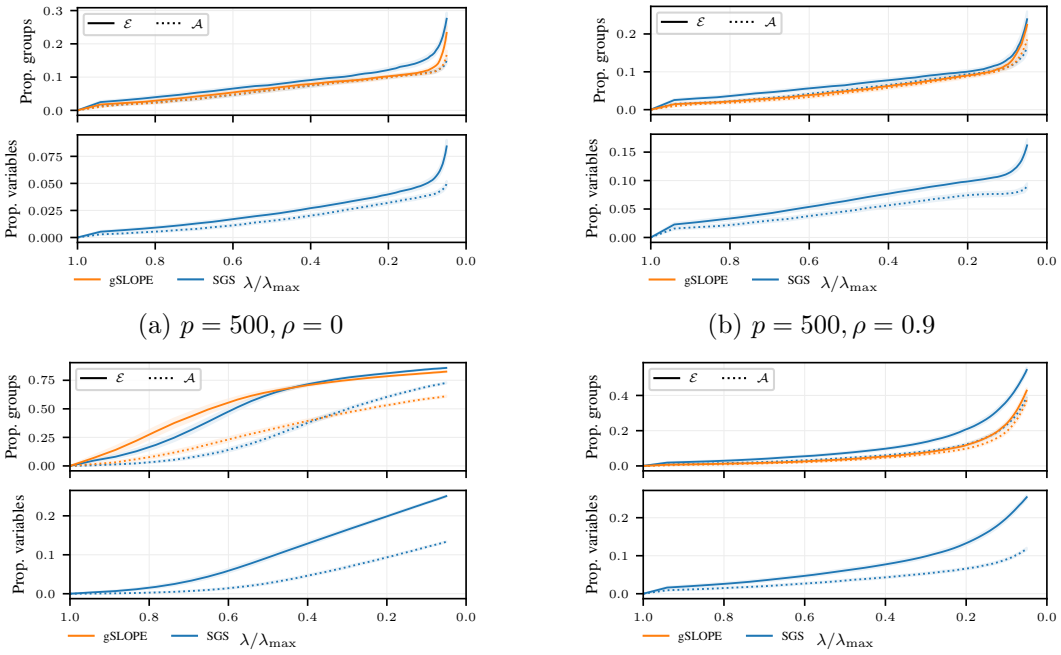
| **Method** | **Type** | card($\mathcal{A}_v$) | card($\mathcal{S}_v$) | card($\mathcal{E}_v$) | card($\mathcal{K}_v$) |
|---|---|---|---|---|---|
| SGS | Linear | 179 (3) | 313 (5) | 363 (6) | 51 (1) |
| SGS | Logistic | 230 (3) | 405 (5) | 472 (6) | 66 (1) |

Table C.1: Variable screening metrics for SGS using linear and logistic models for the simulation study presented in Section 4.6. The number of variables in $\mathcal{A}_v, \mathcal{S}_v, \mathcal{E}_v$, and $\mathcal{K}_v$ are shown, averaged across all 20 cases of the correlation ($\rho$) and dimensionality ($p$). Standard errors are shown in grey.

| Method | Type | Runtime (s) | | Cardinality | | | | Num. it. | | $\ell_2$ dist. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Screen | No screen | $\mathcal{A}_g$ | $\mathcal{S}_g$ | $\mathcal{E}_g$ | $\mathcal{K}_g$ | Screen | No screen | |
| gSLOPE | Linear | 1016 (21) | 1623 (27) | 55 (1) | 76 (1) | 76 (1) | 0.006 (0.004) | 333 (6) | 351 (6) | $2 \times 10^{-6}$ ($1\times10^{-6}$) |
| gSLOPE | Logistic | 814 (8) | 1409 (11) | 71 (1) | 97 (1) | 97 (1) | 0.014 (0.014) | 78 (1) | 83 (1) | $1 \times 10^{-8}$ ($1\times10^{-8}$) |
| SGS | Linear | 735 (15) | 1830 (34) | 61 (1) | 84 (1) | 91 (1) | - | 91 (3) | 708 (12) | $2 \times 10^{-9}$ ($3\times10^{-9}$) |
| SGS | Logistic | 407 (2) | 859 (6) | 84 (1) | 107 (1) | 118 (1) | - | 7 (0.2) | 51 (0.8) | $4 \times 10^{-9}$ ($3\times10^{-10}$) |

Table C.2: General and group screening metrics for SGS and gSLOPE using linear and logistic models for the synthetic study. General metrics: runtime (seconds), iteration count (with failed convergences in brackets), and $\ell_2$ distance between screening and no screening. Group metrics: numbers of groups in $\mathcal{A}_g$, $\mathcal{S}_g$, $\mathcal{E}_g$, and $\mathcal{K}_g$. Results are averaged over all 20 cases of the correlation ($\rho$) and dimensionality ($p$), with standard errors in grey.

### Additional Results for the Linear Model




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

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

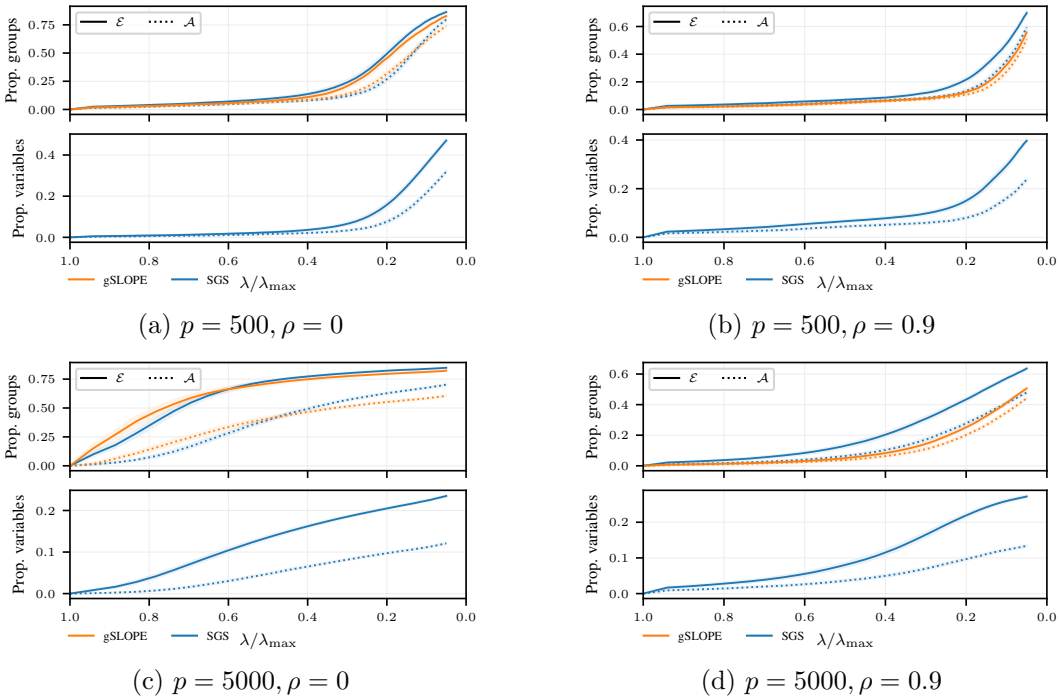
(c) $p = 5000, \rho = 0$

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

Figure C.8: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$ as a function of the regularisation path for the linear model with SGS and gSLOPE, shown for different values of the correlation ($\rho$) and dimensionality ($p$). The results are averaged over 100 repetitions, with 95% confidence intervals shown.

### Additional Results for the Logistic Model

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




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

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

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

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

Figure C.9: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$ as a function of the regularisation path for the logistic model with SGS and gSLOPE, shown for different values of the correlation ($\rho$) and dimensionality ($p$). The results are averaged over 100 repetitions, with 95% confidence intervals shown. This figure is the equivalent of Figure C.8 for the logistic model.

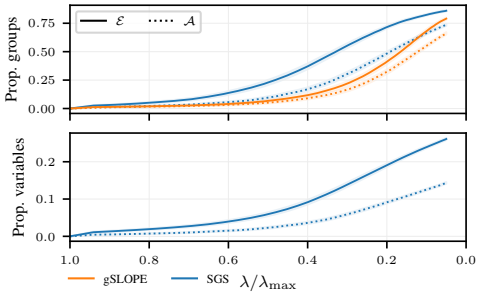



Figure C.10: The proportion of groups/variables in $\mathcal{E}, \mathcal{A}$ as a function of the regularisation path for the logistic model with $p = 2750, \rho = 0.6, m = 197$, shown for gSLOPE and SGS. The results are averaged over 100 repetitions, with 95% confidence intervals shown. This figure is the equivalent of Figure 4.2 for the logistic model.

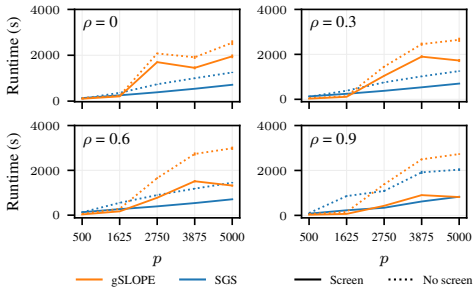



Figure C.11: Runtime (in seconds) for screening against no screening as a function of input dimensionality ($p$), broken down into different correlation cases, for the logistic model. The results are averaged over 100 repetitions, with standard errors shown. This figure is the equivalent of Figure 4.5 for the logistic model.

### C.5.4 Additional Results from the Real Data Experiments

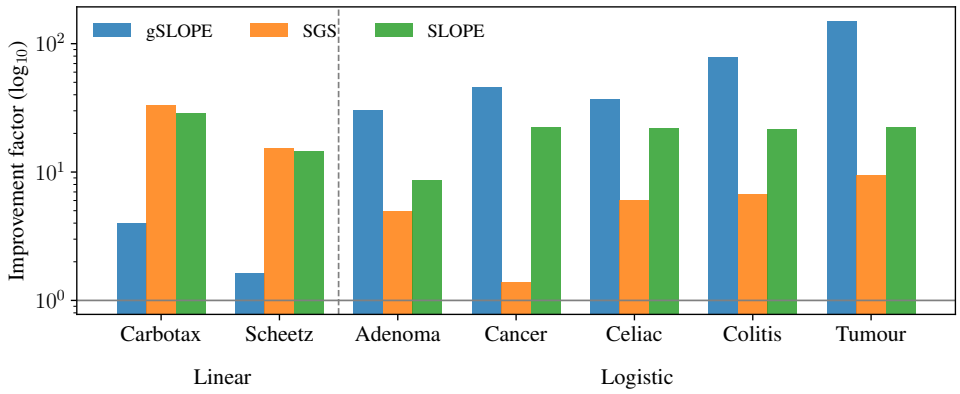



Figure C.12: The improvement factor of SLOPE, gSLOPE, and SGS applied to the real datasets, for fitting 100 path models, split into response type. The horizontal grey line represents no screening improvement.

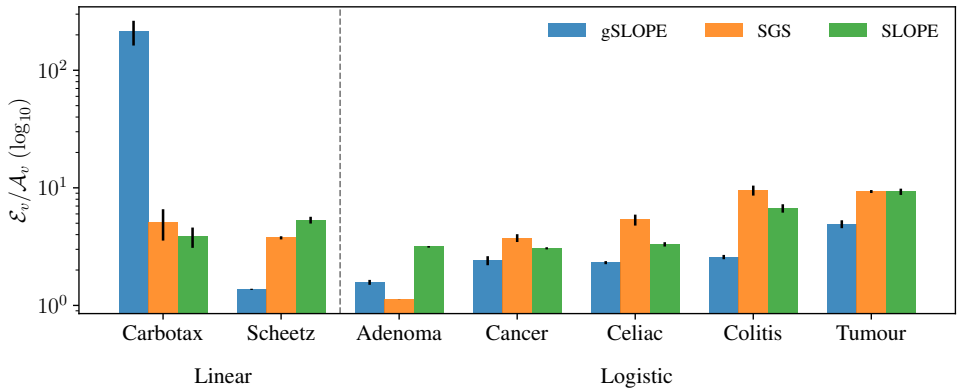



Figure C.13: The ratio of the fitting set ($\mathcal{E}_v$) to the active set ($\mathcal{A}_v$) of SLOPE, gSLOPE, and SGS applied to the real datasets, for fitting 100 path models, split into response type.

| **Method** | **Dataset** | card($\mathcal{A}_v$) | card($\mathcal{S}_v$) | card($\mathcal{E}_v$) | card($\mathcal{K}_v$) |
|---|---|---|---|---|---|
| SGS | Carbotax | 144 (6) | 361 (10) | 364 (10) | 2 (0.50) |
| SGS | Scheetz | 612 (74) | 1908 (174) | 1915 (175) | 7 (1.28) |
| SGS | Adenoma | 2174 (23) | 2451 (4) | 2463 (5) | 12 (0.22) |
| SGS | Cancer | 86 (3) | 273 (7) | 273 (7) | 1 (0.05) |
| SGS | Celiac | 175 (10) | 545 (22) | 546 (22) | 1 (0.07) |
| SGS | Colitis | 68 (4) | 409 (7) | 411 (7) | 2 (0.18) |
| SGS | Tumour | 128 (5) | 1189 (59) | 1200 (59) | 12 (0.54) |

Table C.3: Variable screening metrics for SGS applied to the real datasets. The number of variables in $\mathcal{A}_v, \mathcal{S}_v, \mathcal{E}_v$, and $\mathcal{K}_v$ are shown. The results are averaged across the nine pathway collections, with standard errors shown in grey.

| Method | Dataset | Runtime (s) | | Cardinality | | | | Num. it. | | $\ell_2$ dist. |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Screen | No screen | $\mathcal{A}_g$ | $\mathcal{S}_g$ | $\mathcal{E}_g$ | $\mathcal{K}_g$ | Screen | No screen | |
| gSLOPE | Carbotax | 865 | 3456 | 71 (3) | 124 (2) | 124 (2) | 0.05 (0.05) | 2367(1) (176) | 6433(10) (303) | $1 \times 10^{-9}$ ($2\times10^{-10}$) |
| gSLOPE | Scheetz | 4731 | 7735 | 153 (6) | 212 (6) | 212 (6) | 0 (0) | 6252(20) (313) | 7183(38) (332) | $8 \times 10^{-10}$ ($8\times10^{-10}$) |
| gSLOPE | Adenoma | 415 | 12610 | 71 (6) | 77 (4) | 77 (4) | 0 (0) | 3956(14) (378) | 7864(78) (394) | $9 \times 10^{-6}$ ($1\times10^{-6}$) |
| gSLOPE | Cancer | 109 | 5003 | 46 (3) | 81 (5) | 81 (5) | 0 (0) | 764(0) (74) | 6250(40) (404) | $4 \times 10^{-7}$ ($8\times10^{-8}$) |
| gSLOPE | Celiac | 205 | 7530 | 34 (75) | 62 (9) | 62 (9) | 0 (0) | 834(0) (177) | 4764(27) (428) | $4 \times 10^{-6}$ ($1\times10^{-5}$) |
| gSLOPE | Colitis | 148 | 11567 | 69 (5) | 114 (8) | 114 (8) | 0 (0) | 749(0) (62) | 9090(72) (206) | $4 \times 10^{-6}$ ($7\times10^{-7}$) |
| gSLOPE | Tumour | 222 | 33405 | 24 (1) | 83 (4) | 83 (4) | 0 (0) | 892(0) (67) | 9224(86) (218) | $4 \times 10^{-7}$ ($5\times10^{-8}$) |
| SGS | Carbotax | 63 | 2067 | 40 (1) | 138 (1) | 119 (1) | - | 168(0) (60) | 3572(3) (350) | $8 \times 10^{-10}$ ($3\times10^{-10}$) |
| SGS | Scheetz | 358 | 5476 | 105 (4) | 203 (5) | 186 (6) | - | 10(0) (4) | 5039(28) (360) | $7 \times 10^{-8}$ ($2\times10^{-8}$) |
| SGS | Adenoma | 1234 | 6143 | 601 (6) | 789 (1) | 735 (1) | - | 1172(0) (69) | 4224(20) (338) | $6 \times 10^{-7}$ ($5\times10^{-8}$) |
| SGS | Cancer | 151 | 211 | 70 (2) | 295 (3) | 182 (4) | - | 124(0) (15) | 147(0) (14) | $2 \times 10^{-8}$ ($2\times10^{-9}$) |
| SGS | Celiac | 339 | 2035 | 120 (7) | 460 (10) | 288 (10) | - | 55(0) (12) | 1259(0) (112) | $9 \times 10^{-5}$ ($2\times10^{-5}$) |
| SGS | Colitis | 271 | 1821 | 52 (3) | 368 (3) | 230 (3) | - | 54(0) (10) | 1411(0) (188) | $1 \times 10^{-6}$ ($3\times10^{-7}$) |
| SGS | Tumour | 781 | 7451 | 94 (3) | 733 (10) | 479 (14) | - | 33(0) (7) | 6006(23) (287) | $7 \times 10^{-7}$ ($2\times10^{-7}$) |

Table C.4: General and group screening metrics for SGS and gSLOPE on real datasets. General metrics: runtime (seconds), iteration count (with failed convergences in brackets), and $\ell_2$ distance between screening and no screening. Group metrics: numbers of groups in $\mathcal{A}_g$, $\mathcal{S}_g$, $\mathcal{E}_g$, and $\mathcal{K}_g$. Results are averaged over nine pathway collections, with standard errors in grey.

# APPENDIX D

# Additional Material for Chapter 5

## D.1 Sparse-group Lasso

### D.1.1 Variable Reduction

*Proof of Proposition 5.3.* The proof strategy is similar to that of Proposition 5.1. For $k \in [l-1]$, to prove the two sets are equivalent, we need to show that for any $i \in G_j$ such that $j \in \mathcal{A}_g(\lambda_{k+1})$, $i \in \mathcal{A}_v(\lambda_{k+1}) \iff i \in \mathcal{S}_v(\lambda_{k+1})$. We instead prove the contrapositive: $i \notin \mathcal{S}_v(\lambda_{k+1}) \iff i \notin \mathcal{A}_v(\lambda_{k+1})$. So,

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

□

*Proof of Proposition 5.4.* The proof strategy is similar to that of Proposition 5.2. For $k \in [l-1]$, to prove the screened set is a superset of the active set, we need to prove that for any $i \in G_j$ such that $j \in \mathcal{A}_g(\lambda_{k+1})$, $i \in \mathcal{A}_v(\lambda_{k+1}) \implies i \in \mathcal{S}_v(\lambda_{k+1})$. We instead prove the contrapositive: $i \notin \mathcal{S}_v(\lambda_{k+1}) \implies i \notin \mathcal{A}_v(\lambda_{k+1})$. First, we rewrite the Lipschitz assumption as (using the reverse triangle inequality)

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

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

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

Plugging this into Equation D.1 yields

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

□

### D.1.2 Reduction to (Adaptive) Lasso and (Adaptive) Group Lasso

Under $\alpha = 1$, SGL reduces to the lasso. In this case, group screening does not occur, and the variable screening rule reduces to the lasso strong rule [193]: discard a variable $i \in [p]$ if

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

Under $\alpha = 0$, SGL reduces to the group lasso. Under this scenario, the group screening reduces to the group lasso strong rule [193]: discard a group $j \in [m]$ if

$$\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_2 \leq \sqrt{p_j}(2\lambda_{k+1} - \lambda_k),$$

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

$$\begin{aligned}
\text{Adaptive lasso: } |\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_k))| \leq v_i(2\lambda_{k+1} - \lambda_k) &\implies \hat{\beta}_i(\lambda_{k+1}) = 0. \\
\text{Adaptive group lasso: } \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon'_{j,1}} \leq w_j\sqrt{p_j}(2\lambda_{k+1} - \lambda_k) &\implies \hat{\boldsymbol{\beta}}^{(j)}(\lambda_{k+1}) \equiv \mathbf{0},
\end{aligned}$$

where $\epsilon'_{j,1}$ denotes the $\epsilon$-norm under $\epsilon'_j = 1$.

## D.2 Adaptive Sparse-group Lasso

### D.2.1 Group screening

To derive the group screening rule for aSGL, we compare the formulations of SGL and aSGL in terms of the $\epsilon$-norm (Equations 5.1 and 5.20):

$$J_{\text{sgl}}(\boldsymbol{\beta}) = \sum_{j=1}^{m} \tau_j \|\beta^{(j)}\|_{\epsilon_j}^*, \quad J_{\text{asgl}}(\boldsymbol{\beta}) = \sum_{j=1}^{m} \tau_j' \|\beta^{(j)}\|_{\epsilon_j'}^*.$$

It is clear that the derivation for the group screening rule for aSGL is identical to that of SGL (Section 5.2.1), replacing $\tau_j$ with $\tau_j'$ and $\|\cdot\|_{\epsilon_j}$ with $\|\cdot\|_{\epsilon_j'}$. The group screening rule is formalised in Propositions D.1 and D.2.

**Proposition D.1** (Theoretical aSGL group screening)**.** *For aSGL applied with any $\lambda_{k+1}, k \in [l-1]$, the screened group set,*

$$\mathcal{S}_g(\lambda_{k+1}) = \{j \in [m] : \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))\|_{\epsilon_j'} > \tau_j' \lambda_{k+1}\},$$

*is such that $\mathcal{S}_g(\lambda_{k+1}) = \mathcal{A}_g(\lambda_{k+1})$.*

*Proof.* The proof is identical to that of Proposition 5.1 replacing $\tau_j$ with $\tau_j'$ and $\|\cdot\|_{\epsilon_j}$ with $\|\cdot\|_{\epsilon_j'}$. □

**Proposition D.2** (DFR-aSGL group screening)**.** *For aSGL applied with any $\lambda_{k+1}, k \in [l-1]$, assuming that*

$$\|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) - \nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j'} \leq \tau_j' |\lambda_{k+1} - \lambda_k|, \ \forall j \in [m],$$

*then the screened group set,*

$$\mathcal{S}_g(\lambda_{k+1}) = \{j \in [m] : \|\nabla_j f(\hat{\boldsymbol{\beta}}(\lambda_k))\|_{\epsilon_j'} > \tau_j'(2\lambda_{k+1} - \lambda_k)\},$$

*is such that $\mathcal{A}_g(\lambda_{k+1}) \subset \mathcal{S}_g(\lambda_{k+1})$.*

*Proof.* The proof is identical to that of Proposition 5.2 replacing $\tau_j$ with $\tau_j'$ and $\|\cdot\|_{\epsilon_j}$ with $\|\cdot\|_{\epsilon_j'}$. □

### D.2.2 Variable screening

The construction of the variable screening rule for aSGL is very similar to that of SGL (Section 5.2.2). The KKT stationary conditions for aSGL for an inactive variable in an active group are (in comparison to Equation 5.14 for SGL)

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

The derivation of the rule is identical, replacing $\alpha$ with $\alpha v_i$. The variable screening rule is formalised in Propositions D.3 and D.4.

**Proposition D.3** (Theoretical aSGL variable screening)**.** *For aSGL applied with any $\lambda_{k+1}, k \in [l-1]$, the screened variable set,*

$$\mathcal{S}_v(\lambda_{k+1}) = \{i \in G_j \text{ for } j \in \mathcal{A}_g(\lambda_{k+1}) : |\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1}))| > \lambda_{k+1}\alpha v_i\},$$

*is such that $\mathcal{S}_v(\lambda_{k+1}) = \mathcal{A}_v(\lambda_{k+1})$.*

*Proof.* The proof is identical to that of Proposition 5.3 replacing $\alpha$ with $\alpha v_i$ (see Appendix D.1.1). □

**Proposition D.4** (DFR-aSGL variable screening)**.** *For aSGL applied with any $\lambda_{k+1}, k \in [l-1]$, assuming that*

$$|\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_{k+1})) - \nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_k))| \leq \alpha v_i(\lambda_k - \lambda_{k+1}),\ \forall i \in \mathcal{G}_{\mathcal{A}_g(\lambda_{k+1})},$$

*then the variable screened set,*

$$\mathcal{S}_v(\lambda_{k+1}) = \{i \in G_j \text{ for } j \in \mathcal{A}_g(\lambda_{k+1}) : |\nabla_i f(\hat{\boldsymbol{\beta}}(\lambda_k))| > \alpha v_i(2\lambda_{k+1} - \lambda_k)\},$$

*is such that $\mathcal{A}_v(\lambda_{k+1}) \subset \mathcal{S}_v(\lambda_{k+1})$.*

*Proof.* The proof is identical to that of Proposition 5.4 replacing $\alpha$ with $\alpha v_i$ (see Appendix D.1.1). □

## D.3 Data Experiments

This section complements Sections 5.4 and 5.5 by providing additional information on the simulation setup and further results. Additional tables and figures are pro-

vided that further showcase the effectiveness of DFR, including under a logistic model (Appendix D.3.2).

An intercept term was included only for linear models, as it centres the response variable–a step not applicable to binary outcomes. Since the focus here is on computational cost rather than predictive performance, adding an intercept to logistic models offers no practical advantage. See Appendix C.5.1 for a description of the metrics used in this section.

### D.3.1 Additional Results for the Linear Model

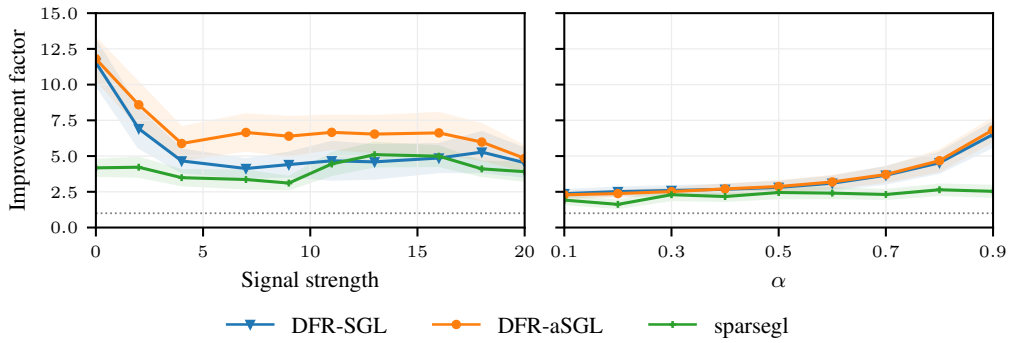



Figure D.1: The improvement factor for the strong rules applied to synthetic data, under the linear model, as a function of the signal strength (left) and $\alpha$ (right) under uneven groups, with 95% confidence intervals.

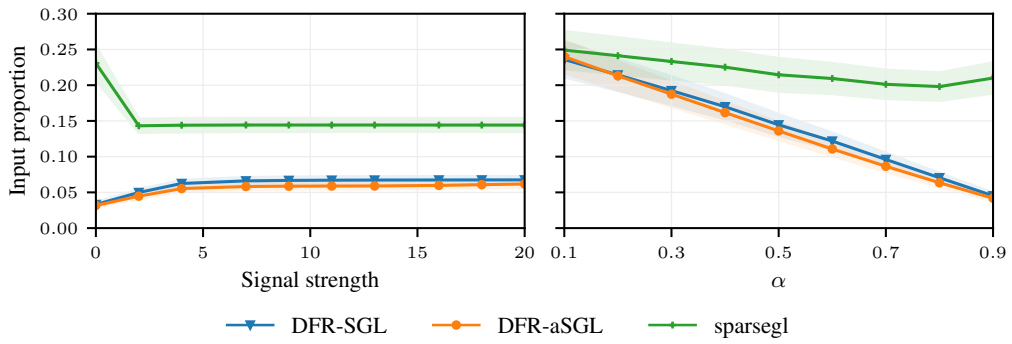



Figure D.2: The input proportion for the strong rules applied to synthetic data, under the linear model, as a function of the signal strength (left) and $\alpha$ (right) under uneven groups, with 95% confidence intervals.

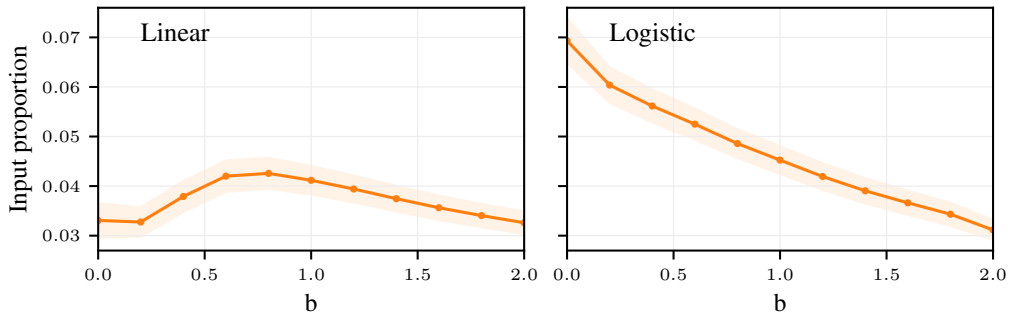



Figure D.3: The input proportion of DFR-aSGL as a function of the weights $b_1 = b_2$ under uneven groups, shown for the linear (left) and logistic (right) models, with 95% confidence intervals.

### Tables for GAP Safe Simulations

| Method | Cardinality | | | | Input proportion | |
|---|---|---|---|---|---|---|
| | $\mathcal{A}_g$ | $\mathcal{S}_g$ | $\mathcal{E}_g$ | $\mathcal{K}_g$ | $\mathcal{E}_g \,/\, \mathcal{A}_g$ | $\mathcal{E}_g \,/\, m$ |
| DFR-aSGL | 7.86 (0.02) | 10.00 (0.03) | 10.00 (0.03) | – | 1.1751 ($7\times10^{-4}$) | 0.2001 ($5\times10^{-4}$) |
| DFR-SGL | 8.16 (0.02) | 10.47 (0.02) | 10.47 (0.02) | – | 1.1937 ($8\times10^{-4}$) | 0.2094 ($5\times10^{-4}$) |
| DFR-SGL BCD | 8.45 (0.02) | 10.07 (0.02) | 10.07 (0.02) | – | 1.1893 ($8\times10^{-4}$) | 0.2015 ($4\times10^{-4}$) |
| sparsegl | 8.16 (0.02) | 11.72 (0.04) | 11.72 (0.04) | $3\times10^{-5}$ ($2\times10^{-5}$) | 1.5837 (0.0111) | 0.2344 ($7\times10^{-4}$) |
| GAP sequential | 8.55 (0.02) | 10.20 (0.02) | 10.20 (0.02) | – | 1.1934 (0.0018) | 0.2040 ($4\times10^{-4}$) |
| GAP dynamic | 8.55 (0.02) | 10.20 (0.02) | 10.20 (0.02) | – | 1.1917 (0.0017) | 0.2039 ($4\times10^{-4}$) |

Table D.1: Group screening metrics corresponding to the GAP safe simulations (Figures 5.2 and 5.3), averaged over all cases and path points, with standard errors shown in grey.

| **Method** | **Cardinality** | | | | **Input proportion** | |
|---|---|---|---|---|---|---|
| | $\mathcal{A}_v$ | $\mathcal{S}_v$ | $\mathcal{E}_v$ | $\mathcal{K}_v$ | $\mathcal{E}_v$ / $\mathcal{A}_v$ | $\mathcal{E}_v$ / $p$ |
| DFR-aSGL | 61.77 (0.22) | 36.70 (0.34) | 95.99 (0.42) | 0.0092 ($3\times10^{-4}$) | 1.3464 (0.0011) | 0.0960 ($4\times10^{-4}$) |
| DFR-SGL | 65.37 (0.22) | 38.30 (0.34) | 101.07 (0.43) | 0 (0) | 1.3576 (0.0011) | 0.1011 ($4\times10^{-4}$) |
| DFR-SGL BCD | 72.69 (0.25) | 23.35 (0.09) | 93.11 (0.32) | 0.0250 (0.0054) | 1.3281 (0.0011) | 0.0931 ($3\times10^{-4}$) |
| sparsegl | 65.41 (0.22) | 234.39 (0.70) | 234.39 (0.70) | – | 11.5604 (0.2001) | 0.2344 ($7\times10^{-4}$) |
| GAP sequential | 73.01 (0.25) | 93.85 (0.33) | 93.85 (0.33) | – | 1.3292 (0.0022) | 0.0938 ($3\times10^{-4}$) |
| GAP dynamic | 73.01 (0.25) | 93.38 (0.33) | 93.38 (0.33) | – | 1.2927 (0.0020) | 0.0934 ($3\times10^{-4}$) |

Table D.2: Variable screening metrics corresponding to the GAP safe simulations (Figures 5.2 and 5.3), averaged over all cases and path points, with standard errors shown in grey.

| **Method** | **Timings** | | | **Iterations** | | $\ell_2$ **dist.** | **Failed convergence** | |
|---|---|---|---|---|---|---|---|---|
| | No screen (s) | Screen (s) | IF | No screen | Screen | | No screen | Screen |
| DFR-aSGL | 659.89 (6.50) | 154.92 (2.39) | 7.01 (0.15) | 271.60 (2.36) | 174.55 (2.49) | $4\times10^{-4}$ ($3\times10^{-6}$) | 0 (0) | 0 (0) |
| DFR-SGL | 685.49 (6.27) | 157.10 (2.24) | 7.98 (0.20) | 286.13 (2.04) | 185.47 (2.43) | $4\times10^{-4}$ ($3\times10^{-6}$) | 0 (0) | 0 (0) |
| DFR-SGL BCD | 131.46 (1.12) | 32.02 (0.55) | 5.55 (0.08) | 252.26 (1.94) | 139.39 (1.15) | $2\times10^{-7}$ ($1\times10^{-8}$) | 0 (0) | 0 (0) |
| sparsegl | 685.49 (6.27) | 275.37 (4.43) | 3.44 (0.08) | 286.13 (2.04) | 278.18 (2.28) | $4\times10^{-4}$ ($3\times10^{-6}$) | 0 (0) | 0 (0) |
| GAP sequential | 0.11 ($3\times10^{-3}$) | 0.11 ($3\times10^{-3}$) | 0.98 (0.01) | – | – | – | – | – |
| GAP dynamic | 0.11 ($3\times10^{-3}$) | 0.11 ($3\times10^{-3}$) | 1.00 (0.01) | – | – | – | – | – |

Table D.3: Model fitting metrics corresponding to the GAP safe simulations (Figures 5.2 and 5.3), averaged over all cases and path points, with standard errors shown in grey. The timing results are the average time taken to evaluate the full path on a dataset.

**Tables for Other Simulations**

| Method | Cardinality | | | | Input proportion | |
|---|---|---|---|---|---|---|
| | $\mathcal{A}_g$ | $\mathcal{S}_g$ | $\mathcal{E}_g$ | $\mathcal{K}_g$ | $\mathcal{E}_g$ / $\mathcal{A}_g$ | $\mathcal{E}_g$ / $m$ |
| DFR-aSGL | 8.46 (0.03) | 10.55 (0.04) | 10.54 (0.04) | – | 1.1414 ($5\times10^{-4}$) | 0.1968 ($4\times10^{-4}$) |
| DFR-SGL | 8.51 (0.03) | 11.16 (0.04) | 11.16 (0.04) | – | 1.1792 ($6\times10^{-4}$) | 0.2084 ($4\times10^{-4}$) |
| sparsegl | 8.51 (0.03) | 10.25 (0.04) | 10.25 (0.04) | $8\times10^{-5}$ ($2\times10^{-5}$) | 1.2191 (0.0026) | 0.2083 ($4\times10^{-4}$) |

Table D.4: Group screening metrics corresponding to the other linear model simulations, averaged over all cases and path points, with standard errors shown in grey.

| Method | Cardinality | | | | Input proportion | |
|---|---|---|---|---|---|---|
| | $\mathcal{A}_v$ | $\mathcal{S}_v$ | $\mathcal{E}_v$ | $\mathcal{K}_v$ | $\mathcal{E}_v$ / $\mathcal{A}_v$ | $\mathcal{E}_v$ / $p$ |
| DFR-aSGL | 50.95 (0.13) | 29.90 (0.15) | 78.92 (0.22) | 0.0297 ($4\times10^{-4}$) | 1.4805 (0.0011) | 0.0634 ($2\times10^{-4}$) |
| DFR-SGL | 54.40 (0.13) | 32.78 (0.15) | 85.16 (0.22) | $4\times10^{-6}$ ($4\times10^{-6}$) | 1.5057 (0.0014) | 0.0676 ($2\times10^{-4}$) |
| sparsegl | 54.42 (0.13) | 331.75 (0.86) | 331.75 (0.86) | – | 11.4845 (0.0882) | 0.2205 ($5\times10^{-4}$) |

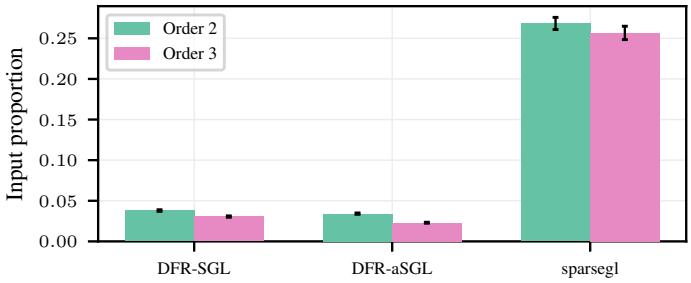
Table D.5: Variable screening metrics corresponding to the other linear model simulations, averaged over all cases and path points, with standard errors shown in grey.

| Method | Timings | | | Iterations | | $\ell_2$ dist. | Failed convergence | |
|---|---|---|---|---|---|---|---|---|
| | No screen (s) | Screen (s) | IF | No screen | Screen | | No screen | Screen |
| DFR-aSGL | 723.99 (34.04) | 77.00 (2.23) | 11.19 (0.41) | 415.87 (7.41) | 227.93 (3.51) | $7\times10^{-5}$ ($5\times10^{-7}$) | 0 (0) | 0 (0) |
| DFR-SGL | 251.59 (3.93) | 51.73 (0.89) | 9.15 (0.15) | 413.10 (7.12) | 240.78 (3.78) | $6\times10^{-5}$ ($5\times10^{-7}$) | 0 (0) | 0 (0) |
| sparsegl | 251.59 (3.93) | 138.18 (5.62) | 3.59 (0.05) | 413.10 (7.12) | 353.66 (5.79) | $6\times10^{-5}$ ($5\times10^{-7}$) | 0 (0) | 0 (0) |

Table D.6: Model fitting metrics corresponding to the other linear model simulations, averaged over all cases and path points, with standard errors shown in grey. The timing results are the average time taken to evaluate the full path on a dataset.

### D.3.2 Results for the Logistic Model

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




(a) Input proportion.

| Method | Interaction | |
|---|---|---|
| | 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) |

(b) Improvement factor.

Figure D.4: The strong rules applied to synthetic interaction data, under the logistic model, with standard errors shown. The parameters of the data were set as $p = 400, n = 80$, and $m = 52$ groups of sizes in $[3, 15]$. The interaction input dimensionality was $p_{O_2} = 2111$ and $p_{O_3} = 7338$, with no interaction hierarchy imposed. The sparsity proportion of interaction variables was set to 0.7 (with the same signal as the marginal effects).

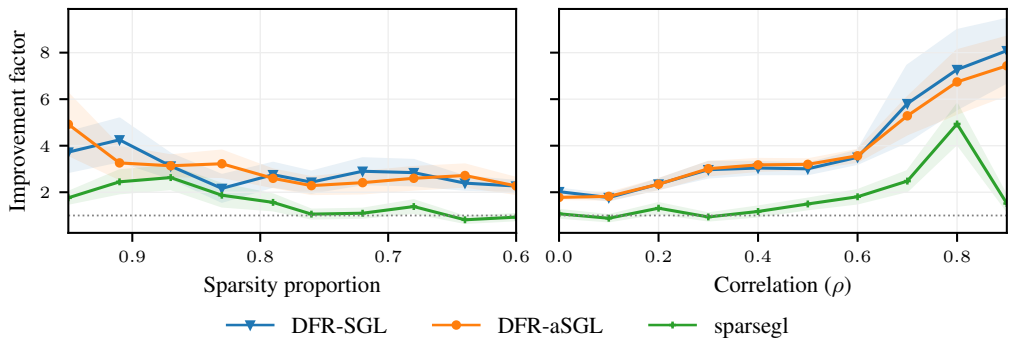



Figure D.5: The improvement factor for the strong rules applied to synthetic data, under the logistic model, as a function of the sparsity proportion (left) and data correlation (right) under uneven groups, with 95% confidence intervals.

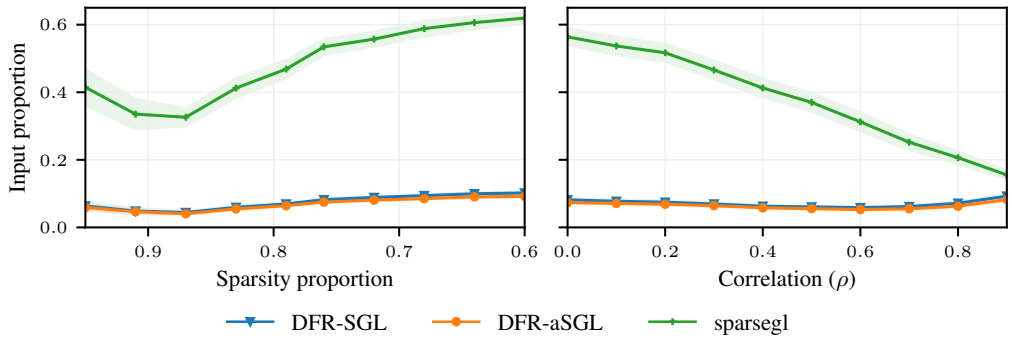



Figure D.6: The input proportion for the strong rules applied to synthetic data, under the logistic model, as a function of the sparsity proportion (left) and data correlation (right) under uneven groups, with 95% confidence intervals.

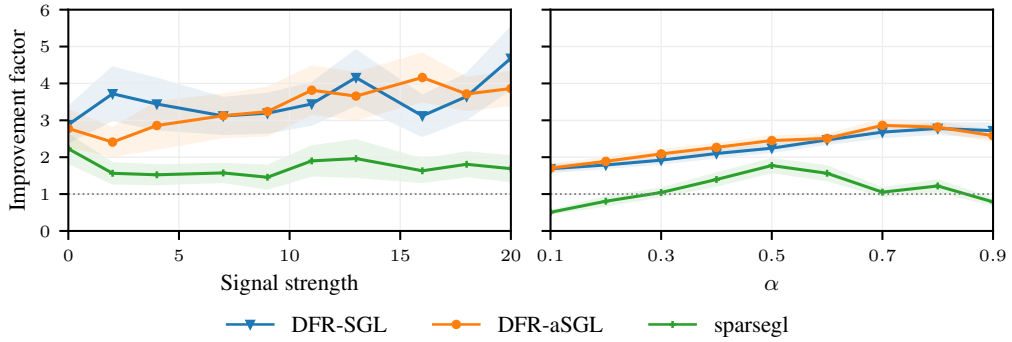



Figure D.7: The improvement factor for the strong rules applied to synthetic data, under the logistic model, as a function of the signal strength (left) and $\alpha$ (right) under uneven groups, with 95% confidence intervals.

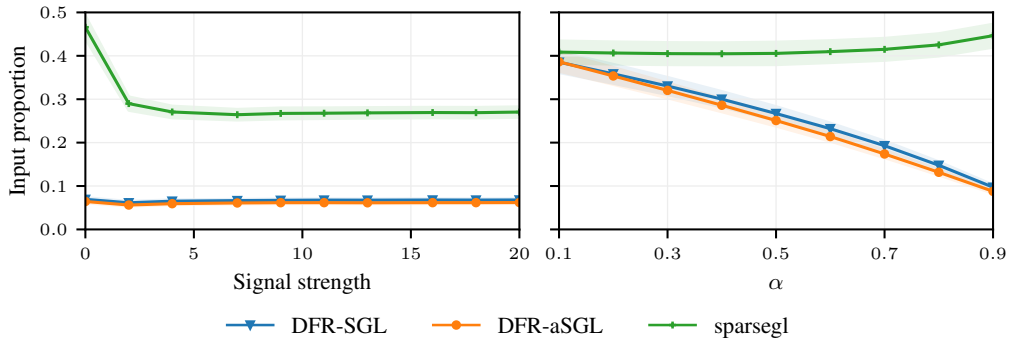



Figure D.8: The input proportion for the strong rules applied to synthetic data, under the logistic model, as a function of the signal strength (left) and $\alpha$ (right) under uneven groups, with 95% confidence intervals.

**Tables for the Logistic Models**

| Method | Cardinality | | | | Input proportion | |
|---|---|---|---|---|---|---|
| | $\mathcal{A}_g$ | $\mathcal{S}_g$ | $\mathcal{E}_g$ | $\mathcal{K}_g$ | $\mathcal{E}_g$ / $\mathcal{A}_g$ | $\mathcal{E}_g$ / $m$ |
| DFR-aSGL | 8.08 (0.01) | 8.55 (0.02) | 8.55 (0.02) | – | 1.0292 ($6\times10^{-4}$) | 0.3528 ($6\times10^{-4}$) |
| DFR-SGL | 8.56 (0.02) | 9.30 (0.02) | 9.30 (0.02) | – | 1.0554 ($7\times10^{-4}$) | 0.3823 ($7\times10^{-4}$) |
| sparsegl | 8.56 (0.02) | 8.87 (0.02) | 8.87 (0.02) | $5\times10^{-5}$ ($2\times10^{-5}$) | 1.0808 (0.0029) | 0.3697 ($7\times10^{-4}$) |

Table D.7: Group screening metrics corresponding to the logistic model simulations, averaged over all cases and path points, with standard errors shown in grey.

| Method | Cardinality | | | | Input proportion | |
|---|---|---|---|---|---|---|
| | $\mathcal{A}_v$ | $\mathcal{S}_v$ | $\mathcal{E}_v$ | $\mathcal{K}_v$ | $\mathcal{E}_v$ / $\mathcal{A}_v$ | $\mathcal{E}_v$ / $p$ |
| DFR-aSGL | 79.76 (0.24) | 35.27 (0.10) | 111.79 (0.32) | 0.0155 ($3\times10^{-4}$) | 1.5263 (0.0012) | 0.1075 ($3\times10^{-4}$) |
| DFR-SGL | 84.23 (0.25) | 40.13 (0.12) | 120.84 (0.34) | $9\times10^{-6}$ ($7\times10^{-6}$) | 1.5503 (0.0013) | 0.1154 ($3\times10^{-4}$) |
| sparsegl | 84.26 (0.25) | 445.21 (1.05) | 445.21 (1.05) | – | 11.6349 (0.1015) | 0.4004 ($7\times10^{-4}$) |

Table D.8: Variable screening metrics corresponding to the logistic model simulations, averaged over all cases and path points, with standard errors shown in grey.

| Method | Timings | | | Iterations | | $\ell_2$ dist. | Failed convergence | |
|---|---|---|---|---|---|---|---|---|
| | No screen (s) | Screen (s) | IF | No screen | Screen | | No screen | Screen |
| DFR-aSGL | 132.10 (7.21) | 34.75 (0.66) | 3.37 (0.05) | 125.73 (2.77) | 78.95 (1.10) | $9\times10^{-10}$ ($2\times10^{-11}$) | 0 (0) | 0 (0) |
| DFR-SGL | 103.16 (3.37) | 34.27 (0.56) | 3.43 (0.05) | 141.42 (2.92) | 87.88 (1.30) | $2\times10^{-10}$ ($8\times10^{-12}$) | 0 (0) | 0 (0) |
| sparsegl | 103.16 (3.37) | 109.93 (4.69) | 1.51 (0.03) | 141.42 (2.92) | 121.29 (2.11) | $2\times10^{-10}$ ($3\times10^{-12}$) | 0 (0) | 0 (0) |

Table D.9: Model fitting metrics corresponding to the logistic model simulations, averaged over all cases and path points, with standard errors shown in grey. The timing results are the average time taken to evaluate the full path on a dataset.

### D.3.3 Additional Results for the Real Data

| Method | Dataset | Cardinality | | | | Input proportion | |
|---|---|---|---|---|---|---|---|
| | | $\mathcal{A}_g$ | $\mathcal{C}_g$ | $\mathcal{E}_g$ | $\mathcal{K}_g$ | $\mathcal{E}_g$ / $\mathcal{A}_g$ | $\mathcal{E}_g$ / $m$ |
| DFR-aSGL | Adenoma | 1.41 (±0.08) | 1.38 (±0.08) | 1.42 (±0.08) | – | 1.0053 (±0.0053) | 0.0046 ($\pm 3\times10^{-4}$) |
| 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 | BRCA1 | 3.84 (±0.26) | 4.17 (±0.33) | 4.16 (±0.33) | – | 1.0439 (±0.0121) | 0.0171 (±0.0013) |
| DFR-SGL | BRCA1 | 5.60 (±0.48) | 6.90 (±0.60) | 6.90 (±0.60) | – | 1.2023 (±0.0218) | 0.0284 (±0.0025) |
| sparsegl | BRCA1 | 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) | – | 1.0000 (±0.0000) | 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) | 1.0000 (±0.0000) | 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 ($\pm 1\times10^{-3}$) |
| 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 ($\pm 9\times10^{-4}$) |

Table D.10: Group screening metrics corresponding to the real data studies, averaged over all path points, with standard errors shown in grey.

| **Method** | **Dataset** | **Cardinality** | | | | **Input proportion** | |
|---|---|---|---|---|---|---|---|
| | | $\mathcal{A}_v$ | $\mathcal{C}_v$ | $\mathcal{E}_v$ | $\mathcal{K}_v$ | $\mathcal{E}_v \,/\, \mathcal{A}_v$ | $\mathcal{E}_v \,/\, p$ |
| DFR-aSGL | Adenoma | 3.38 (±0.21) | 0.81 (±0.11) | 4.15 (±0.30) | 0.0404 (±0.0199) | 1.1828 (±0.0259) | $2 \times 10^{-4}$ ($\pm 2\times10^{-5}$) |
| DFR-SGL | Adenoma | 14.38 (±1.11) | 61.47 (±2.90) | 75.51 (±3.85) | 0 (±0) | 8.9655 (±0.7776) | 0.0041 ($\pm 2\times10^{-4}$) |
| 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 (±0.0004) |
| DFR-SGL | Celiac | 37.13 (±3.92) | 26.11 (±3.05) | 61.94 (±6.77) | 0 (±0) | 1.6187 (±0.0455) | 0.0042 (±0.0005) |
| sparsegl | Celiac | 37.26 (±3.94) | 1019.13 (±106.78) | 1019.13 (±106.78) | – | 27.6753 (±1.1176) | 0.0695 (±0.0073) |
| DFR-aSGL | BRCA1 | 135.75 (±4.69) | 31.93 (±2.77) | 165.83 (±6.19) | 0.0505 (±0.0221) | 1.1998 (±0.0116) | 0.0096 (±0.0004) |
| DFR-SGL | BRCA1 | 241.69 (±7.34) | 63.42 (±4.26) | 301.8 (±8.81) | 0 (±0) | 3.9386 (±2.6945) | 0.0174 (±0.0005) |
| sparsegl | BRCA1 | 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) | $6 \times 10^{-4}$ ($\pm 5\times10^{-5}$) |
| DFR-SGL | Tumour | 10.70 (±0.72) | 9.87 (±0.50) | 20.34 (±1.09) | 0 (±0) | 2.3432 (±0.1189) | 0.0011 ($\pm 6\times10^{-5}$) |
| sparsegl | Tumour | 10.70 (±0.72) | 246.8 (±15.39) | 246.8 (±15.39) | – | 25.9945 (±0.9193) | 0.0133 ($\pm 8\times10^{-4}$) |

Table D.11: Variable screening metrics corresponding to the real data studies, averaged over all path points, with standard errors shown in grey.

| Method | Dataset | Timings | | | Iterations | | $\ell_2$ dist. | Failed convergence | |
|---|---|---|---|---|---|---|---|---|---|
| | | No screen (s) | Screen (s) | I.F. | No screen | Screen | | No screen | Screen |
| DFR-aSGL | Adenoma | 8476.04 | 7.91 | 1072.10 | 8903.76 (±270.64) | 119.18 (±14.69) | $7 \times 10^{-5}$ ($\pm 6\times10^{-6}$) | 0.8384 (±0.0372) | 0 (±0) |
| DFR-SGL | Adenoma | 9017.70 | 149.10 | 60.48 | 9272.90 (±205.31) | 4374.64 (±286.62) | $3 \times 10^{-5}$ ($\pm 2\times10^{-6}$) | 0.8687 (±0.0341) | 0 (±0) |
| sparsegl | Adenoma | 9017.70 | 198.36 | 45.46 | 9272.90 (±205.31) | 5140.16 (±390.91) | $3 \times 10^{-5}$ ($\pm 2\times10^{-6}$) | 0.8687 (±0.0341) | 0.0404 (±0.0199) |
| DFR-aSGL | Celiac | 1188.14 | 15.86 | 74.89 | 1042.07 (±90.79) | 120.96 (±14.31) | $1 \times 10^{-6}$ ($\pm 2\times10^{-7}$) | 0 (±0) | 0 (±0) |
| DFR-SGL | Celiac | 1391.78 | 10.31 | 134.95 | 1195.34 (±98.13) | 75.37 (±9.60) | $2 \times 10^{-7}$ ($\pm 1\times10^{-8}$) | 0 (±0) | 0 (±0) |
| sparsegl | Celiac | 1391.78 | 16.49 | 84.40 | 1195.34 (±98.13) | 93.29 (±8.86) | $8 \times 10^{-8}$ ($\pm 3\times10^{-9}$) | 0 (±0) | 0 (±0) |
| DFR-aSGL | BRCA1 | 21889.33 | 119.16 | 183.69 | 1653.45 (±35.90) | 243.27 (±11.70) | $2 \times 10^{-9}$ ($\pm 7\times10^{-10}$) | 0 (±0) | 0 (±0) |
| DFR-SGL | BRCA1 | 22227.01 | 103.78 | 214.17 | 1674.07 (±19.77) | 334.16 (±13.44) | $2 \times 10^{-12}$ ($\pm 3\times10^{-13}$) | 0 (±0) | 0 (±0) |
| sparsegl | BRCA1 | 22227.01 | 4132.04 | 5.38 | 1674.07 (±19.77) | 1580.73 (±44.23) | $4 \times 10^{-14}$ ($\pm 1\times10^{-14}$) | 0 (±0) | 0 (±0) |
| DFR-aSGL | Scheetz | 90569.05 | 132.20 | 685.08 | 6040.81 (±457.00) | 1030.48 (±90.27) | $1 \times 10^{-6}$ ($\pm 3\times10^{-7}$) | 0.5657 (±0.0501) | 0 (±0) |
| DFR-SGL | Scheetz | 68084.77 | 2246.13 | 30.31 | 1891.21 (±386.40) | 642.38 (±136.12) | $3 \times 10^{-9}$ ($\pm 8\times10^{-10}$) | 0.1818 (±0.0390) | 0 (±0) |
| sparsegl | Scheetz | 68084.77 | 6666.65 | 10.21 | 1891.21 (±386.40) | 1890.45 (±386.44) | $5 \times 10^{-22}$ ($\pm 2\times10^{-22}$) | 0.1818 (±0.0390) | 0.1818 (±0.0390) |
| DFR-aSGL | Trust-experts | 5.96 | 3.10 | 1.92 | 76.74 (±1.60) | 85.11 (±3.06) | $1 \times 10^{-11}$ ($\pm 3\times10^{-12}$) | 0 (±0) | 0 (±0) |
| DFR-SGL | Trust-experts | 7.29 | 3.20 | 2.28 | 98.36 (±4.17) | 92.55 (±5.93) | $4 \times 10^{-11}$ ($\pm 9\times10^{-12}$) | 0 (±0) | 0 (±0) |
| sparsegl | Trust-experts | 7.29 | 4.52 | 1.61 | 98.36 (±4.17) | 104.24 (±5.16) | $2 \times 10^{-6}$ ($\pm 2\times10^{-6}$) | 0 (±0) | 0 (±0) |
| DFR-aSGL | Tumour | 7027.69 | 86.25 | 81.48 | 8783.1 (±265.36) | 82.75 (±5.61) | $3 \times 10^{-8}$ ($\pm 9\times10^{-9}$) | 0.6768 (±0.0472) | 0 (±0) |
| DFR-SGL | Tumour | 7466.83 | 89.43 | 83.49 | 9272.02 (±224.00) | 186.27 (±7.64) | $3 \times 10^{-9}$ ($\pm 2\times10^{-10}$) | 0.8586 (±0.0352) | 0 (±0) |
| sparsegl | Tumour | 7466.83 | 90.03 | 82.93 | 9272.02 (±224.00) | 197.40 (±9.38) | $2 \times 10^{-9}$ ($\pm 2\times10^{-10}$) | 0.8586 (±0.0352) | 0 (±0) |

Table D.12: Model fitting metrics corresponding to the real data studies, averaged over all path points, with standard errors shown in grey.

# APPENDIX E

# Additional Material for Chapter 7

## E.1 Group-based Bayesian SLOPE Models

### E.1.1 Bayesian gSLOPE (BGSLOPE)

*Proof of Proposition 7.1.* Assume we have a variable $\mathbf{z} = (z_1, \dots, z_p)$ with the gSLOPE prior

$$\pi(\boldsymbol{\beta} \mid \sigma^2, \mathbf{w}) \propto \prod_{j=1}^{m} \exp\left\{ -\frac{1}{\sigma} w_{r_g(\boldsymbol{\beta}, j)} \sqrt{p_j} \|\boldsymbol{\beta}^{(j)}\|_2 \right\}. \tag{E.1}$$

Now, define $\boldsymbol{\beta} = \tilde{\mathbf{A}}^{-1}\mathbf{z}$ such that $z_i = \beta_i \tilde{a}_j$ for $i \in G_j$. By the transformation of variables, we have that the prior for $\boldsymbol{\beta}$ is given by

$$\pi\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, c, \sigma^2; \mathbf{w}\right) \propto \left|\det(d\mathbf{z}/d\boldsymbol{\beta})\right| \pi\left(\mathbf{z} \mid \boldsymbol{\gamma}, c, \sigma^2; \mathbf{w}\right).$$

Now, we have that $(d\mathbf{z}/d\boldsymbol{\beta})_i = \tilde{a}_j$, and as $\tilde{\mathbf{A}}$ is diagonal,

$$\det(d\mathbf{z}/d\boldsymbol{\beta}) = \det(\tilde{\mathbf{A}}) = \prod_{j=1}^{m} \tilde{a}_j^{p_j} = c^{\sum_{j=1}^{m} p_j \mathbf{1}(\gamma_j = 1)}.$$

The final equality follows by the definition of $\tilde{\mathbf{A}}$. Therefore,

$$\begin{aligned}\pi\left(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, c, \sigma^2; \mathbf{w}\right) &\propto \left|\det\left(\frac{d\mathbf{z}}{d\boldsymbol{\beta}}\right)\right| \pi\left(\mathbf{z} \mid \boldsymbol{\gamma}, c, \sigma^2; \mathbf{w}\right) \\ &= c^{\sum_{j=1}^m p_j \mathbf{1}(\gamma_j=1)} \prod_{j=1}^m \exp\left\{-\frac{1}{\sigma} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)} \sqrt{p_j} \|\mathbf{z}^{(j)}\|_2\right\} \\ &= c^{\sum_{j=1}^m p_j \mathbf{1}(\gamma_j=1)} \prod_{j=1}^m \exp\left\{-\frac{1}{\sigma} \tilde{a}_j w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)} \sqrt{p_j} \|\boldsymbol{\beta}^{(j)}\|_2\right\},\end{aligned}$$

which corresponds to the BGSLOPE prior. The final step follows as the entries in $\tilde{\mathbf{A}}$ are identical within each group, so that it can be treated as a scaling term. □

**Theorem E.1.** *The gSLOPE normalising constant is given by*

$$\begin{aligned}C^{-1} = m! \prod_{g=1}^m \frac{2\pi^{p_g/2}}{\Gamma(p_g/2)} \\ \times \sum_{\mathbf{k} \in \mathbb{K}} \prod_{j=1}^{m-1} \frac{(p_j + k_{j-1} - 1)! (\sigma^{-1}\sqrt{p_j} \sum_{l=1}^j w_l)^{k_j} (p_m + k_{m-1} - 1)!}{k_j! (\sigma^{-1}\sqrt{p_j} \sum_{l=1}^j w_l)^{p_j + k_{j-1}} (\sigma^{-1}\sqrt{p_m} \sum_{l=1}^m w_l)^{p_m + k_{m-1}}},\end{aligned}$$

*where* $\mathbb{K} = \{\mathbf{k} \in \mathbb{Z}^{m-1} : k_1 \in [0, p_1 - 1],\ k_2 \in [0, p_2 + k_1 - 1],\ \ldots,\ k_{m-1} \in [0, p_{m-1} + k_{m-2} - 1]\}$, *with* $k_0 = 0$.

*Proof.* The proof strategy is to transform the problem into spherical coordinates, expressing the integrals via Gamma functions, and thus exploiting their exponential representation.

The normalising constant is given by the integral [173]

$$I = \int e^{-\sigma^{-1} \sum_{g=1}^m p_g w_g \|\boldsymbol{\beta}^{(g)}\|_2} d\boldsymbol{\beta}.$$

We restrict ourselves to the setting $\sqrt{p_1}\|\boldsymbol{\beta}^{(1)}\|_2 \geq \cdots \geq \sqrt{p_m}\|\boldsymbol{\beta}^{(m)}\|_2$. There are $m!$ possible permutations of this ordering, so that

$$I = m! \int_{\sqrt{p_1}\|\boldsymbol{\beta}^{(1)}\|_2 \geq \cdots \geq \sqrt{p_m}\|\boldsymbol{\beta}^{(m)}\|_2} e^{-\sigma^{-1} \sum_{g=1}^m p_g w_g \|\boldsymbol{\beta}^{(g)}\|_2} d\boldsymbol{\beta}^{(1)} \cdots d\boldsymbol{\beta}^{(m)}.$$

We can rewrite the exponential term as

$$e^{-\sigma^{-1} w_m \sqrt{p_m} \|\boldsymbol{\beta}^{(m)}\|_2} e^{-\sigma^{-1} \sum_{g=1}^{m-1} w_g \sqrt{p_g} \|\boldsymbol{\beta}^{(g)}\|_2},$$

as $\sqrt{p_1}\|\boldsymbol{\beta}^{(1)}\|_2 \geq \cdots \geq \sqrt{p_m}\|\boldsymbol{\beta}^{(m)}\|_2$ and by the fact that the penalty sequence is decreasing. Therefore, the integral becomes

$$I = m! \int_{\sqrt{p_1}\|\boldsymbol{\beta}^{(1)}\|_2 \geq \cdots \geq \sqrt{p_m}\|\boldsymbol{\beta}^{(m)}\|_2} e^{-\sigma^{-1} w_m \sqrt{p_m}\|\boldsymbol{\beta}^{(m)}\|_2} \times e^{-\sigma^{-1}\sum_{g=1}^{m-1} w_g \sqrt{p_g}\|\boldsymbol{\beta}^{(g)}\|_2} d\boldsymbol{\beta}^{(1)} \cdots d\boldsymbol{\beta}^{(m)}.$$

Continuing the pattern, we get nested integrals

$$I = m! \int_{\mathcal{R}^{p_m}|_{m-1}} e^{-\sigma^{-1} w_m \sqrt{p_m}\|\boldsymbol{\beta}^{(m)}\|_2} \int_{\mathcal{R}^{p_{m-1}}|_{m-2}} e^{-\sigma^{-1} w_m \sqrt{p_m}\|\boldsymbol{\beta}^{(m)}\|_2} \cdots \int_{\mathcal{R}^{p_2}|_1} e^{-\sigma^{-1} w_1 \sqrt{p_1}\|\boldsymbol{\beta}^{(1)}\|_2} d\boldsymbol{\beta}^{(1)} \cdots d\boldsymbol{\beta}^{(m)},$$

where $\mathcal{R}^{p_g}|_{g-1}$ is the space of all possible vectors of $\boldsymbol{\beta}^{(g)}$ restricted to $\sqrt{p_{g-1}}\|\boldsymbol{\beta}^{(g-1)}\|_2 \geq \sqrt{p_g}\|\boldsymbol{\beta}^{(g)}\|_2$.

To find this, spherical coordinates are applied. Using the change of variables $r = \|\boldsymbol{\beta}^{(g)}\|_2$ allows each vector to be represented as $\boldsymbol{\beta}^{(g)} = r\mathbf{u}$, where $r$ is the radius (magnitude) of the vector and $\mathbf{u} \in \mathbb{R}^{m_g}$ is a unit vector (giving direction). These two components allow any vector to be expressed in spherical form. Using these coordinates, the volume element transforms as (decomposed into radial and surface area components)

$$d\boldsymbol{\beta}^{(g)} = r_g^{p_g - 1} dr_g d\Omega_g.$$

Therefore, the first nested integral (rightmost one) is given by, with the restriction that $r_1\sqrt{p_1} \geq r_2\sqrt{p_2} \implies r_1 \geq r_2\sqrt{p_2}/\sqrt{p_1}$,

$$I_1 = \int_{r_2\sqrt{p_2}/\sqrt{p_1}}^{\infty} r_1^{p_1 - 1} e^{-\sigma^{-1} w_1 \sqrt{p_1} r_1} dr_1 \int_{\mathbb{S}^{p_1 - 1}} \Omega_1,$$

where $\mathbb{S}$ is the space of surface areas to integrate over. These can be dealt with separately as they do not rely on $r$ (and therefore any sorting). We have

$$I_1^S = \int_{\mathbb{S}^{p_1 - 1}} d\Omega_1 = \frac{2\pi^{p_1/2}}{\Gamma(p_1/2)},$$

where $I_1^{\mathbb{S}}$ refers to the first surface area integral and $\Gamma$ denotes a complete gamma function. Considering all groups, we have that the directional part of the integral

will contribute (by their independence of each other)

$$I^S = \int_{\mathbb{S}^{p_m-1}} \cdots \int_{\mathbb{S}^{p_2-1}} \int_{\mathbb{S}^{p_1-1}} d\Omega_1 d\Omega_2 \cdots d\Omega_m = \prod_{g=1}^{m} \frac{2\pi^{p_g/2}}{\Gamma(p_g/2)}.$$

As these can be dealt with separately, the problem is that of calculating $C^{-1} = m! I_m^r I^S$, where $I_m^r$ is the calculation of all $m$ nested radial integrals; it is a function of all previous integrals $I_m^r = f(I_1^r, \ldots, I_{m-1}^r)$.

Now, the radial part of the integral is in the form of a (scaled) upper incomplete gamma function [1]

$$\int_x^\infty t^{a-1} e^{-bt} dt = \frac{\tilde{\Gamma}(a, bx)}{b^a},$$

where $\tilde{\Gamma}$ denotes an upper incomplete gamma function. Applied to the first radial integral gives, denoted $I_1^r$,

$$\begin{aligned} I_1^r = \int_{r_2\sqrt{p_2}/\sqrt{p_1}}^\infty r_1^{p_1-1} e^{-\sigma^{-1} w_1 \sqrt{p_1} r_1} dr_1 &= \frac{\tilde{\Gamma}\left(p_1, \frac{w_1\sqrt{p_1}}{\sigma} \cdot \frac{r_2\sqrt{p_2}}{\sqrt{p_1}}\right)}{\left(\sigma^{-1} w_1 \sqrt{p_1}\right)^{p_1}} \\ &= \frac{\tilde{\Gamma}\left(p_1, w_1 r_2 \sqrt{p_2} \sigma^{-1}\right)}{\left(\sigma^{-1} w_1 \sqrt{p_1}\right)^{p_1}}. \end{aligned}$$

Using the following incomplete gamma function relation [1]

$$\tilde{\Gamma}(n, x) = (n-1)! e^{-x} \sum_{k=0}^{n-1} \frac{x^k}{k!},$$

we have

$$\begin{aligned} I_1^r = \int_{r_2\sqrt{p_2}/\sqrt{p_1}}^\infty r_1^{p_1-1} e^{-\sigma^{-1} w_1 \sqrt{p_1} r_1} dr_1 &= \frac{\tilde{\Gamma}\left(p_1, w_1 r_2 \sqrt{p_2} \sigma^{-1}\right)}{\left(\sigma^{-1} w_1 \sqrt{p_1}\right)^{p_1}} \\ &= \frac{(p_1-1)! e^{-\sigma^{-1} w_1 r_2 \sqrt{p_2}} \sum_{k_1=0}^{p_1-1} (w_1 \sqrt{p_2} \sigma^{-1})^{k_1} r_2^{k_1} / k_1!}{\left(\sigma^{-1} w_1 \sqrt{p_1}\right)^{p_1}}. \end{aligned}$$

Plugging this term into the second integrand and ignoring the directional terms

$$\begin{aligned}
I_2^r &= \int_{r_3\sqrt{p_3}/\sqrt{p_2}}^{\infty} r_2^{p_2-1} e^{-\sigma^{-1}w_2\sqrt{p_2}r_2} I_1^r dr_2, \\
&= \frac{(p_1-1)!}{\left(\sigma^{-1}w_1\sqrt{p_1}\right)^{p_1}} \int_{r_3\sqrt{p_3}/\sqrt{p_2}}^{\infty} r_2^{p_2-1} e^{-\sigma^{-1}w_2\sqrt{p_2}r_2} e^{-\sigma^{-1}w_1 r_2\sqrt{p_2}} \\
&\qquad\qquad \times \sum_{k_1=0}^{p_1-1} (w_1\sqrt{p_2}\sigma^{-1})^{k_1} r_2^{k_1}/k_1! dr_2 \\
&= \sum_{k_1=0}^{p_1-1} \frac{(p_1-1)!(w_1\sqrt{p_2}\sigma^{-1})^{k_1}}{k_1!\left(\sigma^{-1}w_1\sqrt{p_1}\right)^{p_1}} \int_{r_3\sqrt{p_3}/\sqrt{p_2}}^{\infty} r_2^{k_1} r_2^{p_2-1} e^{-\sigma^{-1}w_2\sqrt{p_2}r_2} e^{-\sigma^{-1}w_1 r_2\sqrt{p_2}} dr_2 \\
&= \sum_{k_1=0}^{p_1-1} \frac{(p_1-1)!(w_1\sqrt{p_2}\sigma^{-1})^{k_1}}{k_1!\left(\sigma^{-1}w_1\sqrt{p_1}\right)^{p_1}} \int_{r_3\sqrt{p_3}/\sqrt{p_2}}^{\infty} r_2^{p_2+k_1-1} e^{-\sigma^{-1}\sqrt{p_2}r_2(w_1+w_2)} dr_2
\end{aligned}$$

where we have swapped the integral and summation by Fubini's theorem. Again, by the incomplete gamma relation, the integral term becomes

$$\int_{r_3\sqrt{p_3}/\sqrt{p_2}}^{\infty} r_2^{p_2+k_1-1} e^{-\sigma^{-1}\sqrt{p_2}r_2(w_1+w_2)} dr_2 = \frac{\Gamma\left(p_2+k_1, (w_1+w_2)r_3\sqrt{p_3}\sigma^{-1}\right)}{\left(\sigma^{-1}(w_1+w_2)\sqrt{p_2}\right)^{p_2+k_1}}.$$

Therefore,

$$I_2^r = \sum_{k_1=0}^{p_1-1} \frac{(p_1-1)!(w_1\sqrt{p_2}\sigma^{-1})^{k_1}}{k_1!\left(\sigma^{-1}w_1\sqrt{p_1}\right)^{p_1}} \times \frac{\tilde{\Gamma}\left(p_2+k_1, (w_1+w_2)r_3\sqrt{p_3}\sigma^{-1}\right)}{\left(\sigma^{-1}(w_1+w_2)\sqrt{p_2}\right)^{p_2+k_1}}.$$

Each time we evaluate an integrand, a summation term is generated. Therefore, generalising this up until the $m-1$ integrand, we have

$$\begin{aligned}
I_{m-1}^r = \sum_{k_1=0}^{p_1-1} \cdots \sum_{k_{m-2}=0}^{p_{m-2}+k_{m-3}-1} & \frac{(p_1-1)!(\sigma^{-1}\sqrt{p_2}w_1)^{k_1}}{k_1!\left(\sigma^{-1}\sqrt{p_1}w_1\right)^{p_1}} \\
& \times \cdots \times \frac{(p_{m-2}+k_{m-3}-1)!(\sigma^{-1}\sqrt{p_{m-1}}\sum_{l=1}^{m-2}w_l)^{k_{m-2}}}{k_{m-2}!\left(\sigma^{-1}\sqrt{p_{m-2}}\sum_{l=1}^{m-2}w_l\right)^{p_{m-2}+k_{m-3}}} \\
& \times \frac{\tilde{\Gamma}\left(p_{m-1}+k_{m-2}, \sigma^{-1}\sqrt{p_m}r_m\sum_{l=1}^{m-1}w_l\right)}{\left(\sigma^{-1}\sqrt{p_{m-1}}\sum_{l=1}^{m-1}w_l\right)^{p_{m-1}+k_{m-2}}}.
\end{aligned}$$

The $m$th and final integrand is given by (ignoring the constant terms before the

integral)

$$
\begin{aligned}
I_m^r &= \int_0^\infty r_m^{p_m-1} e^{-\sigma^{-1}\sqrt{p_m} r_m w_m} I_{m-1}^r dr_m \\
&\propto \int_0^\infty r_m^{p_m-1} e^{-\sigma^{-1}\sqrt{p_m} r_m w_m} \tilde{\Gamma}\left(p_{m-1}+k_{m-2}, \sigma^{-1}\sqrt{p_m} r_m \sum_{l=1}^{m-1} w_l\right) dr_m, \\
&= \sum_{k_{m-1}=0}^{p_{m-1}+k_{m-2}-1} \frac{(p_{m-1}+k_{m-2}-1)!}{k_{m-1}!} \left(\sigma^{-1}\sqrt{p_m}\sum_{l=1}^{m-1} w_l\right)^{k_{m-1}} \\
&\qquad \times \int_0^\infty r_m^{p_m-1} e^{-\sigma^{-1}\sqrt{p_m} r_m w_m} e^{-\sigma^{-1}\sqrt{p_m} r_m \sum_{l=1}^{m-1} w_l} r_m^{k_{m-1}} dr_m \\
&= \sum_{k_{m-1}=0}^{p_{m-1}+k_{m-2}-1} \frac{(p_{m-1}+k_{m-2}-1)!}{k_{m-1}!} \left(\sigma^{-1}\sqrt{p_m}\sum_{l=1}^{m-1} w_l\right)^{k_{m-1}} \\
&\qquad \times \int_0^\infty r_m^{p_m+k_{m-1}-1} e^{-\sigma^{-1}\sqrt{p_m} r_m \sum_{l=1}^{m} w_l} dr_m \\
&= \sum_{k_{m-1}=0}^{p_{m-1}+k_{m-2}-1} \frac{(p_{m-1}+k_{m-2}-1)!}{k_{m-1}!} \left(\sigma^{-1}\sqrt{p_m}\sum_{l=1}^{m-1} w_l\right)^{k_{m-1}} \\
&\qquad \times \frac{\Gamma(p_m+k_{m-1})}{\left(\sigma^{-1}\sqrt{p_m}\sum_{l=1}^{m} w_l\right)^{p_m+k_{m-1}}} \\
&= \sum_{k_{m-1}=0}^{p_{m-1}+k_{m-2}-1} \frac{(p_{m-1}+k_{m-2}-1)!}{k_{m-1}!} \left(\sigma^{-1}\sqrt{p_m}\sum_{l=1}^{m-1} w_l\right)^{k_{m-1}} \\
&\qquad \times \frac{(p_m+k_{m-1}-1)!}{\left(\sigma^{-1}\sqrt{p_m}\sum_{l=1}^{m} w_l\right)^{p_m+k_{m-1}}}
\end{aligned}
$$

where the complete gamma function satisfies $\int_0^\infty r^{a-1}e^{-br}dr = \Gamma(a)/b^a$, justifying the second to last step, and $\Gamma(n) = (n-1)!$, if $n$ is a positive integer, which is the case for $p_m + k_{m-1}$, justifying the last step.

Putting this all together, including the directional integrand, $I^S$, gives

$$
\begin{aligned}
C^{-1} = m! \prod_{g=1}^{m} \frac{2\pi^{p_g/2}}{\Gamma(p_g/2)} & \\
\times \sum_{k_1=0}^{p_1-1} \cdots \sum_{k_{m-2}=0}^{p_{m-2}+k_{m-3}-1} \sum_{k_{m-1}=0}^{p_{m-1}+k_{m-2}-1} & \left[ \frac{(p_1-1)!(\sigma^{-1}\sqrt{p_2}w_1)^{k_1}}{k_1!(\sigma^{-1}\sqrt{p_1}w_1)^{p_1}} \times \cdots \right. \\
\times \frac{(p_{m-1}+k_{m-2}-1)!(\sigma^{-1}\sqrt{p_m}\sum_{l=1}^{m-1} w_l)^{k_{m-1}}}{k_{m-1}!(\sigma^{-1}\sqrt{p_{m-1}}\sum_{l=1}^{m-1} w_l)^{p_{m-1}+k_{m-2}}} & \\
\left. \times \frac{(p_m+k_{m-1}-1)!}{(\sigma^{-1}\sqrt{p_m}\sum_{l=1}^{m} w_l)^{p_m+k_{m-1}}} \right]. &
\end{aligned}
$$

This can be simplified to

$$
\begin{aligned}
C^{-1} = m! \prod_{g=1}^{m} \frac{2\pi^{p_g/2}}{\Gamma(p_g/2)} & \\
\times \sum_{\mathbf{k}\in\mathbb{K}} \prod_{j=1}^{m-1} \frac{(p_j+k_{j-1}-1)!(\sigma^{-1}\sqrt{p_j}\sum_{l=1}^{j} w_l)^{k_j}(p_m+k_{m-1}-1)!}{k_j!(\sigma^{-1}\sqrt{p_j}\sum_{l=1}^{j} w_l)^{p_j+k_{j-1}}(\sigma^{-1}\sqrt{p_m}\sum_{l=1}^{m} w_l)^{p_m+k_{m-1}}}, &
\end{aligned}
$$

where $\mathbb{K} = \{\mathbf{k} \in \mathbb{Z}^{m-1} : k_1 \in [0, p_1 - 1],\ k_2 \in [0, p_2 + k_1 - 1],\ \ldots,\ k_{m-1} \in [0, p_{m-1} + k_{m-2} - 1]\}$, with $k_0 = 0$. □

### E.1.2 Bayesian SGS (BSGS)

*Proof of Proposition 7.2.* We assume a random variable $\mathbf{z} = (z_1, \ldots, z_p)$ has an SGS Laplace prior, given by

$$
\pi(\mathbf{z} \mid \sigma^2, \mathbf{w}, \mathbf{v}) \propto \exp\left\{ -\frac{1}{\sigma}\alpha \sum_{i=1}^{p} v_{r_v(\mathbf{z},i)}|z_i| - \frac{1}{\sigma}(1-\alpha)\sum_{j=1}^{m} w_{r_g(\mathbf{z},j)}\sqrt{p_j}\|\mathbf{z}^{(j)}\|_2 \right\}.
$$

Now, define $\boldsymbol{\beta} = \hat{\mathbf{A}}^{-1}\mathbf{z}$ such that $z_i = \beta_i \hat{a}_i$ for $i \in G_j$. Then, by the transformation of variables, the prior distribution for $\boldsymbol{\beta}$ is given by

$$
\pi\left(\boldsymbol{\beta} \mid c_v, c_g, \sigma^2; \mathbf{w}, \mathbf{v}\right) \propto \left|\det(d\mathbf{z}/d\boldsymbol{\beta})\right| \pi\left(\mathbf{z} \mid c_v, c_g, \sigma^2; \mathbf{w}, \mathbf{v}\right).
$$

We have that $(d\mathbf{z}/d\boldsymbol{\beta})_i = \hat{a}_i$, so that

$$\begin{aligned}\det(d\mathbf{z}/d\boldsymbol{\beta}) &= \prod_{i=1}^{p} \hat{a}_i = \prod_{j=1}^{m} c_g^{p_j \mathbf{1}\{\gamma_j=1\}} \prod_{i \in G_j} c_v^{\mathbf{1}\{\delta_i=1\}} \\ &= \prod_{j=1}^{m} c_g^{p_j \mathbf{1}\{\gamma_j=1\}} c_v^{\sum_{i \in G_j} \mathbf{1}\{\delta_i=1\}} = c_g^{\sum_{j=1}^{m} p_j \mathbf{1}(\gamma_j=1)} c_v^{\sum_{i=1}^{p} \mathbf{1}(\delta_i=1)}.\end{aligned}$$

We can rewrite the prior on $z$ in terms of $\boldsymbol{\beta}$:

$$\begin{aligned}&\pi\left(\mathbf{z} \mid c_v, c_g, \sigma^2; \mathbf{w}, \mathbf{v}\right) \\ &\propto \exp\left\{-\frac{1}{\sigma}\alpha \sum_{i=1}^{p} v_{r_v(\mathbf{z},i)}|z_i| - \frac{1}{\sigma}(1-\alpha)\sum_{j=1}^{m} w_{r_g(\mathbf{z},j)}\sqrt{p_j}\|\mathbf{z}^{(j)}\|_2\right\} \\ &= \prod_{i=1}^{p} \exp\left\{-\frac{1}{\sigma}\alpha v_{r_v(\mathbf{z},i)}|z_i|\right) \prod_{j=1}^{m} \exp\left\{-\frac{1}{\sigma}(1-\alpha) w_{r_g(\mathbf{z},j)}\sqrt{p_j}\|\mathbf{z}^{(j)}\|_2\right\} \\ &= \prod_{i=1}^{p} \exp\left\{-\frac{1}{\sigma}\hat{a}_i \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)}|\beta_i|\right\} \prod_{j=1}^{m} \exp\left\{-\frac{1}{\sigma}(1-\alpha) w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta},j)}\sqrt{p_j}\|\hat{\mathbf{A}}^{(j)}\boldsymbol{\beta}^{(j)}\|_2\right\}.\end{aligned}$$

Therefore, the prior for $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\pi\left(\boldsymbol{\beta} \mid c_v, c_g, \sigma^2; \mathbf{w}, \mathbf{v}\right) \propto c_g^{\sum_{j=1}^{m} p_j \mathbf{1}(\gamma_j=1)} c_v^{\sum_{i=1}^{p} \mathbf{1}(\delta_i=1)} &\prod_{i=1}^{p} \exp\left\{-\frac{1}{\sigma}|\beta_i|\hat{a}_i \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)}\right\} \\ &\times \prod_{j=1}^{m} \exp\left\{-\frac{1}{\sigma}\|\hat{\mathbf{A}}^{(j)}\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j}(1-\alpha) w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta},j)}\right\},\end{aligned}$$

corresponding to the BSGS prior. □

## E.2 SAEM for BGSLOPE

### E.2.1 Simulation Step Derivations

Here, the conditional distributions required for the Gibbs sampler are derived for each latent variable.

**Inclusion Vector ($\boldsymbol{\gamma}$)**

According to the dependency graph (Figure 7.2), we simulate $\gamma_j$, for $j \in [m]$, via

$$\gamma_j \sim \pi(\gamma_j \mid \boldsymbol{\gamma}_{-j}, c, \mathbf{y}, \boldsymbol{\beta}, \sigma, \theta) = \pi(\gamma_j \mid \boldsymbol{\gamma}_{-j}, c, \boldsymbol{\beta}, \sigma, \theta).$$

As $\gamma_j$ is binary, the posterior is a Bernoulli distribution with active probability

$$\mathbb{P}(\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}, c, \boldsymbol{\beta}, \sigma, \theta) = \frac{\mathbb{P}(\gamma_j = 1 \mid \theta)\pi(\boldsymbol{\beta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c, \sigma)}{\sum_{\gamma_j \in \{0,1\}} \mathbb{P}(\gamma_j \mid \theta)\pi(\boldsymbol{\beta} \mid \gamma_j, \boldsymbol{\gamma}_{-j}, c, \sigma)}.$$

This describes the posterior probability of a binary signal indicator for the $j$th group, using the prior $\mathbb{P}(\gamma_j = 1 \mid \theta) = \theta$ and the conditional likelihood of $\boldsymbol{\beta}$ given $\gamma_j = 1$ and $\gamma_j = 0$.

We have that $\mathbb{P}(\gamma_j = 1 \mid \theta) = \theta$ and $\mathbb{P}(\gamma_j = 0 \mid \theta) = 1 - \theta$ via the Bernoulli distribution on $\gamma_j$, and,

$$\begin{aligned}
\pi(\boldsymbol{\beta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c, \sigma) &= c^{p_j} \exp\left\{\frac{-c}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}^1\boldsymbol{\beta}, j)}\right\} \\
&\quad \times c^{\sum_{k\in[m]\setminus\{j\}} p_k \mathbf{1}(\gamma_k = 1)} \prod_{k \in [m]\setminus\{j\}} \exp\left\{\frac{-1}{\sigma}\tilde{a}_k \|\boldsymbol{\beta}^{(k)}\|_2 \sqrt{p_k} w_{r_g(\tilde{\mathbf{A}}^1\boldsymbol{\beta}, k)}\right\}, \\
\pi(\boldsymbol{\beta} \mid \gamma_j = 0, \boldsymbol{\gamma}_{-j}, c, \sigma) &= \exp\left\{\frac{-1}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}^0\boldsymbol{\beta}, j)}\right\} \\
&\quad \times c^{\sum_{k\in[m]\setminus\{j\}} p_k \mathbf{1}(\gamma_k = 1)} \prod_{k \in [m]\setminus\{j\}} \exp\left\{\frac{-1}{\sigma}\tilde{a}_k \|\boldsymbol{\beta}^{(k)}\|_2 \sqrt{p_k} w_{r_g(\tilde{\mathbf{A}}^0\boldsymbol{\beta}, k)}\right\},
\end{aligned}$$

where $\tilde{\mathbf{A}}^1$ and $\tilde{\mathbf{A}}^0$ are the same as $\tilde{\mathbf{A}}$ except for the $j$th group, for which $\tilde{a}_j^1 = c$ and $\tilde{a}_j^0 = 1$. This leads to

$$\begin{aligned}
\mathbb{P}(\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}, c, \boldsymbol{\beta}, \sigma, \theta) &= \frac{\mathbb{P}(\gamma_j = 1 \mid \theta)\pi(\boldsymbol{\beta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c, \sigma)}{\sum_{\gamma_j \in \{0,1\}} \mathbb{P}(\gamma_j \mid \theta)\pi(\boldsymbol{\beta} \mid \gamma_j, \boldsymbol{\gamma}_{-j}, c, \sigma)} \\
&= \left[\frac{\mathbb{P}(\gamma_j = 1 \mid \theta)\pi(\boldsymbol{\beta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c, \sigma) + \mathbb{P}(\gamma_j = 0 \mid \theta)\pi(\boldsymbol{\beta} \mid \gamma_j = 0, \boldsymbol{\gamma}_{-j}, c, \sigma)}{\mathbb{P}(\gamma_j = 1 \mid \theta)\pi(\boldsymbol{\beta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c, \sigma)}\right]^{-1} \\
&= \left[1 + \frac{\mathbb{P}(\gamma_j = 0 \mid \theta)\pi(\boldsymbol{\beta} \mid \gamma_j = 0, \boldsymbol{\gamma}_{-j}, c, \sigma)}{\mathbb{P}(\gamma_j = 1 \mid \theta)\pi(\boldsymbol{\beta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c, \sigma)}\right]^{-1} \\
&= \left[1 + \frac{(1-\theta)\pi(\boldsymbol{\beta} \mid \gamma_j = 0, \boldsymbol{\gamma}_{-j}, c, \sigma)}{\theta\pi(\boldsymbol{\beta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c, \sigma)}\right]^{-1},
\end{aligned}$$

where the $c^{\sum_{k\in[m]\setminus\{j\}} p_k \mathbf{1}(\gamma_k=1)}$ term present in both has been cancelled, and

$$\begin{aligned}
&\pi(\boldsymbol{\beta} \mid \gamma_j = 0, \boldsymbol{\gamma}_{-j}, c, \sigma) \\
&= \exp\left\{\frac{-1}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2\sqrt{p_j}w_{r_g(\tilde{\mathbf{A}}^0\boldsymbol{\beta},j)}\right\}\prod_{k\in[m]\setminus\{j\}}\exp\left\{\frac{-1}{\sigma}\tilde{a}_{(k)}\|\boldsymbol{\beta}^{(k)}\|_2\sqrt{p_k}w_{r_g(\tilde{\mathbf{A}}^0\boldsymbol{\beta},k)}\right\}, \\
&\pi(\boldsymbol{\beta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c, \sigma) \\
&= c^{p_j}\exp\left\{\frac{-c}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2\sqrt{p_j}w_{r_g(\tilde{\mathbf{A}}^1\boldsymbol{\beta},j)}\right\}\prod_{k\in[m]\setminus\{j\}}\exp\left\{\frac{-1}{\sigma}\tilde{a}_{(k)}\|\boldsymbol{\beta}^{(k)}\|_2\sqrt{p_k}w_{r_g(\tilde{\mathbf{A}}^1\boldsymbol{\beta},k)}\right\}.
\end{aligned}$$

Sampling here requires updating an ordered list at each group $j \in [m]$, which is expensive, so $\tilde{\mathbf{A}}^1$ and $\tilde{\mathbf{A}}^0$ are approximated by using the previous estimate of $\tilde{\mathbf{A}}$. This partially retains information on $\gamma_j$ and stabilises updates; a similar approach is taken for ABSLOPE. The approximation allows the $\prod_{k\in[m]\setminus\{j\}}$ terms to cancel, so that

$$\begin{aligned}
\mathbb{P}(\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}, c, \boldsymbol{\beta}, \sigma, \theta, \tilde{\mathbf{A}}) &= \left[1 + \frac{(1-\theta)\exp\left\{\frac{-1}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2\sqrt{p_j}w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta},j)}\right\}}{\theta c^{p_j}\exp\left\{\frac{-c}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2\sqrt{p_j}w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta},j)}\right\}}\right]^{-1} \\
&= \frac{\theta c^{p_j}\exp\left\{\frac{-c}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2\sqrt{p_j}w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta},j)}\right\}}{(1-\theta)\exp\left\{\frac{-1}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2\sqrt{p_j}w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta},j)}\right\} + \theta c^{p_j}\exp\left\{\frac{-c}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2\sqrt{p_j}w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta},j)}\right\}}.
\end{aligned}$$

This leads to posterior

$$\begin{aligned}
&\gamma_j \sim \text{Bernoulli}\left(\frac{L_1}{L_1+L_2}\right), \text{ where,} \\
&L_1 = \theta c^{p_j}\exp\left\{\frac{-c}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2\sqrt{p_j}w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta},j)}\right\}, \\
&L_2 = (1-\theta)\exp\left\{\frac{-1}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2\sqrt{p_j}w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta},j)}\right\}.
\end{aligned}$$

**Mixture Proportion ($\theta$)**

We have that

$$\theta \sim \pi(\theta \mid \boldsymbol{\gamma}, c, \mathbf{y}, \boldsymbol{\beta}, \sigma, \tilde{\mathbf{A}}) = \pi(\theta \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma, \tilde{\mathbf{A}}) \propto \pi(\theta)\pi(\boldsymbol{\gamma} \mid \theta),$$

where $\pi(\theta)$ is a Beta$(d_1, d_2)$ distribution and $\pi(\boldsymbol{\gamma} \mid \theta)$ is a multivariate Bernoulli distribution. Therefore, by conjugacy, the posterior is given by a Beta distribution:

$$\theta \sim \text{Beta}\left(d_1 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1), d_2 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 0)\right).$$

**Signal Strength Ratio ($c$)**

We sample $c$ from

$$c \sim \pi(c \mid \boldsymbol{\gamma}, \mathbf{y}, \boldsymbol{\beta}, \sigma, \theta, \tilde{\mathbf{A}}) = \pi(c \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma, \tilde{\mathbf{A}}) \propto \pi(c)\pi(\boldsymbol{\beta} \mid c, \boldsymbol{\gamma}, \sigma, \tilde{\mathbf{A}}).$$

As $c \sim \mathcal{U}[0, 1]$, this reduces to the BGSLOPE prior on $\boldsymbol{\beta}$ for an active group (the prior for non-active groups does not contain $c$)

$$\pi(\boldsymbol{\beta} \mid c, \boldsymbol{\gamma}, \sigma, \tilde{\mathbf{A}}) = c^{\sum_{j=1}^{m} p_j \mathbf{1}(\gamma_j = 1)} \exp\left\{\sum_{j=1}^{m} -\frac{c}{\sigma} \|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)} \mathbf{1}(\gamma_j = 1)\right\}.$$

This is in the form of a Gamma distribution, truncated to $[0, 1]$ to satisfy the constraint on $c$,

$$c \sim \text{Gamma}\left(1 + \sum_{j=1}^{m} p_j 1(\gamma_j = 1), \frac{1}{\sigma} \sum_{j=1}^{m} \|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}\boldsymbol{\beta}, j)} \mathbf{1}(\gamma_j = 1)\right).$$

### E.2.2 Stochastic and Maximisation Step Derivations

**Update for $\sigma$**

Considering only the terms of the log-likelihood (Equation 7.7) containing $\sigma$

$$\begin{aligned}
\sigma_{[t]}^{\text{MLE}} &= \underset{\sigma \in \mathbb{R}_{>0}}{\arg\max} \Bigg\{ -(n+2)\log\sigma - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{[t]}\|_2^2 \\
&\qquad\qquad - \frac{1}{\sigma}\sum_{j=1}^{m} (\tilde{a}_j)_{[t]} \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]}, j)} \|\boldsymbol{\beta}_{[t]}^{(j)}\|_2 \Bigg\} \\
&= \underset{\sigma \in \mathbb{R}_{>0}}{\arg\max} \left\{ -(n+2)\log\sigma - \frac{K_1}{2\sigma^2} - \frac{1}{\sigma} K_2 \right\},
\end{aligned}$$

where $K_1 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{[t]}\|_2^2$ and $K_2 = \sum_{j=1}^{m} (\tilde{a}_j)_{[t]} \|\boldsymbol{\beta}_{[t]}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\tilde{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]}, j)}$. Differentiating this with respect to $\sigma$ gives

$$K_1 + K_2\sigma = (n+2)\sigma^2 \implies \sigma = \frac{K_2 \pm \sqrt{K_2^2 + 4K_1(n+2)}}{2(n+2)}.$$

The negative root is negative if

$$\begin{aligned} K_2 - \sqrt{K_2^2 + 4K_1(n+2)} < 0 &\implies K_2 < \sqrt{K_2^2 + 4K_1(n+2)} \\ &\implies 4K_1(n+2) > 0, \end{aligned}$$

which is true as $K_1 > 0$ (unless we have perfect $\mathbf{y}$ recovery, which leads to a model with no variance). As $\sigma > 0$, the positive root is used:

$$\sigma_{[t]}^{\text{MLE}} = \frac{K_2 + \sqrt{K_2^2 + 4K_1(n+2)}}{2(n+2)}.$$

Omitting the gSLOPE penalty, the estimator simplifies to $\sigma_{[t]} = \sqrt{K_1/(n+2)}$, which matches the classical MLE formula of $\sigma$ when $\boldsymbol{\beta}$ is also estimated via MLE [99].

## E.3 SAEM for BSGS

### E.3.1 Simulation Step Derivations

#### Group Inclusion Vector ($\boldsymbol{\gamma}$)

By Figure 7.3, $\boldsymbol{\gamma}$ is simulated by, for $j \in [m]$,

$$\gamma_j \sim \pi(\gamma_j \mid \boldsymbol{\gamma}_{-j}, c_v, c_g, \mathbf{y}, \boldsymbol{\beta}, \sigma, \theta_v, \theta_g, \boldsymbol{\delta}) = \pi(\gamma_j \mid \boldsymbol{\gamma}_{-j}, c_v, c_g, \boldsymbol{\beta}, \sigma, \theta_v, \theta_g, \boldsymbol{\delta}).$$

As $\gamma_j$ is binary, the posterior is given by a Bernoulli distribution with success probability

$$\begin{aligned} &\mathbb{P}(\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}, c_v, c_g, \boldsymbol{\beta}, \sigma, \theta_v, \theta_g, \boldsymbol{\delta}) \\ &\qquad = \frac{\mathbb{P}(\gamma_j = 1 \mid \theta_g)\pi(\boldsymbol{\beta}, \boldsymbol{\delta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c_v, c_g, \sigma, \theta_v, \theta_g)}{\sum_{\gamma_j \in \{0,1\}} \mathbb{P}(\gamma_j \mid \theta_g)\pi(\boldsymbol{\beta}, \boldsymbol{\delta} \mid \gamma_j, \boldsymbol{\gamma}_{-j}, c_v, c_g, \sigma, \theta_v, \theta_g)}. \end{aligned}$$

This can be rewritten as

$$\mathbb{P}(\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}, c_v, c_g, \boldsymbol{\beta}, \sigma, \theta_v, \theta_g, \boldsymbol{\delta}) \tag{E.2}$$
$$= \left[1 + \frac{\mathbb{P}(\gamma_j = 0 \mid \theta_g)\pi(\boldsymbol{\beta}, \boldsymbol{\delta} \mid \gamma_j = 0, \boldsymbol{\gamma}_{-j}, c_v, c_g, \sigma, \theta_v, \theta_g)}{\mathbb{P}(\gamma_j = 1 \mid \theta_g)\pi(\boldsymbol{\beta}, \boldsymbol{\delta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c_v, c_g, \sigma, \theta_v, \theta_g)}\right]^{-1},$$

where $\mathbb{P}(\gamma_j = 1 \mid \theta_g) = \theta_g$ and $\mathbb{P}(\gamma_j = 0 \mid \theta_g) = 1 - \theta_g$. The joint conditional distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ can be decomposed into the active and non-active cases

$$\pi(\boldsymbol{\beta}, \boldsymbol{\delta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c_v, c_g, \sigma, \theta_v, \theta_g) = \pi(\boldsymbol{\delta} \mid \gamma_j = 1, \theta_v)\pi(\boldsymbol{\beta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c_v, c_g, \sigma, \boldsymbol{\delta}),$$
$$\pi(\boldsymbol{\beta}, \boldsymbol{\delta} \mid \gamma_j = 0, \boldsymbol{\gamma}_{-j}, c_v, c_g, \sigma, \theta_v, \theta_g) = \pi(\boldsymbol{\delta} \mid \gamma_j = 0, \theta_v)\pi(\boldsymbol{\beta} \mid \gamma_j = 0, \boldsymbol{\gamma}_{-j}, c_v, c_g, \sigma, \boldsymbol{\delta}).$$

The first term is given by

$$\pi(\boldsymbol{\delta} \mid \gamma_j = 1, \theta_v) = \prod_{i \in G_j} \theta_v^{\delta_i}(1 - \theta_v)^{1-\delta_i}$$
$$\times \prod_{k \in [m]\setminus\{j\}} \prod_{i \in G_k} \left[\theta_v^{\delta_i}(1 - \theta_v)^{1-\delta_i}\mathbf{1}(\gamma_k = 1) + \mathbf{1}(\gamma_k = 0, \delta_i = 0)\right],$$
$$\pi(\boldsymbol{\delta} \mid \gamma_j = 0, \theta_v) = \prod_{k \in [m]\setminus\{j\}} \prod_{i \in G_k} \left[\theta_v^{\delta_i}(1 - \theta_v)^{1-\delta_i}\mathbf{1}(\gamma_k = 1) + \mathbf{1}(\gamma_k = 0, \delta_i = 0)\right].$$

The second term is the prior of the regression coefficients, which under an active group is

$$\pi(\boldsymbol{\beta} \mid \gamma_j = 1, \boldsymbol{\gamma}_{-j}, c_v, c_g, \sigma, \boldsymbol{\delta}) = c_g^{p_j} c_v^{\sum_{i \in G_j} \mathbf{1}\{\delta_i = 1\}}$$
$$\times \prod_{i \in G_j} \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g c_v^{\mathbf{1}(\delta_i = 1)} v_{r_v(\hat{\mathbf{A}}^1 \boldsymbol{\beta}, i)}\right\}$$
$$\times \exp\left\{-\frac{1}{\sigma} c_g \|\boldsymbol{\kappa}^{(j)}\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}^1 \boldsymbol{\beta}, j)}\right\}$$
$$\times \prod_{k \in [m]\setminus\{j\}} c_g^{p_k \mathbf{1}\{\gamma_k = 1\}} c_v^{\sum_{i \in G_k} \mathbf{1}\{\delta_i = 1\}}$$
$$\times \prod_{i \notin G_j} \exp\left\{-\frac{1}{\sigma}|\beta_i| \hat{a}_i^1 v_{r_v(\hat{\mathbf{A}}^1 \boldsymbol{\beta}, i)}\right\}$$
$$\times \prod_{k \in [m]\setminus\{j\}} \exp\left\{-\frac{1}{\sigma}\|(\hat{\mathbf{A}}^1)^{(k)}\boldsymbol{\beta}^{(k)}\|_2 \sqrt{p_k} w_{r_g(\hat{\mathbf{A}}^1 \boldsymbol{\beta}, k)}\right\},$$

where $\boldsymbol{\kappa}^{(j)} = \text{diag}(c_v^{\mathbf{1}(\delta_i=1)})$ for all $i \in G_j$. Matrices $\hat{\mathbf{A}}^{\,1}$ and $\hat{\mathbf{A}}^{\,0}$ have the same entries as $\hat{\mathbf{A}}$ except for all $i \in G_j$, where the entries are $\hat{a}_i^0 = 1$ and $\hat{a}_i^1 = c_g c_v^{\mathbf{1}(\delta_i=1)}$. Under an inactive group, the prior is given by

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid \gamma_j = 0, \boldsymbol{\gamma}_{-j}, c_v, c_g, \sigma, \boldsymbol{\delta}) = &\prod_{i \in G_j} \exp\left\{-\frac{1}{\sigma}|\beta_i| v_{r_v(\hat{\mathbf{A}}^{\,0}\boldsymbol{\beta}, i)}\right\} \\ &\times \exp\left\{-\frac{1}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}^{\,0}\boldsymbol{\beta}, j)}\right\} \\ &\times \prod_{k \in [m]\setminus\{j\}} c_g^{p_k \mathbf{1}\{\gamma_k = 1\}} c_v^{\sum_{i \in G_k} \mathbf{1}\{\delta_i = 1\}} \\ &\times \prod_{i \notin G_j} \exp\left\{-\frac{1}{\sigma}|\beta_i| \hat{a}_i^0 v_{r_v(\hat{\mathbf{A}}^{\,0}\boldsymbol{\beta}, i)}\right\} \\ &\times \prod_{k \in [m]\setminus\{j\}} \exp\left\{-\frac{1}{\sigma}\|(\hat{\mathbf{A}}^{\,0})^{(k)}\boldsymbol{\beta}^{(k)}\|_2 \sqrt{p_k} w_{r_g(\hat{\mathbf{A}}^{\,0}\boldsymbol{\beta}, k)}\right\}.\end{aligned}$$

Therefore, considering first the fraction from Equation E.2 we have

$$\begin{aligned}\text{Numerator} = (1-\theta_g) &\prod_{k \in [m]\setminus\{j\}} \prod_{i \in G_k} \left[\theta_v^{\delta_i}(1-\theta_v)^{1-\delta_i}\mathbf{1}(\gamma_k = 1) + \mathbf{1}(\gamma_k = 0, \delta_i = 0)\right] \\ &\times \prod_{i \in G_j} \exp\left\{-\frac{1}{\sigma}|\beta_i| v_{r_v(\hat{\mathbf{A}}^{\,0}\boldsymbol{\beta}, i)}\right\} \\ &\times \exp\left\{-\frac{1}{\sigma}\|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}^{\,0}\boldsymbol{\beta}, j)}\right\} \\ &\times \prod_{k \in [m]\setminus\{j\}} c_g^{p_k \mathbf{1}\{\gamma_k = 1\}} c_v^{\sum_{i \in G_k} \mathbf{1}\{\delta_i = 1\}} \prod_{i \notin G_j} \exp\left\{-\frac{1}{\sigma}|\beta_i| \hat{a}_i^1 v_{r_v(\hat{\mathbf{A}}^{\,0}\boldsymbol{\beta}, i)}\right\} \\ &\times \prod_{k \in [m]\setminus\{j\}} \exp\left\{-\frac{1}{\sigma}\|(\hat{\mathbf{A}}^{\,0})^{(k)}\boldsymbol{\beta}^{(k)}\|_2 \sqrt{p_k} w_{r_g(\hat{\mathbf{A}}^{\,0}\boldsymbol{\beta}, k)}\right\},\end{aligned}$$

$$
\begin{aligned}
\text{Denominator} = \theta_g & \prod_{i \in G_j} \theta_v^{\delta_i} (1-\theta_v)^{1-\delta_i} \\
& \times \prod_{k \in [m] \setminus \{j\}} \prod_{i \in G_k} \left[ \theta_v^{\delta_i} (1-\theta_v)^{1-\delta_i} \mathbf{1}(\gamma_k = 1) + \mathbf{1}(\gamma_k = 0, \delta_i = 0) \right] \\
& \times c_g^{p_j} c_v^{\sum_{i \in G_j} \mathbf{1}\{\delta_i = 1\}} \prod_{i \in G_j} \exp\left\{ -\frac{1}{\sigma} |\beta_i| c_g c_v^{\mathbf{1}(\delta_i = 1)} v_{r_v(\hat{\mathbf{A}}^1 \boldsymbol{\beta}, i)} \right\} \\
& \times \exp\left\{ -\frac{1}{\sigma} c_g \|\boldsymbol{\kappa}^{(j)} \boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}^1 \boldsymbol{\beta}, j)} \right\} \\
& \times \prod_{k \in [m] \setminus \{j\}} c_g^{p_k \mathbf{1}\{\gamma_k = 1\}} c_v^{\sum_{i \in G_k} \mathbf{1}\{\delta_i = 1\}} \prod_{i \notin G_j} \exp\left\{ -\frac{1}{\sigma} |\beta_i| \hat{a}_i^1 v_{r_v(\hat{\mathbf{A}}^1 \boldsymbol{\beta}, i)} \right\} \\
& \times \prod_{k \in [m] \setminus \{j\}} \exp\left\{ -\frac{1}{\sigma} \|(\hat{\mathbf{A}}^1)^{(k)} \boldsymbol{\beta}^{(k)}\|_2 \sqrt{p_k} w_{r_g(\hat{\mathbf{A}}^1 \boldsymbol{\beta}, k)} \right\}.
\end{aligned}
$$

As with BGSLOPE (Appendix E.2.1), we will use $\hat{\mathbf{A}}$ as an approximation for both $\hat{\mathbf{A}}^0$ and $\hat{\mathbf{A}}^1$. Therefore, several terms can be cancelled to obtain

$$
\begin{aligned}
\text{Numerator} = (1-\theta_g) & \times \prod_{i \in G_j} \exp\left\{ -\frac{1}{\sigma} |\beta_i| v_{r_v(\hat{\mathbf{A}} \boldsymbol{\beta}, i)} \right\} \\
& \times \exp\left\{ -\frac{1}{\sigma} \|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}} \boldsymbol{\beta}, j)} \right\}, \\
\text{Denominator} = \theta_g & \prod_{i \in G_j} \theta_v^{\delta_i} (1-\theta_v)^{1-\delta_i} \\
& \times c_g^{p_j} c_v^{\sum_{i \in G_j} \mathbf{1}\{\delta_i = 1\}} \prod_{i \in G_j} \exp\left\{ -\frac{1}{\sigma} |\beta_i| c_g c_v^{\mathbf{1}(\delta_i = 1)} v_{r_v(\hat{\mathbf{A}} \boldsymbol{\beta}, i)} \right\} \\
& \times \exp\left\{ -\frac{1}{\sigma} c_g \|\boldsymbol{\kappa}^{(j)} \boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}} \boldsymbol{\beta}, j)} \right\}.
\end{aligned}
$$

Finally,

$$\begin{aligned}
&\mathbb{P}(\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j}, c_v, c_g, \boldsymbol{\beta}, \sigma, \theta_v, \theta_g, \boldsymbol{\delta}) = \frac{L_1'}{L_1' + L_2'}, \text{ where,} \\
&L_1' = \theta_g c_g^{p_j} c_v^{\sum_{i \in G_j} \mathbf{1}\{\delta_i = 1\}} \exp\left\{-\frac{c_g}{\sigma} \|\boldsymbol{\kappa}^{(j)} \boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)}\right\} \\
&\qquad \times \prod_{i \in G_j} \theta_v^{\delta_i} (1 - \theta_v)^{1 - \delta_i} \exp\left\{-\frac{c_g}{\sigma} |\beta_i| c_v^{\mathbf{1}(\delta_i = 1)} v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\}, \\
&L_2' = (1 - \theta_g) \exp\left\{-\frac{1}{\sigma} \|\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)}\right\} \prod_{i \in G_j} \exp\left\{-\frac{1}{\sigma} |\beta_i| v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\}.
\end{aligned}$$

Therefore, $\gamma_j \sim \text{Bernoulli}\left(\frac{L_1'}{L_1' + L_2'}\right)$, for $j \in [m]$.

**Variable Inclusion Vector ($\boldsymbol{\delta}$)**

For each $i \in [p]$, the BSGS dependency graph shows that we simulate $\delta_i$ by

$$\delta_i \sim \pi(\delta_i \mid \boldsymbol{\delta}_{-i}, c_v, c_g, \mathbf{y}, \boldsymbol{\beta}, \sigma, \theta_v, \theta_g, \boldsymbol{\gamma}) = \pi(\delta_i \mid \boldsymbol{\delta}_{-i}, c_v, c_g, \boldsymbol{\beta}, \sigma, \theta_v, \theta_g, \boldsymbol{\gamma}).$$

Similar to the derivation for the $\boldsymbol{\gamma}$ simulation step, this can be expressed as, denoting $j$ as the group that variable $i$ sits in,

$$\begin{aligned}
&\mathbb{P}(\delta_i = 1 \mid \boldsymbol{\delta}_{-i}, c_v, c_g, \boldsymbol{\beta}, \sigma, \theta_v, \theta_g, \boldsymbol{\gamma}) \\
&\quad = \frac{\mathbb{P}(\delta_i = 1 \mid \theta_v, \gamma_j) \pi(\boldsymbol{\beta} \mid \delta_i = 1, \boldsymbol{\delta}_{-i}, c_v, c_g, \sigma, \theta_v, \theta_g, \boldsymbol{\gamma})}{\sum_{\delta_i \in \{0,1\}} \mathbb{P}(\delta_i \mid \theta_v, \gamma_j) \pi(\boldsymbol{\beta} \mid \delta_i, \boldsymbol{\delta}_{-i}, c_v, c_g, \sigma, \theta_v, \theta_g, \boldsymbol{\gamma})} \\
&\quad = \left[1 + \frac{\mathbb{P}(\delta_i = 0 \mid \theta_v, \gamma_j) \pi(\boldsymbol{\beta} \mid \delta_i = 0, \boldsymbol{\delta}_{-i}, c_v, c_g, \sigma, \theta_v, \theta_g, \boldsymbol{\gamma})}{\mathbb{P}(\delta_i = 1 \mid \theta_v, \gamma_j) \pi(\boldsymbol{\beta} \mid \delta_i = 1, \boldsymbol{\delta}_{-i}, c_v, c_g, \sigma, \theta_v, \theta_g, \boldsymbol{\gamma})}\right]^{-1}. \qquad \text{(E.3)}
\end{aligned}$$

Now, the prior terms are given by $\mathbb{P}(\delta_i = 1 \mid \theta_v, \gamma_j) = \theta_v \mathbf{1}(\gamma_j = 1)$ and $\mathbb{P}(\delta_i = 0 \mid \theta_v, \gamma_j) = 1 - \theta_v \mathbf{1}(\gamma_j = 1)$. Denoting $\hat{\mathbf{A}}^1$ and $\hat{\mathbf{A}}^0$ as matrices with the same elements as $\hat{\mathbf{A}}$ except for variable $i$, which has entries $\hat{\mathbf{a}}_i^0 = c_g^{\mathbf{1}(\gamma_j = 1)}$ and $\hat{\mathbf{a}}_i^1 = c_v c_g$. Then, the

second term, for an active variable, is given by

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid \delta_i = 1, \boldsymbol{\delta}_{-i}, c_v, c_g, \sigma, \theta_v, \theta_g, \boldsymbol{\gamma}) = c_g^{p_j} c_v c_g^{\sum_{k\in[m]\setminus\{j\}} p_k \mathbf{1}(\gamma_k=1)} c_v^{\sum_{k\in[p]\setminus\{i\}} \mathbf{1}(\delta_k=1)} \\ \times \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g c_v v_{r_v(\hat{\mathbf{A}}^1\boldsymbol{\beta},i)}\right\} \\ \times \exp\left\{-\frac{1}{\sigma} c_g \|\tilde{\boldsymbol{\kappa}}^{(j)}\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}^1\boldsymbol{\beta},j)}\right\} \\ \times \prod_{k\in[p]\setminus\{i\}} \exp\left\{-\frac{1}{\sigma}|\beta_k| \hat{a}_k^1 v_{r_v(\hat{\mathbf{A}}^1\boldsymbol{\beta},k)}\right\} \\ \times \prod_{k\in[m]\setminus\{j\}} \exp\left\{-\frac{1}{\sigma}\|(\hat{\mathbf{A}}^1)^{(k)}\boldsymbol{\beta}^{(k)}\|_2 \sqrt{p_k} w_{r_g(\hat{\mathbf{A}}^1\boldsymbol{\beta},k)}\right\},\end{aligned}$$

where $\tilde{\boldsymbol{\kappa}}^{(j)} = \text{diag}(c_v^{\mathbf{1}(\delta_k=1)})$ for all $k \in G_j, k \neq i$, and the $i$th element is $c_v$. For an inactive variable, we have

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid \delta_i = 0, \boldsymbol{\delta}_{-i}, c_v, c_g, \sigma, \theta_v, \theta_g, \boldsymbol{\gamma}) = & c_g^{\sum_{j=1}^m p_j \mathbf{1}(\gamma_j=1)} c_v^{\sum_{k\in[p]\setminus\{i\}} \mathbf{1}(\delta_k=1)} \\ & \times \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g^{\mathbf{1}(\gamma_j=1)} v_{r_v(\hat{\mathbf{A}}^0\boldsymbol{\beta},i)}\right\} \\ & \times \prod_{k\in[p]\setminus\{i\}} \exp\left\{-\frac{1}{\sigma}|\beta_k| \hat{a}_k^0 v_{r_v(\hat{\mathbf{A}}^0\boldsymbol{\beta},k)}\right\} \\ & \times \prod_{j=1}^m \exp\left\{-\frac{1}{\sigma}\|(\hat{\mathbf{A}}^0)^{(j)}\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}^0\boldsymbol{\beta},j)}\right\}.\end{aligned}$$

A similar approximation as before is used to replace $\hat{\mathbf{A}}^0$ and $\hat{\mathbf{A}}^1$ with $\hat{\mathbf{A}}$. Then, considering the fraction in Equation E.3, the condition $\mathbf{1}(\gamma_j = 1)$ is removed as we are only sampling $\boldsymbol{\delta}$ in active groups, so that

$$\begin{aligned}\text{Numerator} = (1-\theta_v) c_g^{\sum_{j=1}^m p_j \mathbf{1}(\gamma_j=1)} c_v^{\sum_{k\in[p]\setminus\{i\}} \mathbf{1}(\delta_k=1)} \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g^{\mathbf{1}(\gamma_j=1)} v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)}\right\} \\ \times \prod_{k\in[p]\setminus\{i\}} \exp\left\{-\frac{1}{\sigma}|\beta_k| \hat{a}_k v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},k)}\right\} \\ \times \prod_{j=1}^m \exp\left\{-\frac{1}{\sigma}\|\hat{\mathbf{A}}^{(j)}\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta},j)}\right\},\end{aligned}$$

$$\begin{aligned}\text{Denominator} = \theta_v c_g^{p_j} c_v c_g^{\sum_{k\in[m]\setminus\{j\}} p_k \mathbf{1}(\gamma_k=1)} c_v^{\sum_{k\in[p]\setminus\{i\}} \mathbf{1}(\delta_k=1)} \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g c_v v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)}\right\} \\ \times \prod_{k\in[p]\setminus\{i\}} \exp\left\{-\frac{1}{\sigma}|\beta_k|\hat{a}_k v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},k)}\right\} \\ \times \prod_{k\in[m]\setminus\{j\}} \exp\left\{-\frac{1}{\sigma}\|\hat{\mathbf{A}}^{(k)}\boldsymbol{\beta}^{(k)}\|_2\sqrt{p_k} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta},k)}\right\} \\ \times \exp\left\{-\frac{1}{\sigma} c_g \|\tilde{\boldsymbol{\kappa}}^{(j)}\boldsymbol{\beta}^{(j)}\|_2\sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta},j)}\right\}.\end{aligned}$$

After cancellations, and combining the group terms in the denominator, we obtain

$$\begin{aligned}\text{Numerator} &= (1-\theta_v) c_g^{p_j \mathbf{1}(\gamma_j=1)} \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g^{\mathbf{1}(\gamma_j=1)} v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)}\right\}, \\ \text{Denominator} &= \theta_v c_g^{p_j} c_v \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g c_v v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)}\right\}.\end{aligned}$$

Therefore, for $i \in [p]$,

$$\begin{aligned}&\delta_i \sim \text{Bernoulli}\left(\frac{\tilde{L}_1}{\tilde{L}_1+\tilde{L}_2}\right), \text{ where,} \\ &\tilde{L}_1 = \theta_v c_v \exp\left\{-\frac{1}{\sigma}|\beta_i| c_g c_v \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)}\right\},\ \tilde{L}_2 = (1-\theta_v)\exp\left\{-\frac{1}{\sigma}|\beta_i| c_g \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)}\right\}.\end{aligned}$$

**Group Mixture Proportion ($\theta_g$)**

We have that

$$\theta_g \sim \pi(\theta \mid \boldsymbol{\gamma}, c_v, c_g, \mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma, \hat{\mathbf{A}}, \theta_v) = \pi(\theta_g \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma, \hat{\mathbf{A}}, \boldsymbol{\delta}) \propto \pi(\theta_g)\pi(\boldsymbol{\gamma} \mid \theta_g),$$

such that $\pi(\theta_g)$ is a Beta$(d_1, d_2)$ distribution. Now, as $\pi(\boldsymbol{\gamma} \mid \theta_g)$ follows a Bernoulli distribution, the posterior is given by a Beta distribution

$$\theta_g \sim \text{Beta}\left(d_1 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1), d_2 + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 0)\right).$$

**Variable Mixture Proportion ($\theta_v$)**

Due to the dependency between $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$, we have

$$\theta_v \sim \pi(\theta_v \mid \boldsymbol{\gamma}, c_v, c_g, \mathbf{y}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma, \hat{\mathbf{A}}, \theta_g) = \pi(\theta_v \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma, \hat{\mathbf{A}}, \boldsymbol{\delta}) \propto \pi(\theta_v)\pi(\boldsymbol{\delta} \mid \boldsymbol{\gamma}, \theta_v).$$

As $\pi(\theta_v)$ is a Beta$(e_1, e_2)$ distribution, the posterior is given by

$$\begin{aligned}\pi(\theta_v)\pi(\boldsymbol{\delta} \mid \boldsymbol{\gamma}, \theta_v) = \theta_v^{e_1-1}(1-\theta_v)^{e_2-1} \\ \times \prod_{j=1}^{m} \prod_{i \in G_j} \left[\theta_v^{\delta_i}(1-\theta_v)^{1-\delta_i}\mathbf{1}(\gamma_j = 1) + \mathbf{1}(\gamma_j = 0, \delta_i = 0)\right].\end{aligned}$$

Considering the two cases of whether the group is active

$$\begin{aligned}\text{Non-active:} \quad \pi(\theta_v)\pi(\boldsymbol{\delta} \mid \gamma_j = 0, \theta_v) &= \theta_v^{e_1-1}(1-\theta_v)^{e_2-1} \\ \text{Active:} \quad \pi(\theta_v)\pi(\boldsymbol{\delta} \mid \gamma_j = 1, \theta_v) &= \theta_v^{e_1-1}(1-\theta_v)^{e_2-1} \prod_{i \in G_j} \left[\theta_v^{\delta_i}(1-\theta_v)^{1-\delta_i}\right] \\ &= \theta_v^{e_1-1+\sum_{i \in G_j} \mathbf{1}(\delta_i=1)}(1-\theta_v)^{e_2-1+\sum_{i \in G_j} \mathbf{1}(\delta_i=0)}.\end{aligned}$$

Combining the two cases yields

$$\begin{aligned}&\pi(\theta_v)\pi(\boldsymbol{\delta} \mid \boldsymbol{\gamma}, \theta_v) \\ &= \prod_{j=1}^{m} \theta_v^{e_1-1+\mathbf{1}(\gamma_j=1)\sum_{i \in G_j} \mathbf{1}(\delta_i=1)}(1-\theta_v)^{e_2-1+\mathbf{1}(\gamma_j=1)\sum_{i \in G_j} \mathbf{1}(\delta_i=0)} \\ &= \theta_v^{me_1-m+\sum_{j=1}^{m} \mathbf{1}(\gamma_j=1)\sum_{i \in G_j} \mathbf{1}(\delta_i=1)}(1-\theta_v)^{me_2-m+\sum_{j=1}^{m} \mathbf{1}(\gamma_j=1)\sum_{i \in G_j} \mathbf{1}(\delta_i=0)}.\end{aligned}$$

This of the form of a Beta distribution, so that

$$\begin{aligned}\theta_v \sim \text{Beta}\Bigg(me_1 - m + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1) \sum_{i \in G_j} \mathbf{1}(\delta_i = 1) + 1, \\ me_2 - m + \sum_{j=1}^{m} \mathbf{1}(\gamma_j = 1) \sum_{i \in G_j} \mathbf{1}(\delta_i = 0) + 1\Bigg).\end{aligned}$$

**Group Signal Strength Ratio ($c_g$)**

We have

$$c_g \sim \pi(c_g \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{y}, \boldsymbol{\beta}, \sigma, \theta_v, \theta_g, c_v, \hat{\mathbf{A}}) = \pi(c_g \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma, c_v, \hat{\mathbf{A}}) \propto \pi(c_g)\pi(\boldsymbol{\beta} \mid c_g, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma, c_v, \hat{\mathbf{A}}).$$

The first term is uniform, so we can ignore it. To isolate on the components of the second term containing $c_g$ we need only consider the case $\gamma_j = 1$, as it is not possible to discriminate between $\boldsymbol{\delta} \in \{0, 1\}$ here, so that

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid c_g, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma, c_v, \hat{\mathbf{A}}) &= c_g^{\sum_{j=1}^m p_j \mathbf{1}\{\gamma_j = 1\}} \prod_{i=1}^p \exp\left\{\frac{-1}{\sigma}|\beta_i| c_g c_v \mathbf{1}(\delta_i = 1) v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\} \\ &\quad \times \prod_{j=1}^m \exp\left\{-\frac{c_g \mathbf{1}(\gamma_j = 1)}{\sigma}\|\boldsymbol{\kappa}^{(j)}\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)}\right\} \\ &= c_g^{\sum_{j=1}^m p_j \mathbf{1}(\gamma_j = 1)} \exp\left\{ - \frac{c_g}{\sigma}\left[\sum_{i=1}^p c_v |\beta_i| v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)} \mathbf{1}(\delta_i = 1) \right.\right. \\ &\quad \left.\left. + \sum_{j=1}^m \|\boldsymbol{\kappa}^{(j)}\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)} \mathbf{1}(\gamma_j = 1)\right]\right\}.\end{aligned}$$

This of the form of a Gamma distribution, so that, truncated at $[0, 1]$,

$$c_g \sim \text{Gamma}\Bigg(1 + \sum_{j=1}^m p_j \mathbf{1}(\gamma_j = 1), \\ \frac{1}{\sigma}\left[\sum_{i=1}^p c_v |\beta_i| \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)} \mathbf{1}(\delta_i = 1) + \sum_{j=1}^m \|\boldsymbol{\kappa}^{(j)}\boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j}(1 - \alpha) w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)} \mathbf{1}(\gamma_j = 1)\right]\Bigg).$$

**Variable Signal Strength Ratio ($c_v$)**

Here, we justify our choice of using MH to sample $c_v$. We would like to sample $c$ from

$$c_v \sim \pi(c_v \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{y}, \boldsymbol{\beta}, \sigma, \theta_v, \theta_g, c_g, \hat{\mathbf{A}}) = \pi(c_v \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma, c_g, \hat{\mathbf{A}}) \propto \pi(c_v)\pi(\boldsymbol{\beta} \mid c_v, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\beta}, \sigma, c_g, \hat{\mathbf{A}}).$$

The first term is uniform, so it vanishes. Therefore, considering only the terms containing $c_v$ in the second component, which occur when $\gamma_j = 1$ and $\delta_i = 1$,

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid c_v, \boldsymbol{\gamma}, \boldsymbol{\delta}, \sigma, c_g, \hat{\mathbf{A}}) &= c_v^{\sum_{i=1}^p \mathbf{1}\{\delta_i=1\}} \prod_{i=1}^p \exp\left\{\frac{-c_g}{\sigma}|\beta_i| c_v \mathbf{1}(\delta_i = 1) v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right\} \\ &\quad \times \prod_{j=1}^m \exp\left\{-\frac{c_g \mathbf{1}(\gamma_j = 1)}{\sigma} \|\boldsymbol{\kappa}^{(j)} \boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)}\right\} \\ &= c_v^{\sum_{i=1}^p \mathbf{1}(\delta_i=1)} \exp\left\{\sum_{i=1}^p \frac{-c_g}{\sigma} c_v |\beta_i| \mathbf{1}(\delta_i = 1) v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta}, i)}\right. \\ &\quad \left. + \sum_{j=1}^m \frac{-c_g \mathbf{1}(\gamma_j = 1)}{\sigma} \|\boldsymbol{\kappa}^{(j)} \boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta}, j)}\right\}. \end{aligned} \tag{E.4}$$

However, this is not of any clear distributional form. The $\boldsymbol{\kappa}$ term in the $\ell_2$ norm contains $c_v$, and we are unable to extract it from the norm. There are two potential solutions: 1. Approximate $\|\boldsymbol{\kappa}^{(j)} \boldsymbol{\beta}^{(j)}\|_2$ as some function of $c_v$, via an approximation such as an affine one. 2. Use Metropolis-Hastings to sample from the distribution. We opt for the latter option, using a Gamma$(2, 2)$ proposal, truncated at $[0, 1]$; this is a relatively uninformative prior that favours sparsity. This proposal is chosen as Equation E.4 would be of the form of a Gamma distribution if we approximated the $\ell_2$ term, and as $c_g$ is also sampled from a Gamma distribution.

### E.3.2 Stochastic and Maximisation Step Derivations

#### Update for $\sigma$

The solution is as for BGSLOPE (Appendix E.2), but using the SGS penalty instead:

$$\begin{aligned}\sigma_{[t]}^{\text{MLE}} = \frac{1}{2(n+2)} \Bigg[ &\sum_{i=1}^p |(\beta_i)_{[t]}| (\hat{a}_i)_{[t]} v_{r_v(\hat{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]}, i)} + \sum_{j=1}^m \|\hat{\mathbf{A}}_{[t]}^{(j)} \boldsymbol{\beta}_{[t]}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]}, j)} \\ &+ \Bigg( \left( \sum_{i=1}^p |(\beta_i)_{[t]}| (\hat{a}_i)_{[t]} v_{r_v(\hat{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]}, i)} + \sum_{j=1}^m \|\hat{\mathbf{A}}_{[t]}^{(j)} \boldsymbol{\beta}_{[t]}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]}, j)} \right)^2 \\ &+ 4(n+2)\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{[t]}\|_2^2 \Bigg)^{-1/2} \Bigg], \end{aligned}$$

or more simply

$$\sigma_{[t]}^{\text{MLE}} = \frac{K_2' + \sqrt{(K_2')^2 + 4K_1'(n+2)}}{2(n+2)}, \text{ where,}$$
$$K_1' = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{[t]}\|_2^2,$$
$$K_2' = \sum_{i=1}^{p} |(\beta_i)_{[t]}|(\hat{a}_i)_{[t]} v_{r_v(\hat{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]},i)} + \sum_{j=1}^{m} \|\hat{\mathbf{A}}_{[t]}^{(j)} \boldsymbol{\beta}_{[t]}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}_{[t]}\boldsymbol{\beta}_{[t]},j)}.$$

The classical MLE formula of $\sigma$ is also recovered here if the penalisation is removed.

### E.3.3 BSGS-$\alpha$

The updates follow those of BSGS, with an additional update; given $\alpha \sim \mathcal{U}[0,1]$,

$$\begin{aligned}
&\pi(\alpha \mid \boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{y}, \boldsymbol{\beta}, \sigma, \theta_g, \theta_v, c_g, c_v, \hat{\mathbf{A}}) \propto \pi(\alpha)\pi(\boldsymbol{\beta} \mid \gamma, \delta, \boldsymbol{\beta}, \sigma, c_v, c_g, \hat{\mathbf{A}}) \\
&\propto \prod_{j=1}^{m} c_g^{p_j \mathbf{1}\{\gamma_j = 1\}} c_v^{\sum_{i \in G_j} \mathbf{1}\{\delta_i = 1\}} \prod_{i=1}^{p} \exp\left\{ -\frac{1}{\sigma} |\beta_i| \hat{a}_i \alpha v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)} \right\} \\
&\quad \times \prod_{j=1}^{m} \exp\left\{ -\frac{1}{\sigma} \|\hat{\mathbf{A}}^{(j)} \boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} (1-\alpha) w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta},j)} \right\} \\
&\propto \exp\left\{ -\frac{\alpha}{\sigma} \left[ \sum_{i=1}^{p} |\beta_i| \hat{a}_i v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)} - \sum_{j=1}^{m} \|\hat{\mathbf{A}}^{(j)} \boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta},j)} \right] \right\}.
\end{aligned}$$

This posterior is in the form of an exponential distribution with rate

$$\lambda = \frac{1}{\sigma} \left[ \sum_{i=1}^{p} |\beta_i| \hat{a}_i v_{r_v(\hat{\mathbf{A}}\boldsymbol{\beta},i)} - \sum_{j=1}^{m} \|\hat{\mathbf{A}}^{(j)} \boldsymbol{\beta}^{(j)}\|_2 \sqrt{p_j} w_{r_g(\hat{\mathbf{A}}\boldsymbol{\beta},j)} \right].$$

However, the rate can be non-positive, so MH is used with a Beta$(10, 0.5)$ proposal.

### E.3.4 Algorithm

**Algorithm 9** SAEM for BSGS

**Input:** Initial values $(\boldsymbol{\beta}_{[0]}, \sigma_{[0]}, (c_g)_{[0]}, (c_v)_{[0]}, (\theta_g)_{[0]}, (\theta_v)_{[0]})$, maximum number of iterations $T > 0$, tolerance $\epsilon > 0$

**while** $t \leq T$ **and** $\|\boldsymbol{\beta}_{[t+1]} - \boldsymbol{\beta}_{[t]}\|_2^2 > \epsilon$ **do**

**Simulation step:**

Sample $(\gamma_j)_{[t]}$ from Equation 7.15 for $j = 1, \ldots, m$

Sample $(\delta_i)_{[t]}$ from Equation 7.18 for $i = 1, \ldots, p$

Sample $(\theta_g)_{[t]}$ from Equation 7.19

Sample $(\theta_v)_{[t]}$ from Equation 7.20

Sample $(c_g)_{[t]}$ from Equation 7.21

Sample $(c_v)_{[t]}$ from Equation 7.22

**Stochastic approximation step:**

Calculate $\boldsymbol{\beta}_{[t]}^{\text{MLE}}$ via Equation 7.23

Calculate $\sigma_{[t]}^{\text{MLE}}$ from Equation 7.24

Update parameters

$$\boldsymbol{\beta}_{[t+1]} = \boldsymbol{\beta}_{[t]} + \eta_t(\boldsymbol{\beta}_{[t]}^{\text{MLE}} - \boldsymbol{\beta}_{[t]}), \quad \sigma_{[t+1]} = \sigma_{[t]} + \eta_t(\sigma_{[t]}^{\text{MLE}} - \sigma_{[t]})$$

Update step size $\eta_t = \begin{cases} 1, & \text{if } t \leq 20, \\ \frac{1}{t-20}, & \text{if } t > 20 \end{cases}$

$t = t + 1$

**end while**

**Output:** $(\boldsymbol{\beta}_{[t+1]}, \boldsymbol{\delta}_{[t]})$, such that $\hat{\boldsymbol{\delta}}_{[t]} \leftarrow \frac{1}{20} \sum_{k=t-19}^{t} \boldsymbol{\delta}_{[k]}$ (if $t < 20$ all iterations of $\boldsymbol{\delta}$ are used), giving the fitted coefficients $\hat{\boldsymbol{\beta}}_{[t+1]} = \boldsymbol{\beta}_{[t+1]} \cdot \mathbf{1}\left(\hat{\boldsymbol{\delta}}_{[t]} > 0.5\right)$

## E.4 Illustrative Example

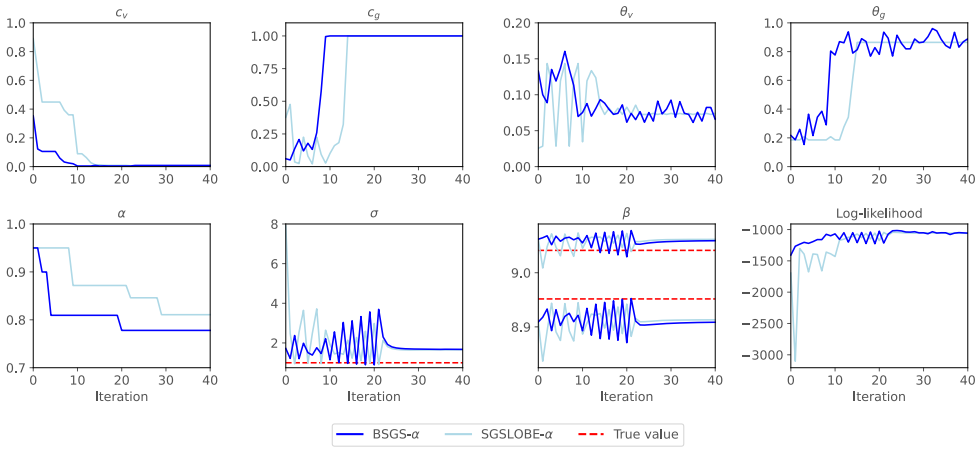



Figure E.1: Bayesian latent variables and model parameters for BSGS-$\alpha$ and SGSLOBE-$\alpha$ for the illustrative example (Section 7.4.5), shown for the first 40 iterations and the true values for $\sigma, \boldsymbol{\beta}$. Two active $\boldsymbol{\beta}$ values are shown. BSGS-$\alpha$ converged in 202 iterations and SGSLOBE-$\alpha$ in 350.

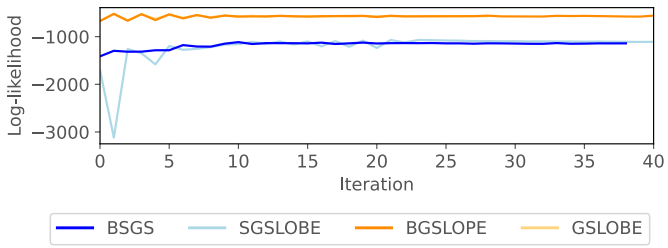



Figure E.2: The log-likelihood for BSGS, SGSLOBE, BGSLOPE, and GSLOBE for the illustrative example (Section 7.4.5).

## E.5 Initialisations and Stability Analysis

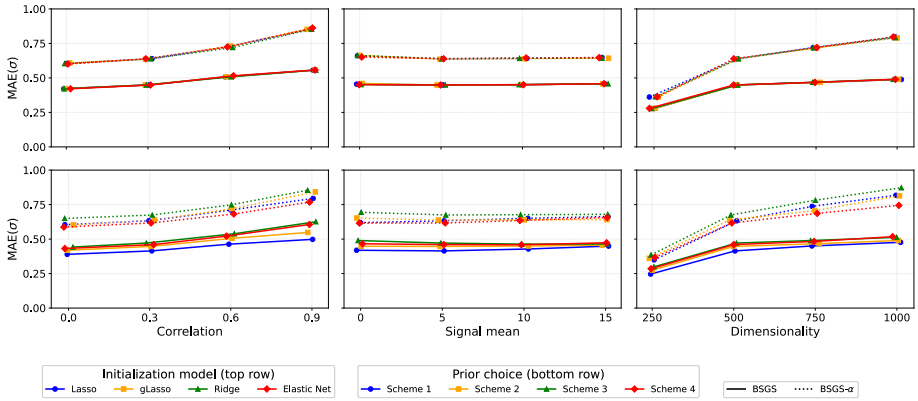



Figure E.3: MAE($\sigma$) for BSGS and BSGS-$\alpha$ under different $\boldsymbol{\beta}$ initialisations (top row) and Beta prior choices (bottom row), with added jitter.

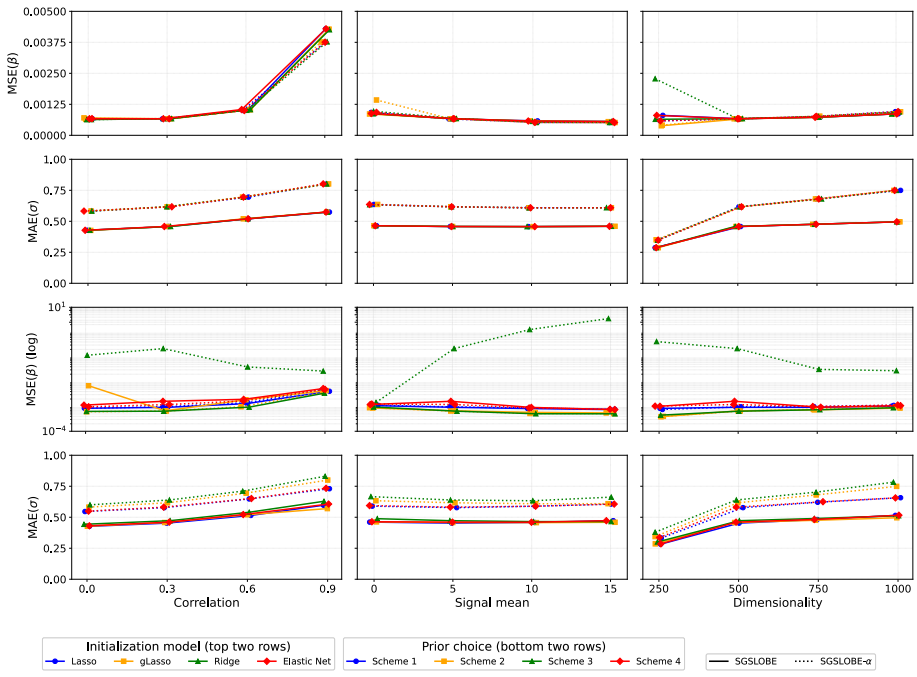



Figure E.4: MSE($\boldsymbol{\beta}$) and MAE($\sigma$) for SGSLOBE and SGSLOBE-$\alpha$ under different $\boldsymbol{\beta}$ initialisations (top rows) and Beta prior choices (bottom rows), with added jitter.

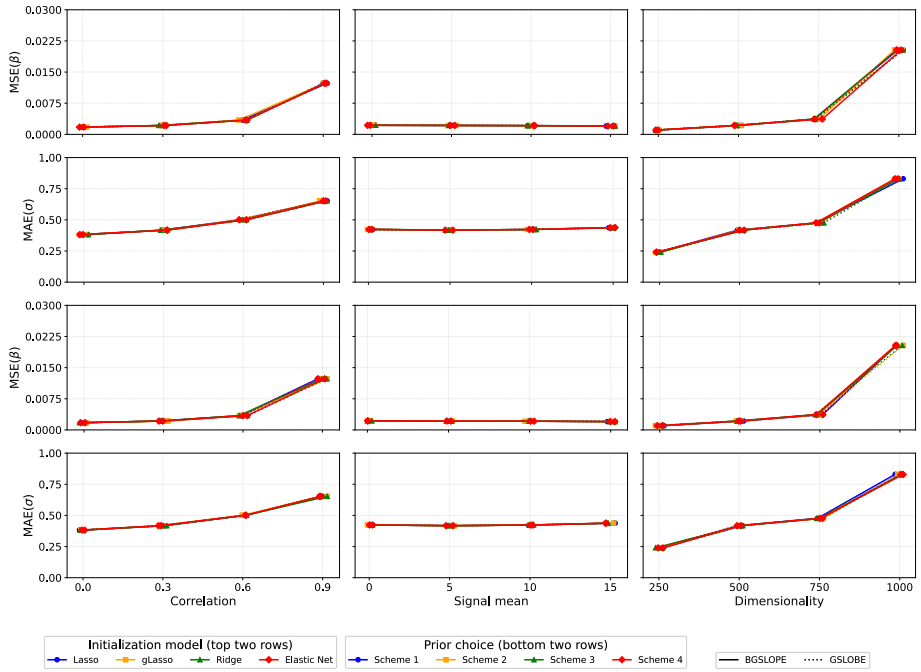



Figure E.5: MSE($\boldsymbol{\beta}$) and MAE($\sigma$) for BGSLOPE and GSLOBE under different $\boldsymbol{\beta}$ initialisations (top rows) and Beta prior choices (bottom rows), with added jitter.

| **Model** | **$\boldsymbol{\beta}$ initialisation** | | | | **Beta prior choice** | | | |
|---|---|---|---|---|---|---|---|---|
| | Lasso | gLasso | Ridge | Elastic net | Scheme 1 | Scheme 2 | Scheme 3 | Scheme 4 |
| BSGS | 82.6 (5.8) | **80.7** (5.6) | 81.6 (5.4) | 82.2 (5.8) | 191.4 (12.2) | 78.1 (5.6) | **73.6** (3.7) | 153.2 (10.1) |
| SGSLOBE | 79.2 (3.9) | 78.4 (4.1) | **76.4** (4.4) | 79.4 (3.8) | **71.3** (5.5) | 78.8 (4.4) | 76.8 (2.3) | **71.3** (5.3) |
| BSGS-$\alpha$ | 124.3 (15.2) | 123.0 (14.8) | **121.7** (15.7) | 123.8 (15.3) | 258.3 (17.8) | **120.2** (15.4) | 128.7 (16.2) | 257.1 (17.8) |
| SGSLOBE-$\alpha$ | **133.1** (2.1) | 133.7 (2.3) | 133.4 (2.0) | **133.1** (2.1) | 141.9 (8.2) | **133.1** (2.0) | 136.8 (3.4) | 141.4 (8.2) |
| BGSLOPE | 155.0 (10.0) | **154.3** (10.0) | 158.6 (9.7) | **154.3** (9.9) | 154.9 (10.0) | 155.5 (10.0) | 155.7 (10.0) | **145.1** (9.8) |
| GSLOBE | 56.4 (5.7) | 57.0 (6.2) | 60.0 (5.4) | **56.2** (5.7) | **56.5** (5.7) | **56.5** (5.7) | 56.7 (5.7) | **56.5** (5.7) |

Table E.1: Mean number of iterations for the $\boldsymbol{\beta}$ initialisations and Beta prior schemes, with standard errors shown in grey.

| Model | MSE($\beta$) | | | | MAE($\sigma$) | | | |
|---|---|---|---|---|---|---|---|---|
| | Lasso | gLasso | Ridge | Elastic net | Lasso | gLasso | Ridge | Elastic net |
| BSGS | **0.0010** (0.0001) | **0.0010** (0.0001) | 0.0018 (0.0043) | 0.0016 (0.0036) | 0.4527 (0.0066) | 0.4532 (0.0075) | **0.4524** (0.0064) | 0.4531 (0.0076) |
| SGSLOBE | **0.0010** (0.0002) | **0.0010** (0.0001) | **0.0010** (0.0001) | **0.0010** (0.0002) | **0.4613** (0.0066) | **0.4613** (0.0066) | 0.4615 (0.0065) | 0.4614 (0.0066) |
| BSGS-$\alpha$ | 0.1120 (0.1814) | **0.0503** (0.1240) | 0.1661 (0.2214) | 0.1020 (0.1773) | 0.6611 (0.0120) | **0.6593** (0.0123) | 0.6598 (0.0126) | 0.6605 (0.0125) |
| SGSLOBE-$\alpha$ | **0.0010** (0.0001) | **0.0010** (0.0003) | 0.0011 (0.0008) | **0.0010** (0.0001) | 0.6302 (0.0083) | 0.6306 (0.0083) | **0.6301** (0.0082) | 0.6302 (0.0083) |
| BGSLOPE | **0.0046** (0.0016) | **0.0046** (0.0016) | **0.0046** (0.0016) | **0.0046** (0.0016) | 0.4680 (0.0212) | 0.4681 (0.0212) | **0.4679** (0.0212) | 0.4680 (0.0212) |
| GSLOBE | **0.0046** (0.0016) | **0.0046** (0.0016) | **0.0046** (0.0016) | **0.0046** (0.0016) | 0.4681 (0.0212) | 0.4681 (0.0212) | **0.4680** (0.0212) | 0.4681 (0.0212) |

Table E.2: Mean MSE($\boldsymbol{\beta}$) and MAE($\sigma$) values for the initialisation schemes, with the best scheme for each model in **bold** and standard errors shown in grey.

| Model | MSE($\beta$) | | | | MAE($\sigma$) | | | |
|---|---|---|---|---|---|---|---|---|
| | Scheme 1 | Scheme 2 | Scheme 3 | Scheme 4 | Scheme 1 | Scheme 2 | Scheme 3 | Scheme 4 |
| BSGS | 0.0014 (0.0001) | 0.0010 (0.0001) | **0.0009** (0.0001) | 0.0015 (0.0004) | **0.4219** (0.0077) | 0.4500 (0.0063) | 0.4782 (0.0069) | 0.4686 (0.0077) |
| SGSLOBE | 0.0012 (0.0001) | 0.0016 (0.0029) | **0.0009** (0.0001) | 0.0016 (0.0004) | 0.4639 (0.0065) | **0.4605** (0.0066) | 0.4792 (0.0070) | 0.4683 (0.0080) |
| BSGS-$\alpha$ | **0.0010** (0.0001) | 0.0331 (0.1023) | 1.5796 (0.4257) | 0.0014 (0.0001) | 0.6537 (0.0111) | 0.6576 (0.0134) | 0.6971 (0.0143) | **0.6327** (0.0087) |
| SGSLOBE-$\alpha$ | 0.0013 (0.0001) | **0.0009** (0.0001) | 0.5055 (0.3309) | 0.0014 (0.0003) | **0.5877** (0.0074) | 0.6292 (0.0081) | 0.6572 (0.0128) | 0.5905 (0.0085) |
| BGSLOPE | **0.0046** (0.0016) | **0.0046** (0.0016) | **0.0046** (0.0016) | **0.0046** (0.0016) | **0.4681** (0.0212) | 0.4683 (0.0212) | 0.4684 (0.0212) | 0.4683 (0.0212) |
| GSLOBE | **0.0046** (0.0016) | **0.0046** (0.0016) | **0.0046** (0.0016) | **0.0046** (0.0016) | 0.4681 (0.0212) | 0.4681 (0.0212) | 0.4682 (0.0213) | **0.4679** (0.0212) |

Table E.3: Mean MSE($\boldsymbol{\beta}$) and MAE($\sigma$) values for the Beta prior schemes, with the best prior scheme for each model in **bold** and standard errors shown in grey.

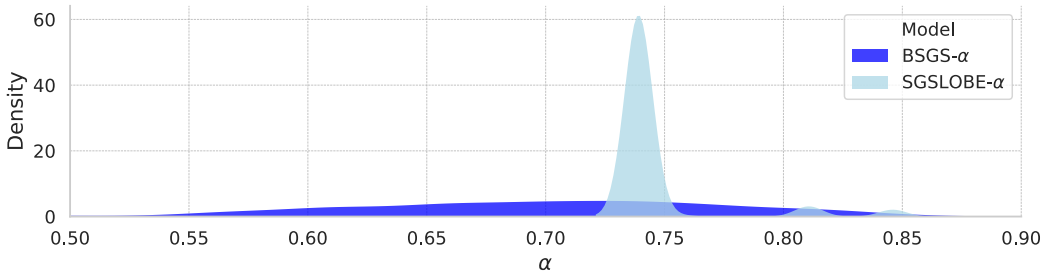



Figure E.6: Distribution of learned $\alpha$ values for BSGS-$\alpha$ and SGSLOBE-$\alpha$ for all simulations performed in Section 7.5.

## E.6 Other Approaches

### E.6.1 A Short Simulation Study

**Setup.** The response was generated using a linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{N}}_n(\mathbf{0}, \mathbf{I}_n)$ with design matrix $\mathbf{X} \sim \boldsymbol{\mathcal{N}}_p(\mathbf{0}, \boldsymbol{\Sigma}) \in \mathbb{R}^{100\times 500}$. Three correlation structures for $\boldsymbol{\Sigma}$ were considered:

1. $\boldsymbol{\Sigma} = \mathbf{I}_p$ (independent features).
2. $\Sigma_{i,i} = 1$, $\Sigma_{i,j} = 0.3$ for $i \neq j$ (moderate correlation).
3. $\Sigma_{i,i} = 1$, $\Sigma_{i,j} = 0.9$ for $i \neq j$ (strong correlation).

The signal was set to $\beta_i \sim \mathcal{N}(0, 10), i \in S_v$, with 0.95 variable sparsity proportion. The following methods are applied using a 50-length lasso path:

- *Stability Selection (SS)* [131] and *Complementary Pairs Stability Selection (CPSS)* [175]: target family-wise error rate of 0.1, $B = 100$ subsamples.
- *Knockoffs* [15]: Using the second-order Knockoffs construction.
- *FDR Stability Selection (FDR-SS)* [2]: $B = 15$ subsamples, $D = 5$ permuted datasets.
- *Kappa Selection (KS)* [190]: $B = 40$ subsamples, threshold of $\alpha_n = 0.1$.
- *ET-LASSO* [219].
- *CV*: includes the min and 1se versions.

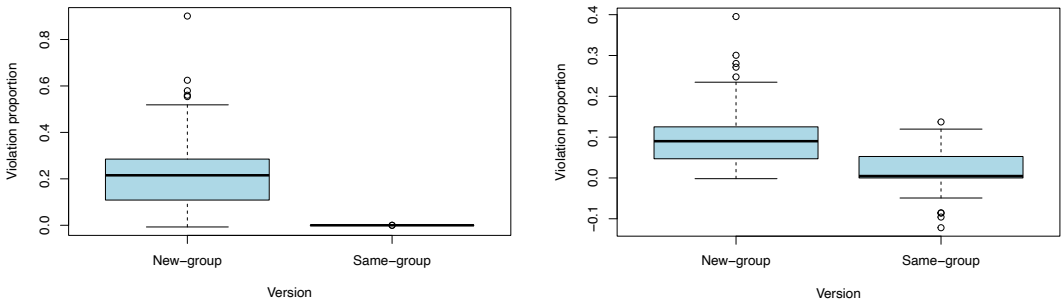
The lasso was used, instead of SLOPE, to ensure a fair comparison, as the majority of methods compared here were developed for the lasso.

**Results.** Table E.4 shows the FDR (with target level 0.1) and sensitivity for the approaches considered: Knockoff is found to have the lowest FDR levels for all cases.

| Case | Metric | CPSS | CV 1se | CV min | ET-LASSO | FDR-SS | Knockoffs | KS | SS |
|---|---|---|---|---|---|---|---|---|---|
| Case 1 | FDR | 0.25 | 0.54 | 0.75 | 0.28 | 0.75 | **0.23** | 0.87 | 0.25 |
| | Sensitivity | 0.17 | 0.24 | 0.40 | 0.17 | 0.31 | 0.15 | **0.54** | 0.18 |
| Case 2 | FDR | 0.32 | 0.62 | 0.76 | 0.58 | 0.80 | **0.30** | 0.88 | 0.33 |
| | Sensitivity | 0.16 | 0.22 | 0.35 | 0.18 | 0.31 | 0.10 | **0.50** | 0.16 |
| Case 3 | FDR | 0.77 | 0.81 | 0.85 | 0.82 | 0.91 | **0.71** | 0.92 | 0.79 |
| | Sensitivity | 0.05 | 0.06 | 0.11 | 0.06 | 0.19 | 0.02 | **0.28** | 0.05 |

Table E.4: FDR and sensitivity for model selection methods under three cases, with the best performance in each case highlighted in **bold**.

### E.6.2 Knockoffs




(a) Sparse-group lasso (SGL) (b) SGS

Figure E.7: The proportion of instances in which the coin flip property did not hold for signal variables, using a second-order construction, shown SGL and SGS. The results are split into whether the knockoffs are placed into a new group or the same group as the original variables. The Knockoffs filter was run twice for each signal variable: once with the original design, $[\mathbf{X}\ \mathbf{X}_{\text{KO}}]$, and once with the signal variable swapped with its knockoff copy.

## E.7 Synthetic Data Experiments

### E.7.1 Results

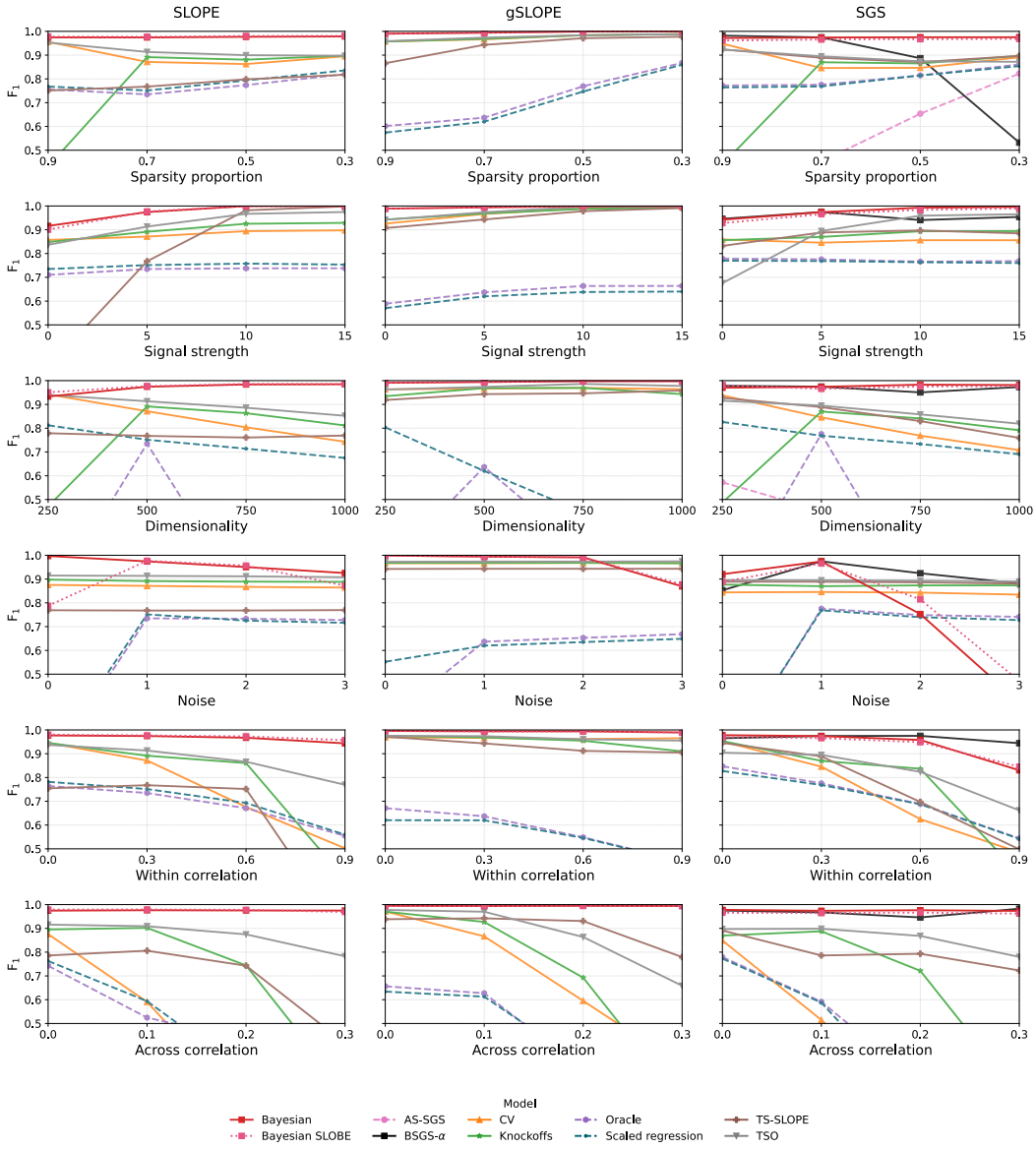



Figure E.8: $F_1$ scores for all model selection approaches, shown for all cases considered, split into the type of model (SLOPE, gSLOPE, SGS).

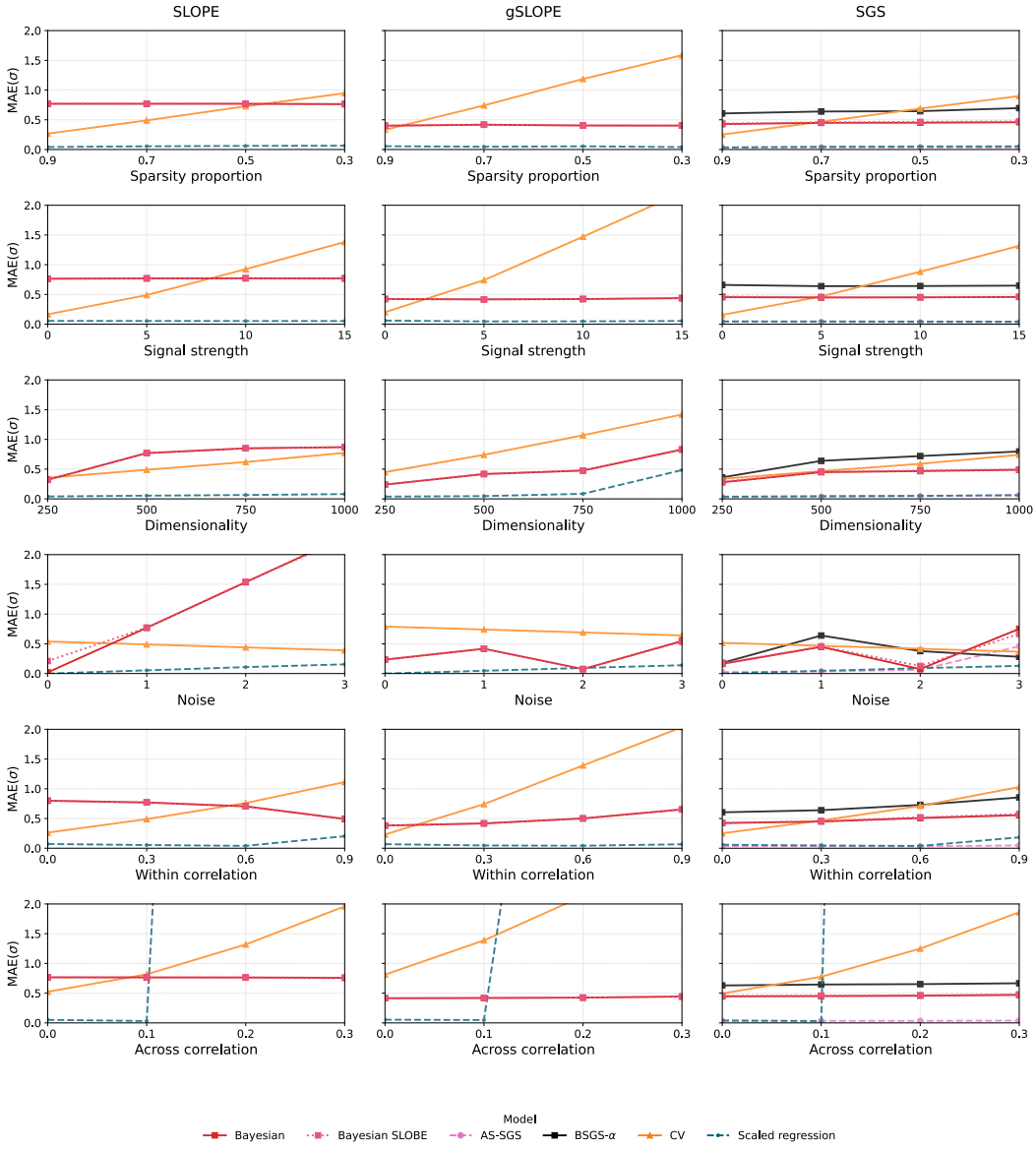



Figure E.9: MAE($\sigma$) for all model selection approaches that estimate the noise, shown for all cases considered, split into the type of model (SLOPE, gSLOPE, SGS).

### Impact of Signal

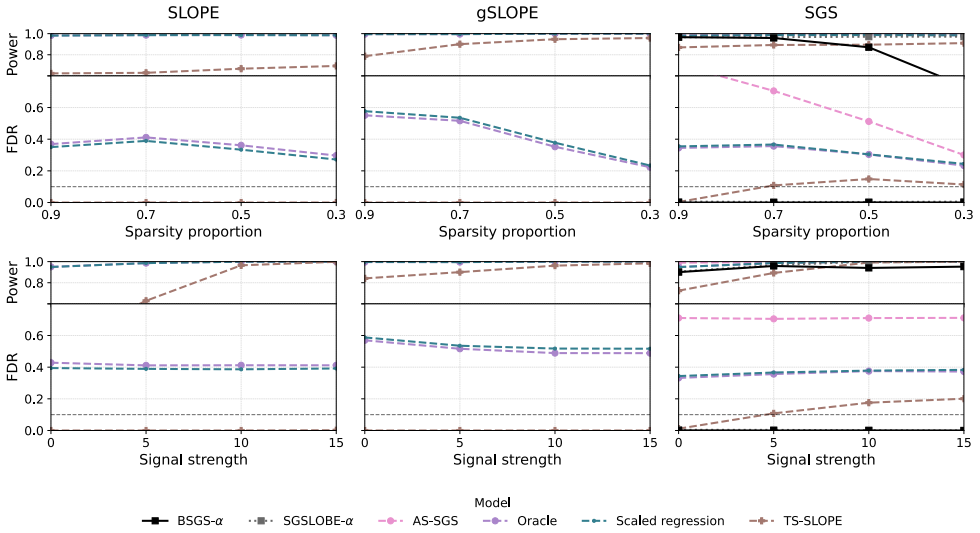



Figure E.10: FDR (bottom plots) and power (top plots) for the other model selection approaches, as functions of the sparsity proportion (top row) and signal strength (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

### Impact of Data-generating Parameters

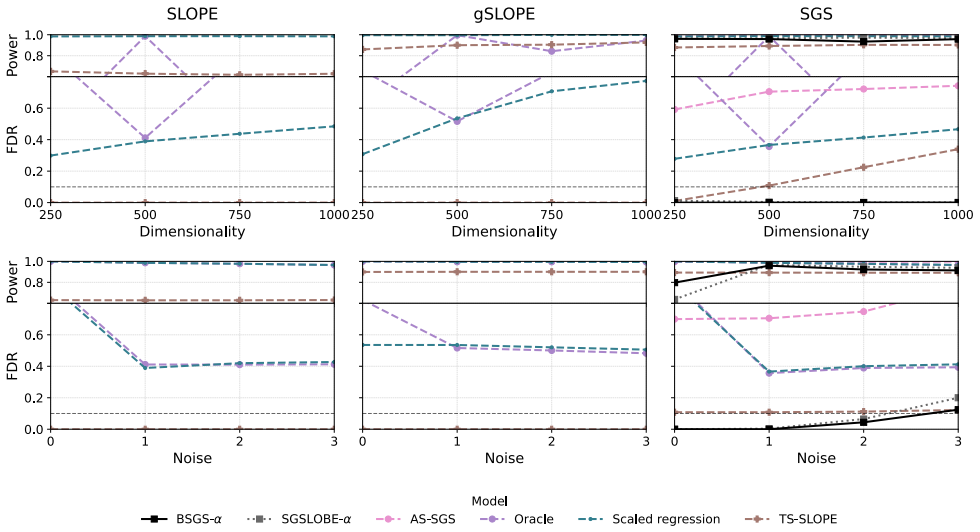



Figure E.11: FDR (bottom plots) and power (top plots) for the other model selection approaches, as functions of the dimensionality (top row) and noise (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

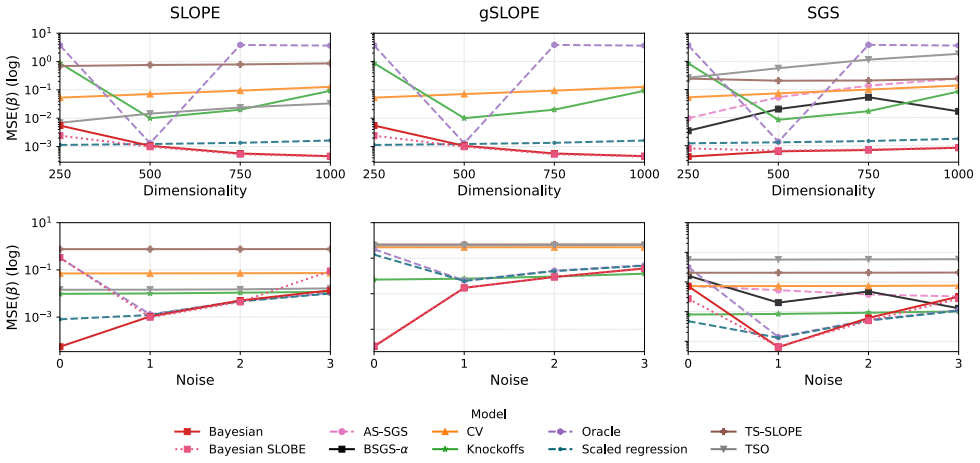



Figure E.12: MSE($\boldsymbol{\beta}$) (log scale) for all model selection approaches, as a function of the dimensionality (top row) and noise (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

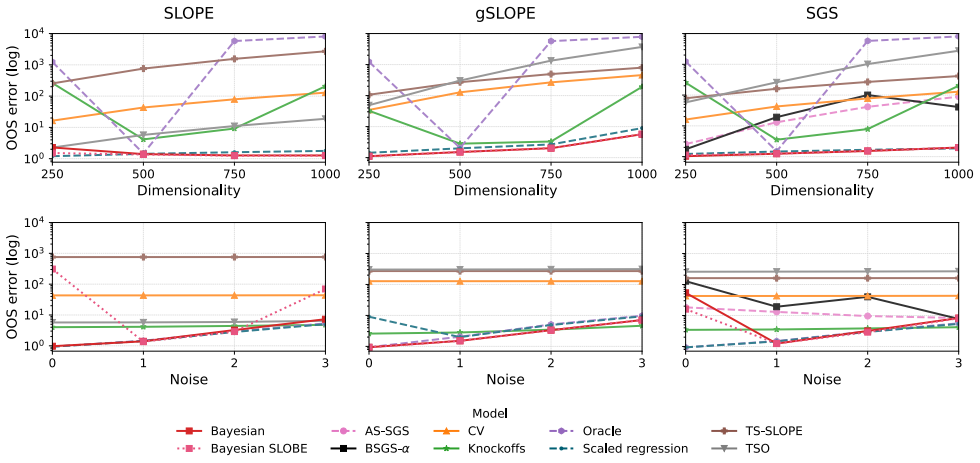



Figure E.13: OOS error (log scale) for all model selection approaches, as a function of the dimensionality (top row) and noise (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

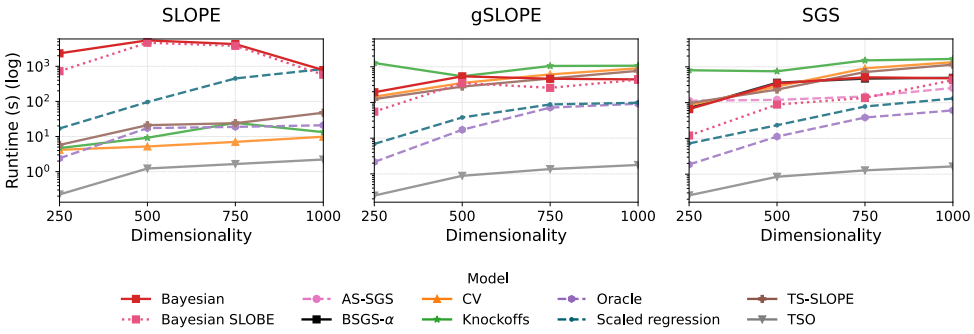



Figure E.14: Runtime in seconds (log scale) for all model selection approaches, as a function of the dimensionality, split into the type of model (SLOPE, gSLOPE, SGS).

## Correlation

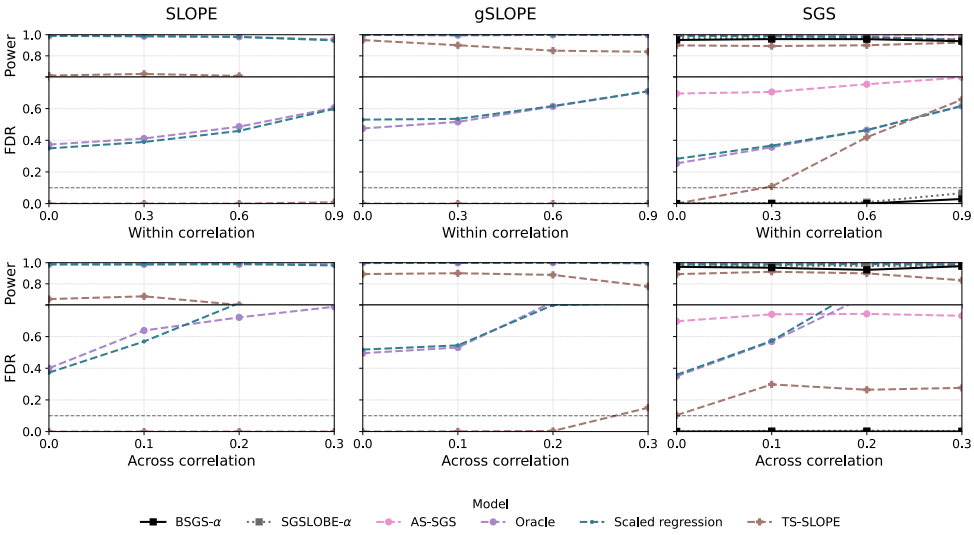



Figure E.15: FDR (bottom plots) and power (top plots) for the other model selection approaches, as functions of within correlation (top row) and across correlation (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

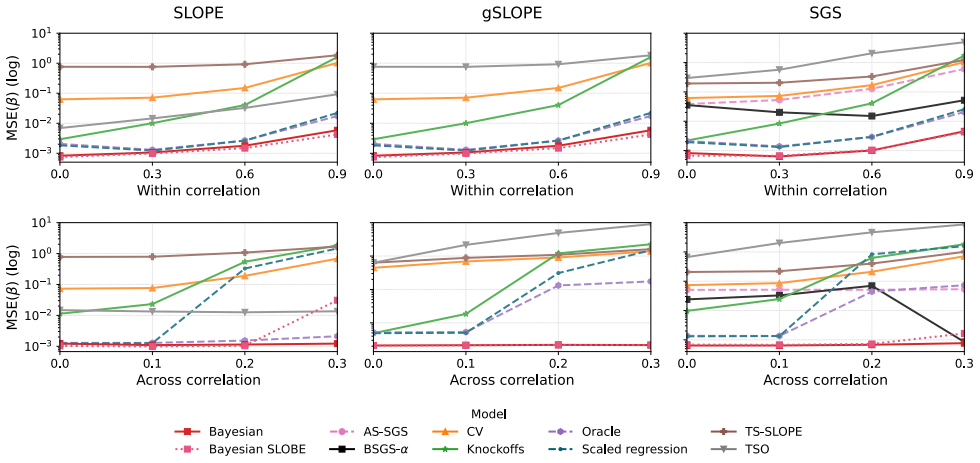



Figure E.16: MSE($\boldsymbol{\beta}$) (log scale) for all model selection approaches, as a function of within correlation (top row) and across correlation (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

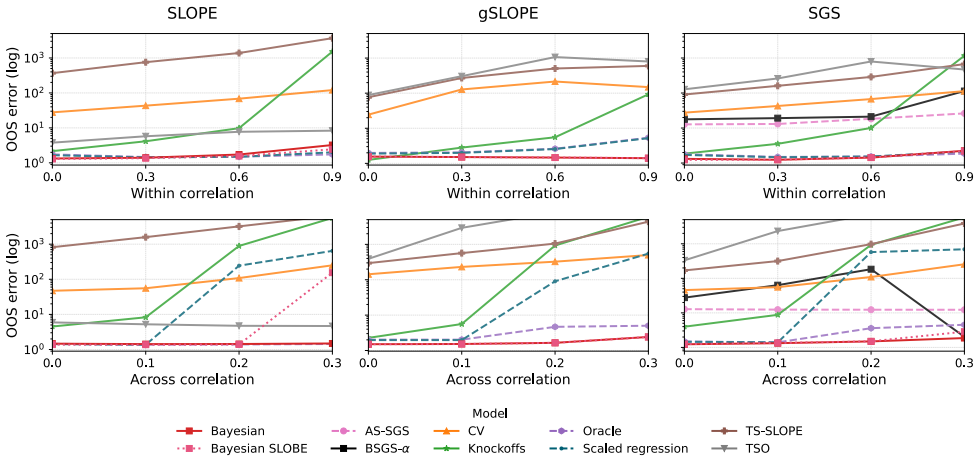



Figure E.17: OOS error (log scale) for all model selection approaches, as a function of within correlation (top row) and across correlation (bottom row), split into the type of model (SLOPE, gSLOPE, SGS).

### E.7.2 Summary

| Model | FDR group | Power group | $F_1$ group |
|---|---|---|---|
| AS-SGS | 0.17 (0.01) | **0.99** ($4\times10^{-3}$) | 0.88 (0.01) |
| BSGS | 0.14 (0.01) | **0.99** ($3\times10^{-3}$) | 0.90 (0.01) |
| SGSLOBE | 0.17 (0.01) | **0.99** ($3\times10^{-3}$) | 0.89 (0.01) |
| BSGS-$\alpha$ | **0.02** ($2\times10^{-3}$) | 0.96 (0.01) | 0.96 (0.01) |
| SGSLOBE-$\alpha$ | **0.03** ($4\times10^{-3}$) | **0.99** ($4\times10^{-3}$) | **0.97** ($4\times10^{-3}$) |
| CV | **0.09** ($3\times10^{-3}$) | 0.98 (0.01) | 0.91 ($4\times10^{-3}$) |
| Knockoffs | **0.02** ($3\times10^{-3}$) | 0.96 (0.01) | **0.97** (0.01) |
| Oracle | 0.46 (0.01) | 0.93 ($4\times10^{-3}$) | 0.65 (0.01) |
| Scaled | 0.42 (0.01) | **0.99** ($2\times10^{-3}$) | 0.71 (0.01) |
| TS-SLOPE | $\mathbf{7\times10^{-4}}$ ($4\times10^{-4}$) | 0.94 (0.01) | 0.96 (0.01) |
| TSO | **0.03** ($2\times10^{-3}$) | 0.95 (0.01) | 0.95 (0.01) |

Table E.5: Group metrics averaged across the six simulation cases considered for SGS, with standard errors shown in grey. The best performing model for each metric is highlighted in **bold** (aside from FDR, for which any that have FDR $\leq 0.1$ are in **bold**).

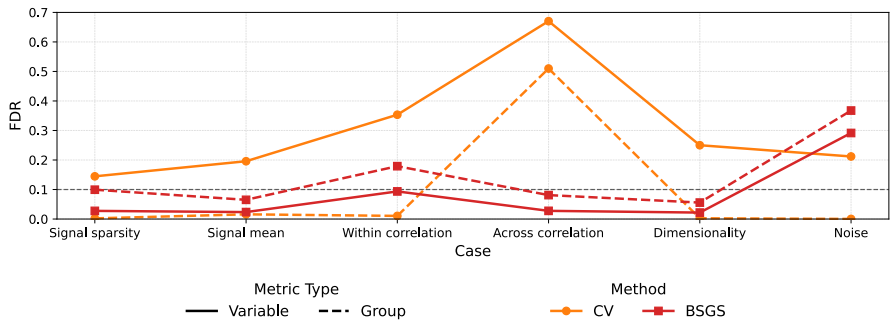



Figure E.18: Variable and group FDR shown for CV and BSGS for all synthetic cases considered.

### E.7.3 Additional Simulations

#### Equal Groups

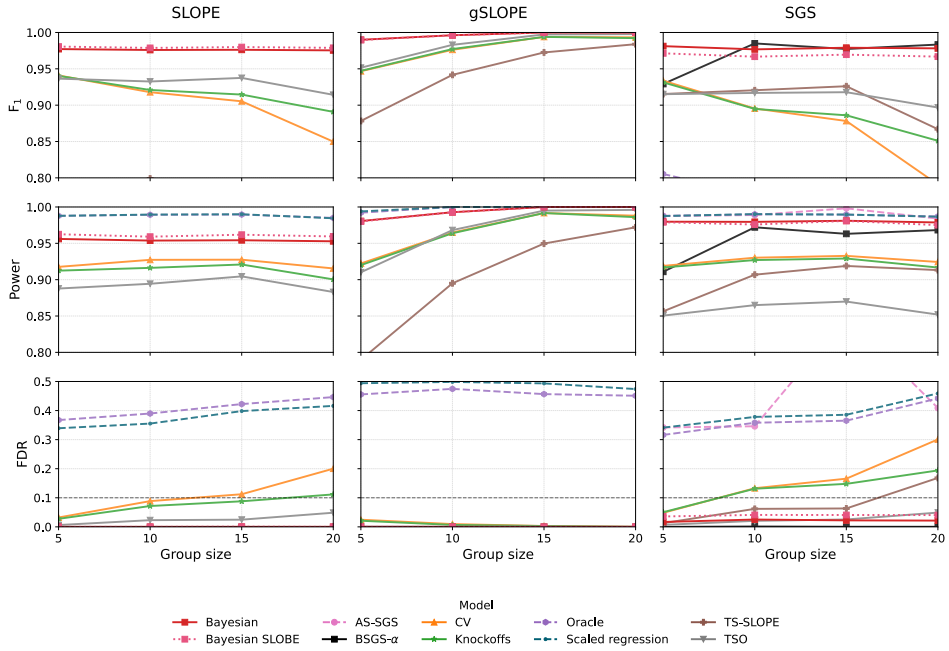



Figure E.19: $F_1$ score (top row), power (middle row), and FDR (bottom plots) for all model selection approaches as a function of group size for equal groups.

### E.7.4 Real Data Experiments

| Dataset | SLOPE | | gSLOPE | | SGS | |
|---|---|---|---|---|---|---|
| | ABSLOPE | SLOBE | BGSLOPE | BGSLOBE | BSGS | SGSLOBE |
| Cancer | 0.803 (0.111) | 1.149 (0.095) | 0.258 (0.031) | 0.403 (0.045) | 0.291 (0.039) | 0.291 (0.039) |
| Colitis | 0.419 (0.052) | 1.228 (0.112) | 0.759 (0.105) | 5.187 (0.735) | 0.401 (0.052) | 0.400 (0.052) |
| BRCA1 | 0.445 (0.043) | 1.028 (0.056) | 0.884 (0.094) | 1.028 (0.056) | 0.467 (0.043) | 0.445 (0.047) |
| Carbotax | 0.838 (0.175) | 0.999 (0.224) | 0.742 (0.156) | 0.999 (0.224) | 0.878 (0.195) | 0.844 (0.177) |
| Rhee | 0.138 (0.007) | 0.970 (0.037) | 0.140 (0.008) | 0.970 (0.037) | 0.138 (0.007) | 0.138 (0.007) |
| Scheetz | 0.442 (0.102) | 1.026 (0.388) | 0.479 (0.112) | 1.026 (0.388) | 0.715 (0.236) | 0.397 (0.079) |
| Trust-experts | 0.345 (0.009) | 0.346 (0.009) | 0.345 (0.009) | 0.345 (0.009) | 0.345 (0.009) | 0.345 (0.009) |

Table E.6: NMSE for the Bayesian methods and the SLOBE variants for each model type (SLOPE, gSLOPE, SGS), with standard errors shown in grey.