Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

GROUP COMBSS: GROUP SELECTION VIA CONTINUOUS OPTIMIZATION

Anant Mathur[1], Sarat Moka[1], Benoit Liquet[2,3], and Zdravko Botev[1]

[1]School of Mathematics and Statistics, University of New South Wales, NSW, AUSTRALIA
[2]School of Mathematical and Physical Sciences, Macquarie University, NSW, AUSTRALIA
[3]Laboratoire de Mathématiques et de leurs Applications, Université de Pau et des Pays de l'Adour, Pau, FRANCE.

ABSTRACT

We present a new optimization method for the group selection problem in linear regression. In this problem, predictors are assumed to have a natural group structure and the goal is to select a small set of groups that best fits the response. The incorporation of group structure in a design matrix is a key factor in obtaining better estimators and identifying associations between response and predictors. Such a discrete constrained problem is well-known to be hard, particularly in high-dimensional settings where the number of predictors is much larger than the number of observations. We propose to tackle this problem by framing the underlying discrete binary constrained problem into an unconstrained continuous optimization problem. The performance of our proposed approach is compared to state-of-the-art variable selection strategies on simulated data sets. We illustrate the effectiveness of our approach on a genetic dataset to identify grouping of markers across chromosomes.

1 INTRODUCTION

Given a dataset $(\boldsymbol{y}, \mathbf{X})$ consisting of a response vector $\boldsymbol{y} \in \mathbb{R}^n$ and a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ with n and p denoting the number of observations and the number of features respectively, the linear regression assumes that $\boldsymbol{y}$ and $\mathbf{X}$ have the linear relationship,

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ denotes the unknown regression coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$ represents a vector of unknown errors, unless otherwise specified, assumed to be independent and identically distributed.

The goal of group selection methods is to identify which groups of features are relevant for predicting the outcome variable and estimate the corresponding regression coefficients. This can help in situations where predictor variables naturally fall into meaningful groups or where there is prior knowledge suggesting that certain groups of variables may be related to the outcome variable. For instance, in genomics, genes belonging to the same pathway typically share similar functionalities and collaborate in regulating biological systems. The collective effect of these genes can be significant, making it feasible to detect them as a group, either at the pathway or gene set level. Incorporating this grouping structure has become increasingly common, largely due to the success of geneset enrichment analysis approaches (Subramanian et al. 2005). Incorporating group structure into regression analysis has proven effective for biomarker identification (Yuan and Lin 2006; Meier et al. 2008; Puig et al. 2009; Simon and Tibshirani 2012).

To formulate this problem, partition the design matrix $\mathbf{X}$ into distinct groups, denoted as $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J]$, where each $\mathbf{X}_j \in \mathbb{R}^{n \times p_j}$ represents j-th group with p_j features. Note that $p = p_1 + \cdots + p_J$. Then, (1) can be re-expressed as

$$\boldsymbol{y} = \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}, \tag{2}$$

979-8-3315-3420-2/24/$31.00 ©2024

where for each j, $\boldsymbol{\beta}_j \in \mathbb{R}^{p_j}$ is the regression coefficients associated with j-th group $\mathbf{X}_j$. To simplify our exposition, we do not include an intercept term in (2), assuming that the response variable $\boldsymbol{y}$ is centered.

The group selection problem then can be stated as a subset selection problem of the form

$$\underset{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_J}{\text{minimize}} \; \frac{1}{n}\|\boldsymbol{y} - \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j\|_2^2, \quad \text{subject to} \sum_{j=1}^{J} I\left(\|\boldsymbol{\beta}_j\|_2 > 0\right) \leq k. \tag{3}$$

where k is the sparsity parameter, $\|\cdot\|_2$ denotes $\mathscr{L}_2$-norm, and $I(\cdot)$ denotes the usual indicator function.

By incorporating group-wise structure into the regression model, group selection methods can improve model interpretability, reduce overfitting, and provide insights into the relationships between different groups of features and the outcome variable. Common approaches for group selection in linear regression include *group Lasso Regression* (Yuan and Lin 2006), a variant of the Lasso regression where the penalty term is applied at the group level rather than at the individual variable level thus encouraging sparsity at the group level, effectively selecting entire groups of features. An extension of group Lasso is *sparse group Lasso* (Simon et al. 2013) which allows for within-group sparsity, meaning not all features within a group are forced to be nonzero simultaneously. A third variant is *hierarchical variable selection*, which can be useful when the groups exhibit a hierarchical organization, such as in gene expression data or nested experimental designs. Relatively recent work (Hazimeh et al. 2023) proposes an efficient approximate algorithm for solving (3) based on a combination of coordinate descent and local search methods.

The paper is organized as follows. In Section 2, we state the group selection problem and formulate our continuous extension. In Section 3, we provide extensive numerical experiments comparing the proposed method with the most popular existing methods. In Section 4 we demonstrate our method using a complex genetic dataset where single nucleotide polymorphisms (SNPs) are utilized to predict gene expression across four distinct tissue types. Concluding remarks and possible future research directions are in Section 5.

2 GROUP SELECTION VIA COMBSS

The goal of this section is to show how the (non-group) model selection approach in (Moka et al. 2024) can be extended to the case of group selection. We call this method Group COMBSS (Continuous Optimization Method Towards Best Subset Selection). To this end, we first restate the exact group selection problem (3) as a binary constrained problem given by

$$\underset{s_1,\ldots,s_J \in \{0,1\}}{\text{minimize}} \; \underset{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_J}{\text{minimize}} \; \frac{1}{n}\|\boldsymbol{y} - \sum_{j=1}^{J} s_j \mathbf{X}_j \boldsymbol{\beta}_j\|_2^2, \quad \text{subject to} \sum_{j=1}^{J} s_j \leq k. \tag{4}$$

For each J-dimensional binary vector $\boldsymbol{s} = (s_1,\ldots,s_J) \in \{0,1\}^J$, let $\mathbf{X}_{[\boldsymbol{s}]}$ be matrix constructed from $\mathbf{X}$ by removing groups $\mathbf{X}_j$ that correspond to all $s_j = 0$. Thus, the number of columns of $\mathbf{X}_{[\boldsymbol{s}]}$ is equal to $\sum_{j=1}^{J} p_j I(s_j = 1)$. Similarly, let $\boldsymbol{\beta}_{[\boldsymbol{s}]}$ be the vector obtained from $\boldsymbol{\beta}$ by removing the elements of $\boldsymbol{\beta}$ indices that correspond all groups with $s_j = 0$. Then, (4) can be expressed as

$$\underset{\boldsymbol{s} \in \{0,1\}^J}{\text{minimize}} \; \underset{\boldsymbol{\beta}_{[\boldsymbol{s}]}}{\text{minimize}} \; \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{[\boldsymbol{s}]}\boldsymbol{\beta}_{[\boldsymbol{s}]}\|_2^2, \quad \text{subject to } |\boldsymbol{s}| \leq k, \tag{5}$$

where $|\boldsymbol{s}|$ denotes the number of 1's in $\boldsymbol{s}$. Now suppose, for a given $\boldsymbol{s}$, $\widehat{\boldsymbol{\beta}}_{[\boldsymbol{s}]}$ is a solution of

$$\underset{\boldsymbol{\beta}_{[\boldsymbol{s}]}}{\text{minimize}} \; \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{[\boldsymbol{s}]}\boldsymbol{\beta}_{[\boldsymbol{s}]}\|_2^2, \tag{6}$$

then (5) is equal to

$$\underset{\boldsymbol{s} \in \{0,1\}^J}{\text{minimize}} \; \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{[\boldsymbol{s}]}\widehat{\boldsymbol{\beta}}_{[\boldsymbol{s}]}\|_2^2, \quad \text{subject to } |\boldsymbol{s}| \leq k. \tag{7}$$

Solving (6) for $\widehat{\boldsymbol{\beta}}_{[\boldsymbol{s}]}$ is relatively easier task compared to solving (7). Indeed, the latter problem is well-known to be NP-hard (Natarajan 1995).

Now, for each $\boldsymbol{t} = [t_1, \ldots, t_J]^\top \in [0,1]^J$, let

$$\mathbf{T}_{\boldsymbol{t}} = \mathrm{Diag}\left([\underbrace{t_1, \ldots, t_1}_{p_1 \text{ times}}, \underbrace{t_2, \ldots, t_2}_{p_2 \text{ times}}, \ldots, \underbrace{t_J, \ldots, t_J}_{p_J \text{ times}}]^\top \right)$$

where $\mathrm{Diag}(\boldsymbol{u})$ is a diagonal matrix with diagonal being $\boldsymbol{u}$. Furthermore, take

$$\mathbf{X}_{\boldsymbol{t}} = \mathbf{X}\mathbf{T}_{\boldsymbol{t}},$$

and define

$$\mathbf{L}_{\boldsymbol{t}} = \frac{\mathbf{X}_{\boldsymbol{t}}^\top \mathbf{X}_{\boldsymbol{t}}}{n} + (\mathbf{I} - \mathbf{T}_{\boldsymbol{t}}^2). \tag{8}$$

One can view $\mathbf{L}_{\boldsymbol{t}}$ as a 'convex combination' of the matrices $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{I}/n$. The term $\mathbf{I} - \mathbf{T}_{\boldsymbol{t}}^2$ ensures that $\mathbf{L}_{\boldsymbol{t}}$ remains non-singular when zero-group exists and when all $t_i = 1$ there is zero impact on the matrix $\mathbf{X}^\top\mathbf{X}/n$. Let $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}$ be a solution of the linear equation (in $\boldsymbol{u}$)

$$\mathbf{L}_{\boldsymbol{t}}\boldsymbol{u} = \left(\frac{\mathbf{X}_{\boldsymbol{t}}^\top \boldsymbol{y}}{n} \right).$$

Then, we consider a Boolean relaxation of (7) given by

$$\underset{\boldsymbol{t} \in [0,1]^J}{\text{minimize}} \ \frac{1}{n} \| \boldsymbol{y} - \mathbf{X}_{\boldsymbol{t}} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}} \|_2^2, \quad \text{subject to} \sum_{j=1}^{J} t_j \le k. \tag{9}$$

We transform the discrete problem (7) into a continuous optimization (9) to take advantage of gradient evaluations. Generally, continuous optimization is acknowledged to be less challenging than combinatorial optimization. An example of this is Linear Programming (LP) vs Mixed-Integer Linear Programming (MILP), see (Fletcher 2000). The following result establishes some key properties of $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}$ and shows the relationship between (7) and (9).

Theorem 1 The following are true.

(i) $\mathbf{L}_{\boldsymbol{t}}$ is non-singular for all $\boldsymbol{t} \in (0,1)^J$.

(ii) For any corner point $\boldsymbol{s} \in \{0,1\}^J$, $\mathbf{X}_{[\boldsymbol{s}]}\widehat{\boldsymbol{\beta}}_{[\boldsymbol{s}]} = \mathbf{X}_{\boldsymbol{s}}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{s}}$.

(iii) For every sequence of vectors $\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, \cdots \in (0,1)^J$ that converges to a point $\boldsymbol{t} \in [0,1]^J$,

$$\| \boldsymbol{y} - \mathbf{X}_{\boldsymbol{t}} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}} \|_2 = \lim_{\ell \to \infty} \| \boldsymbol{y} - \mathbf{X}_{\boldsymbol{t}^{(\ell)}} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}^{(\ell)}} \|_2.$$

The proofs of (i), (ii) and (iii) are natural extensions of the proofs of Theorem 1, 2 and 3 in (Moka et al. 2024), and are thus omitted.

Theorem 1 (i) implies that for all interior points $t \in (0,1)^J$, $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}$ is unique and is given by $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}} = \mathbf{L}_{\boldsymbol{t}}^{-1}\mathbf{X}_{\boldsymbol{t}}^\top \boldsymbol{y}/n$, and (ii) implies that at the corners of the hypercube $[0,1]^J$, the value of objective function in (9) is identical to the value of the objective function in (7). Theorem 1 (iii) establishes the continuity of the objective function of the Boolean relaxation problem (9).

In this paper, instead of solving (9), we consider a relaxation using the Lagrangian form

$$f_\lambda(\boldsymbol{t}) = \frac{1}{n} \| \boldsymbol{y} - \mathbf{X}_{\boldsymbol{t}} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}} \|_2^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j} t_j,$$

for a tuning parameter $\lambda > 0$ and aim to solve

$$\underset{\boldsymbol{t}\in[0,1]^J}{\text{minimize}}\, f_\lambda(\boldsymbol{t}). \tag{10}$$

Instead of the sparsity parameter k, we now have the parameter λ to control the level of the sparsity in the solution. The $\sqrt{p_j}$ term is included to ensure the penalty term is scale-invariant with respect to the group size. The optimization (10) still has unwieldy box constraints. To get rid of these box constraints, we consider the equivalent unconstrained problem:

$$\underset{\boldsymbol{w}\in\mathbb{R}^J}{\text{minimize}}\, g_\lambda(\boldsymbol{w}), \tag{11}$$

where $g_\lambda(\boldsymbol{w}) = f_\lambda(\boldsymbol{t}(\boldsymbol{w}))$, $\boldsymbol{w} \in \mathbb{R}^J$, with $\boldsymbol{t}(\boldsymbol{w}) = 1/(1+\exp(-\boldsymbol{w}))$. That is, for each $i = 1,\ldots,J$, the j-th element t_j is obtained by applying the Sigmoid function on w_j. Since the Sigmoid function is strictly increasing, solving unconstrained problem (11) is equivalent to solving the box-constrained problem (10). We use the Adam optimizer, a popular gradient based approach, for solving (11). See Appendix A for a derivation of the gradient ∇g_λ of the objective function g_λ. Algorithm 1 provides a pseudo-code for the proposed method. It takes the data $(\boldsymbol{y},\mathbf{X})$, group sizes $(p_1,\ldots,p_J)$, penalty parameter λ, initial point $\boldsymbol{w}^{(0)}$, and threshold τ that helps convert the Sigmoid output into a binary one. For the given λ, $\mathsf{Adam}\left(\boldsymbol{w}^{(0)},\nabla g_\lambda\right)$ executes the Adam optimizer, which takes $\boldsymbol{w}^{(0)}$ as an initial point to provide a solution $\boldsymbol{w}$. This $\boldsymbol{w}$ is mapped to a point $\boldsymbol{t} \in [0,1]^J$ using the Sigmoid function and then $\boldsymbol{t}$ is mapped to a binary vector $\boldsymbol{s} \in \{0,1\}^J$ using the threshold parameter $\tau \in (0,1)$.

Algorithm 1 Group COMBSS

```
    Input: $(\boldsymbol{y},\mathbf{X})$, $(p_1,\ldots,p_J)$, $\lambda, \boldsymbol{w}^{(0)}, \tau$
$\boldsymbol{w} \leftarrow \mathsf{Adam}\left(\boldsymbol{w}^{(0)},\nabla g_\lambda\right)$
for $j = 1$ to $j = J$ do
    $t_j \leftarrow 1/(1+\exp(-w_j))$
    $s_j \leftarrow \mathbb{I}(t_j > \tau)$
end for
return $\mathbf{s} = (s_1,\ldots,s_J)^\top$
```

Remark 1 Recent work Hazimeh et al. (2023), Mazumder et al. (2023) suggests that when the signal-to-noise ratio (SNR) is low, additional ridge regularization can improve the prediction performance of the best subset selection. To include such additional ridge penalty in our implementation, we replace $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}$ with

$$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}^{\text{Ridge}} := \left[\mathbf{X}_{\boldsymbol{t}}^\top \mathbf{X}_{\boldsymbol{t}} + n(\mathbf{I} - \mathbf{T}_{\boldsymbol{t}}^2) + \gamma \mathbf{T}_t^2\right]^{-1} \mathbf{X}_{\boldsymbol{t}}^\top \boldsymbol{y}.$$

The parameter γ controls the strength of the ridge penalty. Note that when $\gamma > 0$ the estimator $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}^{\text{Ridge}}$ agrees with the simple ridge estimator at any corner point,

$$\widetilde{\boldsymbol{\beta}}_{\boldsymbol{s}}^{\text{Ridge}} = \left[\mathbf{X}_{[\boldsymbol{s}]}^\top \mathbf{X}_{[\boldsymbol{s}]} + \gamma \mathbf{I}\right]^{-1} \mathbf{X}_{[\boldsymbol{s}]}^\top \boldsymbol{y}, \quad \boldsymbol{s} \in \{0,1\}^J.$$

3 NUMERICAL SIMULATIONS

To compare the performance of a variate of group selection methods, we use datasets simulated from the model:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \quad \text{where} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n), \tag{12}$$

where we generate synthetic predictors $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J]$ with $\mathbf{X}_j \in \mathbb{R}^{n \times p_j}$. The design matrix $\mathbf{X}$ is simulated as a multivariate normal with a between-group correlation ψ and within-group correlation ρ. We run Group COMBSS Algorithm 1 and compare its statistical performance against the state-of-the-art grouped variable selection methods: L0 Group, Group Lasso, Group MCP and Group SCAD. We implement Group Lasso, Group MCP and Group SCAD with the R package `grpreg` (Breheny and Huang 2015). L0 Group is implemented with the Python software accompanying Hazimeh et al. (2023). To tune the parameter λ we generate an independent validation set from the generating process (12) with identical parameter values for ρ and ψ. We then minimize the generalization risk on the validation set over a grid with 100 values. We set the parameter τ to a default value of 10^{-1}. The coefficient $\boldsymbol{\beta}^*$ contains k nonzero groups and the nonzero entries of $\boldsymbol{\beta}^*$ are all set to 1.

After generating a training and validation set in each simulation, we run Group COMBSS to evaluate the λ that minimizes the generalization risk on the validation set. We denote this minimizer as λ^* and the corresponding model coefficient estimate as $\hat{\boldsymbol{\beta}}_{\lambda^*}$. The number of correct and incorrect non-zero groups in $\hat{\boldsymbol{\beta}}_{\lambda^*}$ are referred to as true positives (TP) and false positives (FP), respectively. Likewise, the number of correct and incorrect zero groups in $\hat{\boldsymbol{\beta}}_{\lambda^*}$ are referred to as true negatives (TN) and false negatives (FN), respectively. We consider the following performance measures:

1. **Precision**: Precision is defined as $TP/(TP+FP)$. A precision close to 1 indicates that the method is reliable in its classifications of non-zero groups while minimizing false positives.
2. **Recall**: Recall is defined as $TP/(TP+FN)$. A recall close to 1 indicates that the method is reliable in its classifications of non-zero groups while minimizing false negatives.
3. **Matthews correlation coefficient (MCC)**: MCC is defined as,

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.$$

   The MCC is a balanced measure that ranges from -1 (perfect disagreement) through 0 (no better than random chance) to $+1$ (perfect agreement).
4. **Generalization Risk**: This is defined as $\frac{1}{n}\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\lambda^*} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2$.

We consider the following simulation settings:

- Setting 1: $n = 100$, $p = 40$, $\rho = 0.9$, $\psi = 0.2$, $k = 4$ and $p_j = 4$.
- Setting 2: $n = 100$, $p = 40$, $\rho = 0.9$, $\psi = 0.5$, $k = 4$ and $p_j = 4$.
- Setting 3: $n = 400$, $p = 600$, $\rho = 0.9$, $\psi = 0.2$, $k = 15$ and $p_j = 4$.
- Setting 4: $n = 400$, $p = 600$, $\rho = 0.9$, $\psi = 0.5$, $k = 15$ and $p_j = 4$.

The value of σ^2 is chosen to achieve an SNR of either 1 or 3. For each simulation setting, we replicate the simulation 50 times and report the mean value of each performance measure. Standard errors of the mean are provided in parentheses.

In the low-dimensional and low-group-correlation setting (Setting 1, Table 1), Group COMBSS exhibits the highest MCC, Precision, and Recall scores closely followed by L0 Group. Group LASSO, MCP, and SCAD exhibit lower model risk as these are methods that not only select sparse models but also penalize regression coefficients. However, the performance of these three methods is inferior compared to Group COMBSS and L0 Group. In these simulations, the ridge penalty γ in Group COMBSS is set to zero, thereby excluding any penalization on the regression coefficients. In subsequent simulation efforts, we intend to explore the implications of a non-zero ridge penalty, chosen over a pre-defined grid. In the high-group correlation setting (Setting 2), we observe Group COMBSS achieving the best MCC score. When the signal is strong and group correlation is low, Group COMBSS and L0 group perfectly identify the non-zero groups in the low noise setting (Setting 1, Table 2).

Table 1: Low-dimensional, high noise (Setting 1 and 2).

Method	Setting	SNR = 1			
		MCC	Precision	Recall	Risk
Group COMBSS	1	0.95 (0.02)	0.98 (0.01)	0.95 (0.02)	17.87 (1.06)
L0 Group		0.91 (0.02)	0.97 (0.01)	0.92 (0.02)	20.41 (1.07)
Group LASSO		0.46 (0.03)	0.55 (0.01)	0.99 (0.01)	16.89 (0.76)
Group MCP		0.71 (0.03)	0.74 (0.02)	0.96 (0.01)	4.97 (0.20)
Group SCAD		0.59 (0.03)	0.65 (0.02)	0.98 (0.01)	4.85 (0.20)
Group COMBSS	2	0.74 (0.03)	0.94 (0.02)	0.74 (0.02)	28.62 (1.28)
L0 Group		0.67 (0.03)	0.92 (0.02)	0.66 (0.02)	32.13 (1.18)
Group LASSO		0.41 (0.03)	0.53 (0.01)	0.97 (0.01)	21.43 (0.94)
Group MCP		0.47 (0.03)	0.67 (0.03)	0.74 (0.03)	7.31 (0.23)
Group SCAD		0.49 (0.04)	0.62 (0.02)	0.91 (0.02)	6.67 (0.3)

Table 2: Low-dimensional, low noise (Setting 1 and 2).

Method	Setting	SNR = 3			
		MCC	Precision	Recall	Risk
Group COMBSS	1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	5.52 (0.27)
L0 Group		1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	5.49 (0.27)
Group LASSO		0.41 (0.03)	0.52 (0.01)	0.52 (0.01)	6.75 (0.31)
Group MCP		0.78 (0.02)	0.78 (0.02)	1.00 (0.00)	1.56 (0.07)
Group SCAD		0.57 (0.03)	0.63 (0.02)	1.00 (0.00)	1.61 (0.08)
Group COMBSS	2	0.97 (0.01)	0.98 (0.01)	0.98 (0.01)	9.26 (0.55)
L0 Group		0.91 (0.02)	0.95 (0.02)	0.96 (0.01)	10.65 (0.57)
Group LASSO		0.43 (0.03)	0.53 (0.01)	1.00 (0.00)	9.35 (0.44)
Group MCP		0.65 (0.03)	0.70 (0.02)	0.95 (0.02)	3.09 (0.14)
Group SCAD		0.55 (0.04)	0.62 (0.02)	0.98 (0.01)	3.09 (0.13)

In the high-dimensional regime, as shown in Tables 3 and 4, Group COMBSS achieves the best group selection among all methods, attaining a Precision score that is significantly closer to 1 in comparison to the Lasso, MCP, and SCAD, which tend to select a higher number of false positives. As discussed in Mazumder et al. (2023), it is observed that in cases of high noise (Table 3), the subset selection methods (Group COMBSS, L0 Group) yield higher generalization risk scores. Conversely, in the low-noise, low-group correlation scenario (Setting 3, Table 4), Group COMBSS exhibits the lowest generalization risk.

Table 3: High-dimensional, high noise (Setting 3 and 4).

		SNR = 1			
Method	Setting	MCC	Precision	Recall	Risk
Group COMBSS	3	0.64 (0.01)	0.80 (0.02)	0.56 (0.01)	194.04 (4.89)
L0 Group		0.56 (0.01)	0.83 (0.02)	0.42 (0.01)	238.72 (5.05)
Group LASSO		0.39 (0.01)	0.28 (0.01)	0.87 (0.01)	132.93 (2.82)
Group MCP		0.49 (0.01)	0.43 (0.01)	0.71 (0.02)	157.16 (3.37)
Group SCAD		0.41 (0.01)	0.29 (0.01)	0.86 (0.01)	135.59 (2.88)
Group COMBSS	4	0.30 (0.02)	0.51 (0.02)	0.23 (0.01)	313.71 (7.98)
L0 Group		0.25 (0.02)	0.50 (0.03)	0.17 (0.01)	373.94 (7.04)
Group LASSO		0.21 (0.01)	0.21 (0.01)	0.55 (0.02)	171.99 (3.95)
Group MCP		0.20 (0.01)	0.28 (0.01)	0.30 (0.01)	259.55 (5.99)
Group SCAD		0.21 (0.01)	0.21 (0.01)	0.53 (0.02)	173.23 (4.30)

Table 4: Low-dimensional, low noise (Setting 3 and 4).

		SNR = 3			
Method	Setting	MCC	Precision	Recall	Risk
Group COMBSS	3	0.94 (0.01)	0.95 (0.01)	0.94 (0.01)	55.70 (1.70)
L0 Group		0.88 (0.01)	0.96 (0.01)	0.84 (0.01)	69.27 (2.33)
Group LASSO		0.47 (0.01)	0.30 (0.00)	1.00 (0.00)	58.61 (1.34)
Group MCP		0.73 (0.01)	0.63 (0.01)	0.94 (0.01)	66.06 (2.05)
Group SCAD		0.54 (0.01)	0.36 (0.01)	0.99 (0.00)	64.01 (1.54)
Group COMBSS	4	0.57 (0.02)	0.74 (0.02)	0.49 (0.01)	139.57 (3.44)
L0 Group		0.50 (0.02)	0.77 (0.02)	0.37 (0.01)	172.26 (3.44)
Group LASSO		0.38 (0.01)	0.27 (0.01)	0.85 (0.01)	90.36 (1.95)
Group MCP		0.36 (0.01)	0.38 (0.01)	0.49 (0.01)	153.58 (3.01)
Group SCAD		0.41 (0.01)	0.30 (0.01)	0.83 (0.02)	97.67 (2.63)

4 ILLUSTRATION WITH GENETIC DATA

We demonstrate the application of our approach within the domain of genetic regulation. In expression Quantitative Trait Loci (eQTL) analysis, which aims to identify the genetic factors influencing gene expression variation (i.e., transcription), gene expression data are treated as a quantitative phenotype, while genotype data (SNPs) serve as predictors. In this study, we utilize a dataset extracted from a larger investigation (Heinig et al. 2010) focusing on the Hopx genes, as referenced in Petretto et al. (2010). This dataset has also been analyzed by Liquet et al. (2016), who employed a Bayesian model to identify a concise set of predictors explaining the collective variability of gene expression across four tissues: adrenal gland (ADR), fat, heart, and kidney. Liquet et al. (2017) utilize a sparse group Bayesian multivariate regression model for a similar objective. The Hopx dataset comprises 770 SNPs from 29 inbred rats forming the design matrix ($n = 29$, $p = 770$), with the expression levels measured in the four tissues (ADR, fat, heart, and kidney) serving as outcomes. A comprehensive description of the dataset is also available in Liquet and Chadeau-Hyam (2014) and can be accessed through the R package `R2GUESS`. Table 5 displays how

Table 5: Repartition of the SNPs along the chromosomes.

Chromosome	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Group size	74	67	63	60	39	45	52	43	31	51	21	26	33	22	15	27	18	30	34	19

the SNPs are distributed across the 20 chromosomes of the rats. The chromosome information establishes the grouping structure of the design matrix.

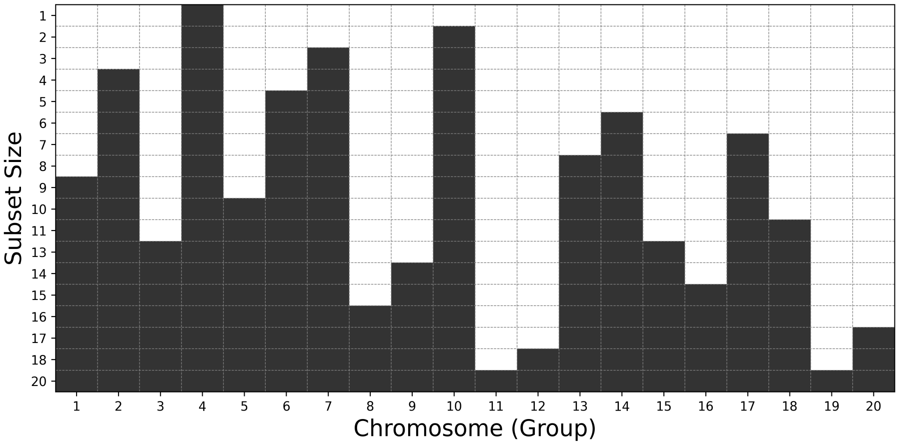

Figure 1: Best Subset Solution Path for ADR variable $(\gamma = 1)$.

We executed Group COMBSS on each tissue separately. The best subset solution path for each tissue has been obtained over a grid with 150 values of λ using Algorithm 1. Due to the high-dimensional aspect of the data $(n = 29, p = 770)$, we add a ridge penalization $(\gamma = 1)$. The solution path for the ADR tissue is presented in Figure 1. As an example, we analyse a parsimonious model with 4 groups, our model picked chromosomes 2, 4, 7 and 10. These chromosomes were also identified as being linked to the ADR tissue (Liquet et al. 2017), which utilized a sparse group Bayesian multivariate regression model. The results of our approach on the other tissues are presented in Figures 2, 3 and 4 in Appendix B. For the Kidney tissue, chromosomes 3, 4, 7 and 10 have been selected (for a model with four groups). Note that the ADR and Kidney outcomes are highly correlated $(r = 0.7)$, which may explain why three out of four groups have common chromosomes. For the Heart Tissue, chromosomes 2, 4, 14 and 15 have been selected (for a model with four groups) while chromosomes 1, 2, 4 and 15 have been selected when analysing the Fat tissue. Note also that the solution path using COMBSS for a partial least squares approach (Liquet et al. 2024), with a multivariate outcome (the four tissues) but without group selection, has identified a parsimonious set of SNPs located on chromosomes 4, 10, and 14. Chromosome 4 was selected in our four separate models, chromosome 10 was selected with the ADR and Kidney models and finally, chromosome 14 was selected when we analysed the heart tissue.

5 CONCLUSION AND FUTURE DIRECTIONS

In this paper, we presented an unconstrained continuous optimization algorithm for the group selection problem in linear regression. Our approach makes it possible to extend the non-group selection method in (Moka et al. 2024) to the group selection setting. We conducted extensive numerical simulations in both high- and low-dimensional settings to compare the performance of the proposed algorithm with the popular grouped variable selection approaches.

We have demonstrated our technique on a complex dataset comprising gene expression data (with four measurements from 29 samples) and SNP explanatory variables (consisting of 770 variables). The dataset exhibits a structured group organization (with 20 groups), delineated by chromosomes. Our current

Group COMBSS selection is not designed yet for analysing a multivariate response. To fully exploit the multivariate response, one can extend the univariate square error loss to accommodate the multivariate outcome. Furthermore, in genetics, it's a common practice to introduce an additional layer of sparsity within selected groups to improve interpretability. This involves identifying the relevant SNPs (variables) within the chosen groups.

Sparse group selection problem is an important generalization of the group selection problem, where in addition to the group selection, it is assumed that only a small number of features in each selected group are active. Similar to Friedman, Hastie, and Tibshirani (2010) that extends Group LASSO to the sparse group selection problem, our method can be extended to this problem. To see this, in addition to $\boldsymbol{t} \in [0,1]^J$, we consider $\boldsymbol{r} = [\boldsymbol{r}_1^\top, \ldots, \boldsymbol{r}_J^\top]^\top \in [0,1]^p$ with $\boldsymbol{r}_j = [r_{j,1}, \ldots, r_{j,p_j}]^\top \in [0,1]^{p_j}$. The vector $\boldsymbol{r}$ acts as binary relaxation for individual features. We can enforce group and within-group sparsity by incorporating appropriate penalties on $\boldsymbol{t}$ and $\boldsymbol{r}$.

In future research, we can also include a ridge penalty as explained in Remark 1 to enhance Group COMBSS's performance when minimizing generalization risk, particularly when SNR is low.

In Moka et al. (2024), an alternative version of COMBSS for linear regression is proposed for best subset selection, i.e., optimization (5) with the number of groups equal to the number of features ($J = p$). Future work can focus on numerical and theoretical study of the extension of this version of COMBSS to the group setting.

A DERIVATIVES OF THE OBJECTIVE FUNCTION

Our goal is to solve (11) using a gradient descent approach. To do that, we need to compute the gradient $\nabla_{\boldsymbol{w}} g_\lambda(\boldsymbol{w}) = (\partial g(\boldsymbol{w})/\partial w_1, \ldots, \partial g(\boldsymbol{w})/\partial w_J)$. With $\odot$ denoting the Hadamard (i.e., element-wise) product between two vectors, observe that

$$\nabla_{\boldsymbol{w}} g_\lambda(\boldsymbol{w}) = \nabla_{\boldsymbol{t}} f_\lambda(\boldsymbol{t}(\boldsymbol{w})) \odot \boldsymbol{t}(\boldsymbol{w})(1 - \boldsymbol{t}(\boldsymbol{w})),$$

where we used the fact that the derivative of the Sigmoid function $\boldsymbol{t}(\boldsymbol{w})$ is $\boldsymbol{t}(\boldsymbol{w})(1 - \boldsymbol{t}(\boldsymbol{w}))$.

Let $\mathbf{Z} = \mathbf{X}^\top \mathbf{X}/n - \mathbf{I}$, so that $\mathbf{L}_{\boldsymbol{t}} = \mathbf{T}_{\boldsymbol{t}} \mathbf{Z} \mathbf{T}_{\boldsymbol{t}} + \mathbf{I}$. Further, let $\mathbf{E}_j$ be a diagonal matrix of dimension p with zeros everywhere except ones along the diagonal at $(\sum_{k=1}^{j-1} p_k) + 1, \ldots, \sum_{k=1}^{j} p_k$. The following result establishes the derivatives $\partial \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}/\partial t_j$. Its proof is similar to the derivation of the gradient in (Moka et al. 2024) and hence ignored.

Lemma 1 Let $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}} = \mathbf{L}_{\boldsymbol{t}}^{-1}\left(\frac{\mathbf{X}_{\boldsymbol{t}}^\top \boldsymbol{y}}{n}\right)$. For any $\boldsymbol{t} \in (0,1)^J$, the derivatives of $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}$ are given by

$$\frac{\partial \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}}{\partial t_j} = \mathbf{L}_{\boldsymbol{t}}^{-1}\left[\mathbf{E}_j - \mathbf{E}_j \mathbf{Z} \mathbf{T}_{\boldsymbol{t}} \mathbf{L}_{\boldsymbol{t}}^{-1} \mathbf{T}_{\boldsymbol{t}} - \mathbf{T}_{\boldsymbol{t}} \mathbf{Z} \mathbf{E}_j \mathbf{L}_{\boldsymbol{t}}^{-1} \mathbf{T}_{\boldsymbol{t}}\right]\left(\frac{\mathbf{X}^\top \boldsymbol{y}}{n}\right), \quad j = 1, \ldots, J.$$

We shall use this Lemma to obtain $\nabla f_\lambda(\boldsymbol{t})$ for $\boldsymbol{t} \in (0,1)^J$. Let $\boldsymbol{\eta}_{\boldsymbol{t}} = \mathbf{T}_{\boldsymbol{t}} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}$. Then,

$$\|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{t}} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}\|_2^2 = \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\eta}_{\boldsymbol{t}}\|_2^2 = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\eta}_{\boldsymbol{t}}^\top \left(\mathbf{X}^\top \boldsymbol{y}\right) + \boldsymbol{\eta}_{\boldsymbol{t}}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\eta}_{\boldsymbol{t}}.$$

Now we focus on the j-th element of $\nabla_{\boldsymbol{t}} f_\lambda(\boldsymbol{t})$, that is,

$$\frac{\partial f_\lambda(\boldsymbol{t})}{\partial t_j} = \frac{\partial}{\partial t_j} \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{t}} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}\|_2^2 + \sqrt{p_j}\lambda.$$

Here,

$$\frac{\partial}{\partial t_j}\left[\frac{1}{n}\|\boldsymbol{y} - \mathbf{X}_{\boldsymbol{t}} \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}\|_2^2\right] = \frac{2}{n}\left(\frac{\partial \boldsymbol{\eta}_{\boldsymbol{t}}}{\partial t_j}\right)^\top \left[(\mathbf{X}^\top \mathbf{X})\boldsymbol{\eta}_{\boldsymbol{t}} - \mathbf{X}^\top \boldsymbol{y}\right] = 2\left(\frac{\partial \boldsymbol{\eta}_{\boldsymbol{t}}}{\partial t_j}\right)^\top \boldsymbol{a}_{\boldsymbol{t}},$$

where $\boldsymbol{a_t} = \left(\mathbf{X}^\top\mathbf{X}/n\right)\boldsymbol{\eta_t} - \left(\mathbf{X}^\top\boldsymbol{y}/n\right)$. From the definition of $\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}$ and $\boldsymbol{\eta_t}$,

$$\frac{\partial \boldsymbol{\eta_t}}{\partial t_j} = \frac{\partial \mathbf{T_t}}{\partial t_j}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}} + \mathbf{T_t}\frac{\partial \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}}{\partial t_j} = \mathbf{E}_j\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}} + \mathbf{T_t}\mathbf{L}_{\boldsymbol{t}}^{-1}\left[\mathbf{E}_j - \mathbf{E}_j\mathbf{Z}\mathbf{T_t}\mathbf{L}_{\boldsymbol{t}}^{-1}\mathbf{T_t} - \mathbf{T_t}\mathbf{Z}\mathbf{E}_j\mathbf{L}_{\boldsymbol{t}}^{-1}\mathbf{T_t}\right]\left(\frac{\mathbf{X}^\top\boldsymbol{y}}{n}\right).$$

Further simplification yields,

$$\begin{aligned}\frac{\partial \boldsymbol{\eta_t}}{\partial t_j} &= \mathbf{E}_j\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}} + \mathbf{T_t}\mathbf{L}_{\boldsymbol{t}}^{-1}\left[\mathbf{E}_j\left(\frac{\mathbf{X}^\top\boldsymbol{y}}{n}\right) - \mathbf{E}_j\mathbf{Z}\boldsymbol{\eta_t} - \mathbf{T_t}\mathbf{Z}\mathbf{E}_j\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}\right]\\ &= \mathbf{E}_j\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}} - \mathbf{T_t}\mathbf{L}_{\boldsymbol{t}}^{-1}\mathbf{E}_j\boldsymbol{b_t} - \mathbf{T_t}\mathbf{L}_{\boldsymbol{t}}^{-1}\mathbf{T_t}\mathbf{Z}\mathbf{E}_j\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}},\end{aligned}$$

where $\boldsymbol{b_t} = \mathbf{Z}\boldsymbol{\eta_t} - \left(\frac{\mathbf{X}^\top\boldsymbol{y}}{n}\right) = \boldsymbol{a_t} - \boldsymbol{\eta_t}$. To further simplify, let $\boldsymbol{c_t} = \mathbf{L}_{\boldsymbol{t}}^{-1}(\boldsymbol{t}\odot\boldsymbol{a_t})$, and $\boldsymbol{d_t} = \mathbf{Z}(\boldsymbol{t}\odot\boldsymbol{c_t})$. Then, the matrix $\frac{\partial\boldsymbol{\eta_t}}{\partial\boldsymbol{t}}$ of dimension $p\times J$ with the j-th column being $\frac{\partial\boldsymbol{\eta_t}}{\partial t_j}$ can be expressed as

$$\frac{\partial \boldsymbol{\eta_t}}{\partial \boldsymbol{t}} = \operatorname{BlkMat}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}) - \mathbf{T_t}\mathbf{L}_{\boldsymbol{t}}^{-1}\operatorname{BlkMat}(\boldsymbol{b_t}) - \mathbf{T_t}\mathbf{L}_{\boldsymbol{t}}^{-1}\mathbf{T_t}\mathbf{Z}\operatorname{BlkMat}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}}),$$

where for a p-dimensional vector $\boldsymbol{a_t} = [\boldsymbol{a}_{\boldsymbol{t},1}^\top,\dots,\boldsymbol{a}_{\boldsymbol{t},J}^\top]^\top$, the $p\times J$ matrix $\operatorname{BlkMat}(\boldsymbol{a_t})$ is defined as

$$\operatorname{BlkMat}(\boldsymbol{a_t}) := \begin{bmatrix}\boldsymbol{a}_{\boldsymbol{t},1} & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{0} & \boldsymbol{a}_{\boldsymbol{t},2} & & \vdots\\ \vdots & \mathbf{0} & \ddots & \vdots\\ \vdots & \vdots & & \vdots\\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{a}_{\boldsymbol{t},J}\end{bmatrix}.$$

Let $\boldsymbol{h} = [\sqrt{p_1},\dots,\sqrt{p_J}]^\top$. Then,

$$\begin{aligned}\nabla f_\lambda(\boldsymbol{t}) &= 2\left(\frac{\partial\boldsymbol{\eta_t}}{\partial t_j}\right)^\top\boldsymbol{a_t} + \lambda\boldsymbol{h}\\ &= 2\operatorname{BlkMat}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}})^\top\boldsymbol{a_t} - 2\operatorname{BlkMat}(\boldsymbol{b_t})^\top\mathbf{L}_{\boldsymbol{t}}^{-1}\mathbf{T_t}\boldsymbol{a_t} - 2\operatorname{BlkMat}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}})^\top\mathbf{Z}\mathbf{T}_{\boldsymbol{t}}^\top\mathbf{L}_{\boldsymbol{t}}^{-1}\mathbf{T_t}\boldsymbol{a_t} + \lambda\boldsymbol{h}\\ &= 2\operatorname{BlkMat}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}})^\top\boldsymbol{a_t} - 2\operatorname{BlkMat}(\boldsymbol{b_t})^\top\boldsymbol{c_t} - 2\operatorname{BlkMat}(\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t}})^\top\boldsymbol{d_t} + \lambda\boldsymbol{h}\\ &= 2\begin{bmatrix}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{t},1}^\top(\boldsymbol{a}_{\boldsymbol{t},1} - \boldsymbol{d}_{\boldsymbol{t},1}) - \boldsymbol{b}_{\boldsymbol{t},1}^\top\boldsymbol{c}_{\boldsymbol{t},1}\\ \vdots\\ \widetilde{\boldsymbol{\beta}}_{\boldsymbol{t},J}^\top(\boldsymbol{a}_{\boldsymbol{t},J} - \boldsymbol{d}_{\boldsymbol{t},J}) - \boldsymbol{b}_{\boldsymbol{t},J}^\top\boldsymbol{c}_{\boldsymbol{t},J}\end{bmatrix} + \lambda\boldsymbol{h}.\end{aligned}$$

B SUPPLEMENT MATERIAL: GENETIC DATA

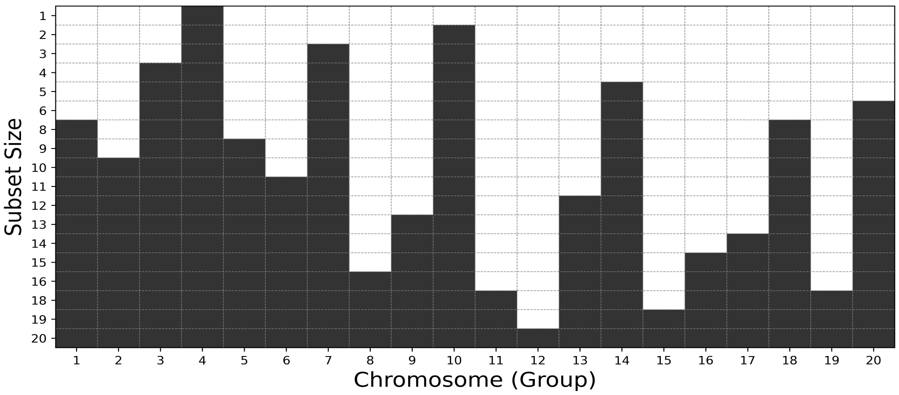

Figure 2: Best Subset Solution Path for Kidney variable $(\gamma = 1)$.

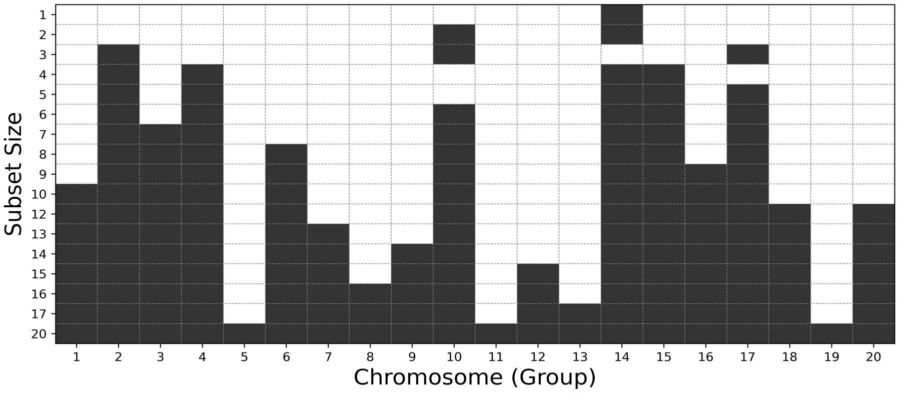

Figure 3: Best Subset Solution Path for Heart variable $(\gamma = 1)$.

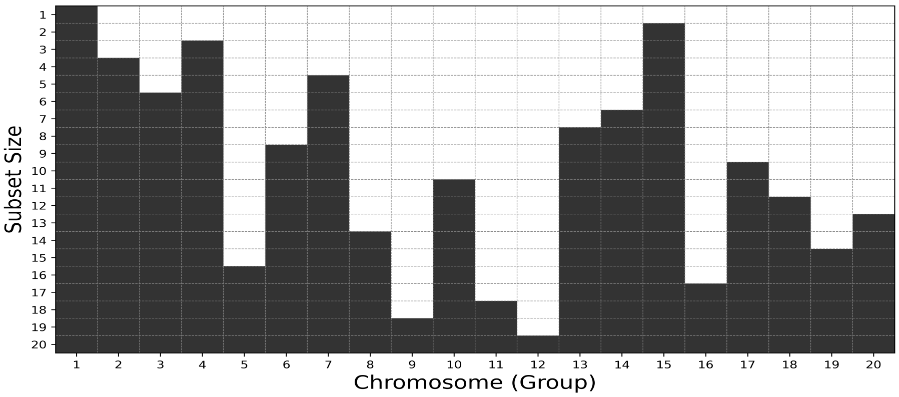

Figure 4: Best Subset Solution Path for Fat variable $(\gamma = 1)$.

REFERENCES

Breheny, P. and J. Huang. 2015. “Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors”. *Statistics and Computing* 25:173–187.

Fletcher, R. 2000. *Practical Methods of Optimization*. 2 ed. John Wiley & Sons.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. “A note on the group lasso and a sparse group lasso”. *arXiv preprint arXiv:1001.0736*.

Hazimeh, H., R. Mazumder, and P. Radchenko. 2023. “Grouped variable selection with discrete optimization: Computational and statistical perspectives”. *The Annals of Statistics* 51(1):1–32.

Heinig, M., E. Petretto, C. Wallace, L. Bottolo, M. Rotival, H. Lu, , , , *et al*. 2010. “A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk”. *Nature* 467(7314):460–464.

Liquet, B., L. Bottolo, G. Campanella, S. Richardson and M. Chadeau-Hyam. 2016. “R2GUESS: a graphics processing unit-based R package for Bayesian variable selection regression of multivariate responses”. *Journal of Statistical Software* 69(2).

Liquet, B. and M. Chadeau-Hyam. 2014. *R2GUESS: Wrapper Functions for GUESS*. R package version 1.4.

Liquet, B., K. Mengersen, A. Pettitt, and M. Sutton. 2017. “Bayesian variable selection regression of multivariate responses for group data”. *Bayesian Analysis* 12(4):1039–1067.

Liquet, B., S. Moka, and S. Muller. 2024. “Best Subset Solution Path for Linear Dimension Reduction Models using Continuous Optimization”. *arXiv preprint arXiv:2403.20007*.

Mazumder, R., P. Radchenko, and A. Dedieu. 2023. “Subset selection with shrinkage: Sparse linear modeling when the SNR is low”. *Operations Research* 71(1):129–147.

Meier, L., S. Van De Geer, and P. Bühlmann. 2008. “The group lasso for logistic regression”. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70(1):53–71.

Moka, S., B. Liquet, H. Zhu, and S. Muller. 2024. “COMBSS: best subset selection via continuous optimization”. *Statistics and Computing* 34(2):75.

Natarajan, B. K. 1995. “Sparse approximate solutions to linear systems”. *SIAM Journal on Computing* 24(2):227–234.

Petretto, E., L. Bottolo, S. R. Langley, M. Heinig, C. McDermott-Roe, R. Sarwar, , , , *et al*. 2010. “New insights into the genetic control of gene expression using a Bayesian multi-tissue approach”. *PLoS computational biology* 6(4):e1000737.

Puig, A. T., A. Wiesel, and A. O. Hero. 2009. “A multidimensional shrinkage-thresholding operator”. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, 113–116. IEEE.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani. 2013. “A sparse-group lasso”. *Journal of Computational and Graphical Statistics* 22(2):231–245.

Simon, N. and R. Tibshirani. 2012. “Standardization and the group lasso penalty”. *Statistica Sinica* 22(3):983.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, , , , *et al*. 2005. “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.

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

AUTHOR BIOGRAPHIES

ANANT MATHUR is a PhD student in the School of Mathematics and Statistics at UNSW Sydney. He received a Bachelor of Science (Hons) from UNSW Sydney. His current work focuses on feature selection algorithms and their efficient implementation. More broadly he is interested in problems in Data Science and Statistics. His email address is anant.mathur@unsw.edu.au.

SARAT MOKA is a Lecturer at the School of Mathematics and Statistics at The University of New South Wales. He received his PhD in Applied Probability from the School of Technology and Computer Science at Tata Institute of Fundamental Research, Mumbai. His research interests encompass applied probability, computational statistics, machine learning, and deep learning. His email address is s.moka@unsw.edu.au and his personal website is https://www.saratmoka.com.

BENOIT LIQUET is Professor of Mathematical and Computational Statistics at the School of Mathematical and Physical Sciences at Macquarie University. His research interests are model selection, machine learning, multi-state models, survival analysis, the multiple testing problem, dimension reduction methods and computational methods to analyse high dimensional and massive data sets. His email address is benoit.liquet-weiland@mq.edu.au

ZDRAVKO BOTEV is the inventor of the widely used diffusion or theta kernel density estimator, as well as the generalized splitting method for rare-event simulation. His research interests include fast algorithms for statistical learning and inference — he is the author of a number of books on Computational Statistics and Data Science. His email address is botev@unsw.edu.au.