Penalized Regression Methods for Interaction and Mixed-Effects Models with Applications to Genomic and Brain Imaging Data

Sahir Rai Bhatnagar

Doctor of Philosophy

Department of Epidemiology, Biostatistics and Occupational Health

McGill University
Montréal, Québec, Canada
July 2018

Chapter 1

Introduction

In this thesis, we consider the prediction of an outcome variable y observed on n individuals from p variables, where p is much larger than n. Challenges in this high-dimensional (HD) context include not only building a good predictor which will perform well in an independent dataset, but also being able to interpret the factors that contribute to the predictions. This latter issue can be very challenging in ultra-high dimensional predictor sets. For example, multiple different sets of covariates may provide equivalent measures of goodness of fit (Fan et al., 2014), and therefore how does one decide which are important?

When $p >> n$, standard generalized linear models (GLMs) methodology cannot uniquely estimate the unknown coefficient vector β. Moreover, even when $p \leq n$ with p close to n, standard errors of GLMs are likely to be inflated and parameter estimates unstable (Reid et al., 2016). In these instances it may be useful to assume the *Bet on Sparsity Principle* which says:

> Use a procedure that does well in sparse problems, since no procedure does well
>
> in dense problems (Friedman et al., 2001).

In fact, sometimes this assumption is necessary since we often do not have a large enough sample size to estimate so many parameters. Even when we do, we might want to identify a

relatively small number of predictors that play an important role on the response. Applied researchers in the health sciences might prefer a simpler model because it can shed light on the etiology of disease. The sparsity assumption may also result in faster computations and more stable predictions on new datasets.

With the advent of high-throughput technologies in genomics and brain imaging studies, computational approaches to variable selection have become increasingly important. Broadly speaking, there are three main approaches to analyze such HD data: 1) univariate regression followed by a filtering step often based on multiple testing corrections, and then subsequently a multivariable model using the reduced set of variables 2) multivariable penalized regression and 3) dimension reduction followed by a multivariable regression.

In this thesis, we focus on the use of penalized regression methods for variable selection and prediction in HD settings. In Chapter 2, we provide a critical review of the literature on penalized regression methods and the computational algorithms used to fit these models. In Chapter 3, we develop the sparse additive interaction learning model `sail` for detecting non-linear interactions with a key environmental or exposure variable in HD data. In Chapter 4, we develop a general penalized linear mixed model framework called `ggmix` that simultaneously selects and estimates variables while accounting for between individual correlations in one step. Chapter 5 explores whether use of exposure-dependent clustering relationships in dimension reduction can improve predictive modelling in a two-step framework that we develop called `eclust`. Chapters 3, 4 and 5 were originally written as stand-alone papers and as a result, there is some inconsistency in notation and overlap with the literature review chapter. Chapter 5 has been published in *Genetic Epidemiology.* Chapters 3 and 4 will be submitted for publication shortly after the submission of the thesis. We have published open source and freely available R packages for each of the methods developed Chapters 3, 4 and 5 (available at https://github.com/greenwoodlab). Table 1.1 provides an overview of our software packages including some of their key characteristics. In Chapter 6, we conclude

with an overview of the three manuscripts.

Table 1.1: Overview of our software packages for the three methods developped in the thesis. **Model** describes the type of loss function that has been developped in the thesis for a given method. **Penalty** are the penalty functions that can applied to the loss function. **Feature** describes the defining characteristics of the method. **Data** describes the inputs required for these methods, where x is the design matrix, y is the response, e is a single exposure variable and $\mathbf{\Psi}$ is an empirical covariance matrix. See Chapter 2 for more details.

	eclust	sail	ggmix
Model			
Least-Squares	✓	✓	✓
Binary Classification	✓		
Penalty			
Ridge	✓		✓
Lasso	✓	✓	✓
Elastic Net	✓		✓
Group Lasso		✓	✓
Feature			
Interactions	✓	✓	
Flexible Modeling	✓	✓	
Random Effects			✓
Data	(x, y, e)	(x, y, e)	$(x, y, \mathbf{\Psi})$

Chapter 2

Literature Review

The Literature Review is comprised of five sections. The first is a description of three general analytic strategies for high-dimensional data. The second and third sections describe two penalization methods that this thesis builds upon, namely the lasso and the group lasso. For each method we detail the algorithms used to fit these models and their convergence properties. In the fourth section we introduce penalized interaction models. This is followed by a brief introduction to linear mixed-effects models.

2.1 High-dimensional regression methods

We briefly introduce three of the main analytic strategies used to analyze HD data using the following notation. For n observations and p covariates, consider the multiple linear regression model

$$y = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

, where $\boldsymbol{y} \in \mathbb{R}^n$ is the response, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix, $\beta_0 \in \mathbb{R}$ is the intercept, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the coefficient vector corresponding to $\mathbf{X}$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of iid random errors.

2.1.1   Univariate regression

Genome-wide association studies (GWAS) have become the standard method for analyzing genetic datasets. A GWAS consists of a series of univariate regressions followed by a multiple testing correction. This approach is simple and easy to implement, and has successfully identified thousands of genetic variants associated with complex diseases (https://www.genome.gov/gwastudies/). Despite these impressive findings, the discovered markers have only been able to explain a small proportion of the phenotypic variance known as the missing heritability problem (Manolio et al., 2009). One plausible explanation is that there are many causal variants that each explain a small amount of variation with small effect sizes (J. Yang et al., 2010). GWAS are likely to miss these true associations due to the stringent significance thresholds required to reduce the number of false positives (Manolio et al., 2009). Most statistical methods for performing multiple testing adjustments assume weak dependence among the variables being tested (Leek & Storey, 2008). Dependence among multiple tests can lead to incorrect Type 1 error rates (Lin et al., 2013) and highly variable significance measures (Leek & Storey, 2008). Even in the presence of weakly dependent variables, adjusting for multiple tests in whole genome studies can result in low power. Furthermore, the univariate regression approach does not allow for modeling the joint effect of many variants which may be biologically more plausible (Schadt, 2009).

Univariate regression results may serve as a filtering step for subsequent multivariable models using the reduced set of predictors. Polygenic risk scores, which combine multiple genetic markers (based on a univariate filter) into a single score, have increased the explanatory power of a set of variables for predicting disease risk (Dudbridge, 2013). However, the final model may not be very good because filtering, i.e., assessing whether to keep a variable in the final model, is performed separately for each predictor. In the next section, we introduce multivariable penalized regression approaches which have been proposed to address some of these limitations.

2.1.2 Multivariable penalized regression

The least squares estimate for $\boldsymbol{\beta}$ in the linear model is given by $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{y}$. In high-dimensional data, the problem is that $\mathbf{X}^T\mathbf{X}$ is singular because the number of covariates greatly exceeds the number of subjects. For example DNA microarrays measure the expression of approximately 20,000 genes. However, due to funding constraints, the sample size is often less than a few hundred. A common solution to this problem is through penalized regression, i.e., applying a constraint on the values of $\boldsymbol{\beta}$. The problem can be formulated as finding the vector $\boldsymbol{\beta}$ that minimizes the penalized sum of squares:

$$\underbrace{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}X_{ij}\beta_j\right)^2}_{\text{goodness of fit}} + \underbrace{\sum_{j=1}^{p}p(\beta_j; \lambda, \gamma)}_{\text{penalty}} \tag{2.2}$$

The first term of (2.2) is the squared loss of the data and can be generalized to any loss function, while the second term is a penalty which depends on non-negative tuning parameters λ and γ that control the amount of shrinkage to be applied to $\boldsymbol{\beta}$ and the degree of concavity of the penalty function, respectively. Several penalty terms have been developed in the literature. Ridge regression places a bound on the square of the coefficients (ℓ_2 penalty) (Hoerl & Kennard, 1970) which has the effect of shrinking the magnitude of the coefficients. This however does not produce parsimonious models as none of the coefficients can be shrunk to exactly 0. The Lasso (Tibshirani, 1996) overcomes this problem by placing a bound on the sum of the absolute values of the coefficients (ℓ_1 penalty) which effectively shrinks some of them to 0, thereby simultaneously performing variable selection. The Lasso, along with other forms of penalization (e.g. SCAD Fan & Li (2001), Fused Lasso (Tibshirani et al., 2005), Adaptive Lasso (Zou, 2006), Relaxed Lasso (Meinshausen, 2007), MCP (Zhang, 2010)) have proven successful in many practical problems. Despite these encouraging results, such methods have low sensitivity for identifying associated variables in the presence of high empirical correlations between covariates, because only one variable tends to be se-

lected from the group of correlated or nearly linearly dependent variables (Bühlmann et al., 2013). As a consequence, there is rarely consistency on which variables are chosen from one dataset to another (e.g. in cross-validation folds). This behavior may not be desirable in high-dimensional datasets in which large sets of predictors are highly correlated and are also associated with the response since it may limit interpretability of results. The elastic net was proposed to benefit from the strengths of ridge regression's treatment of correlated variables and lasso's sparsity (Zou & Hastie, 2005). By placing both an ℓ_1 and ℓ_2 penalty on $\boldsymbol{\beta}$, the elastic net achieves model parsimony while yielding similar regression coefficients for correlated variables. These methods however do not take advantage of any (possibly known) grouping structure in the data. For example, cortical thickness measurements from magnetic resonance imaging (MRI) scans are often grouped into cortical regions of the Automated Anatomical Labelling (AAL) atlas (Tzourio-Mazoyer et al., 2002). Genes involved in the same cellular process (e.g. KEGG pathway (Kanehisa et al., 2008)) can also be placed into biologically meaningful groups. When regularizing with the ℓ_1 penalty, each variable is selected individually regardless of its position in the design matrix. Existing structures between the variables (e.g. spatial, networks, pathways) are ignored even though in many real-life applications the estimation can benefit from this prior knowledge in terms of both prediction accuracy and interpretability (Bach et al., 2012). The group lasso (Yuan & Lin, 2006) (and generalizations thereof) overcomes this problem by producing a structured sparsity (Bach et al., 2012), i.e., given a predetermined grouping of non-overlapping variables, all members of the group are either zero or non-zero. The main drawback when applying these methods is that these groups may not be known *a priori*. Even in genomics, known pathways may not be relevant to the response of interest, and the accuracy of inferring gene networks is still in its infancy.

2.1.3 Dimension reduction together with regression

Due to the unknown grouping problem, several authors have suggested a two-step procedure where they first cluster or group variables in the design matrix and then subsequently proceed to model fitting where the feature space is some summary measure of each group. This idea dates back to 1957 when Kendall (Kendall, 1957) first proposed using principal components in regression. Hierarchical clustering based on the correlation of the design matrix has also been used to create groups of genes in microarray studies and for each level of hierarchy, the cluster average was used as the new set of potential predictors in forward-backward selection (Hastie et al., 2001) or the lasso (Park et al., 2007). Bühlmann et al. (2013) proposed a bottom-up agglomerative clustering algorithm based on canonical correlations and used the group lasso on the derived clusters. There are some advantages to these methods over the ones previously mentioned in Sections 2.1.1 and 2.1.2. First, the results may be more interpretable than the traditional lasso (and related methods) because the non-zero components of the prediction model represent sets of genes as opposed to individual ones. Second, by using genes which cluster well, we bias the inputs towards correlated sets of genes which are more likely to have similar function. Third, taking a summary measure of the resulting clusters can reduce the variance in prediction (overfitting) due to the compressed dimension of the feature space. Lastly, from a practical point of view this approach is flexible and easy to implement because efficient algorithms exist for both clustering (Müllner, 2013) and model fitting (Friedman et al., 2010; Y. Yang & Zou, 2014).

In this context, we introduce a new two-step procedure called `eclust` (Bhatnagar et al., 2018) in Chapter 5 of the thesis. Our method is motivated by the fact that exposure variables (e.g. smoking) can alter correlation patterns between clusters of high-dimensional variables, i.e., alter network properties of the variables. However, it is not well understood whether such altered clustering is informative in prediction. In this paper, we explore whether use of exposure-dependent clustering relationships in dimension reduction can improve predictive

modelling in a two-step framework.

Now that these three general strategies for predictive modelling with $p >> n$ have been described, the next section describes several penalized methods in detail.

2.2   Lasso

For least-squares loss, the lasso estimator (Tibshirani, 1996; Zou, 2006) is defined as

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{(\beta_0, \boldsymbol{\beta})}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} w_i (y_i - \beta_0 - (\mathbf{X}\boldsymbol{\beta})_i)^2 + \lambda \sum_{j=1}^{p} v_j |\beta_j| \tag{2.3}$$

where $(\mathbf{X}\boldsymbol{\beta})_i$ is the ith element of the n-length vector $\mathbf{X}\boldsymbol{\beta}$, $\lambda > 0$ is a tuning parameter which controls the amount of regularization, w_i is a known weight for the ith observation, and v_j is the penalty factor for the jth covariate. These penalty factors are assumed to be known and allow parameters to be penalized differently. In particular, when $v_j = 1$ for $j = 1, \ldots, p$ then all parameters are regularized equally by λ, and when $v_j = 0$ the jth covariate is not penalized, i.e., it will always be included in the model. Note also that the intercept is not penalized. The estimator (2.3) simultaneously does variable selection and shrinks the regression coefficients towards 0. Depending on the choice of λ, $\widehat{\beta}_j(\lambda) = 0$ for some j's, and $\widehat{\beta}_j(\lambda)$ can be thought of as a shrunken least squares estimator (Bühlmann & Van De Geer, 2011). There are two potential benefits brought on by imposing a penalty term to the loss function. First, the least-squares estimate has low bias but a large prediction variance. Prediction accuracy can sometimes be improved by shrinking or biasing the regression coefficients towards 0. This is known as the bias-variance trade-off. Second, interpretation of the model becomes easier when only a small subset of predictors have been identified in being relevant to the response.

The lasso estimator can produce biased estimates for large coefficients (Zou, 2006). In

Section 2.2.4 we present the adaptive lasso which is a method that can mitigate this bias. It is also worth noting that (2.3) is a convex optimization problem and thus can be solved very efficiently using a block coordinate descent algorithm (Friedman et al., 2007; Tseng & Yun, 2009) for which we provide further details in Section 2.2.1. Other algorithms for solving this problem exist, including LARS (Efron et al., 2004) and the homotopy algorithm (Osborne et al., 2000), but these have been largely succeeded by the coordinate descent algorithm due to its speed and computational efficiency.

2.2.1   Block coordinate descent algorithm

In a series of seminal papers, Tseng laid the groundwork for a general purpose block coordinate descent algorithm (CGD) (Tseng, 2001; Tseng et al., 1988; Tseng & Yun, 2009) to minimize the sum of a smooth function f (i.e. continuously differentiable) and a separable convex function P of the form

$$Q_\lambda(\boldsymbol{\Theta}) = \arg \min_{\boldsymbol{\Theta}} \ f(\boldsymbol{\Theta}) + \lambda P(\boldsymbol{\Theta}) \tag{2.4}$$

At each iteration, the algorithm approximates $f(\boldsymbol{\Theta})$ in (2.4) by a strictly convex quadratic function and then applies block coordinate descent to generate a descent direction followed by an inexact line search along this direction (Tseng & Yun, 2009). For continuously differentiable $f(\cdot)$ and convex and block-separable $P(\cdot)$ (e.g. $P(\boldsymbol{\beta}) = \sum_i P_i(\beta_i)$), Tseng & Yun (2009) show that the solution generated by the CGD method is a stationary point of (2.4) if the coordinates are updated in a Gauss-Seidel manner, i.e., $Q_\lambda(\boldsymbol{\Theta})$ is minimized with respect to one parameter while holding all others fixed. The separability of the penalty function into a sum of functions of each individual parameter is the key to applying this algorithm to lasso type problems. Indeed, the CGD algorithm has been successfully applied in fixed effects models (Friedman et al., 2010; Meier et al., 2008) and linear mixed models (Schelldorfer et al., 2011). Following Tseng & Yun (2009), the general purpose CGD algorithm for

solving (2.4) is given by Algorithm 1.

Algorithm 1: Coordinate Gradient Descent Algorithm to solve (2.4)
___

Set the iteration counter $k \leftarrow 0$ and choose initial values for the parameter vector $\mathbf{\Theta}^{(0)}$;

repeat

    Approximate the Hessian $\nabla^2 f(\mathbf{\Theta}^{(k)})$ by a symmetric matix $H^{(k)}$:

$$H^{(k)} = \mathrm{diag}\left[\min\left\{\max\left\{\left[\nabla^2 f(\mathbf{\Theta}^{(k)})\right]_{jj}, 10^{-2}\right\} 10^9\right\}\right]_{j=1,\ldots,p} \tag{2.5}$$

    for $j = 1, \ldots, p$ **do**

        Solve the descent direction $d^{(k)} := d_{H^{(k)}}(\Theta_j^{(k)})$;

$$d_{H^{(k)}}(\Theta_j^{(k)}) \leftarrow \arg\min_d \left\{\nabla f(\Theta_j^{(k)})d + \frac{1}{2}d^2 H_{jj}^{(k)} + \lambda P(\Theta_j^{(k)} + d)\right\} \tag{2.6}$$

        Choose a stepsize;

$$\alpha_j^{(k)} \leftarrow \text{line search given by the Armijo rule}$$

        Update;

$$\widehat{\Theta}_j^{(k+1)} \leftarrow \widehat{\Theta}_j^{(k)} + \alpha_j^{(k)} d^{(k)}$$

    end

    $k \leftarrow k + 1$

until *convergence criterion is satisfied*;

The Armijo rule is defined as follows (Tseng & Yun, 2009):

> Choose $\alpha_{init}^{(k)} > 0$ and let $\alpha^{(k)}$ be the largest element of $\left\{\alpha_{init}^k \delta^r\right\}_{r=0,1,2,\ldots}$ satisfying
>
> $$Q_\lambda(\Theta_j^{(k)} + \alpha^{(k)} d^{(k)}) \le Q_\lambda(\Theta_j^{(k)}) + \alpha^{(k)} \varrho \Delta^{(k)} \tag{2.7}$$
>
> where $0 < \delta < 1$, $0 < \varrho < 1$, $0 \le \gamma < 1$ and
>
> $$\Delta^{(k)} := \nabla f(\Theta_j^{(k)}) d^{(k)} + \gamma(d^{(k)})^2 H_{jj}^{(k)} + \lambda P(\Theta_j^{(k)} + d^{(k)}) - \lambda P(\Theta_j^{(k)}) \tag{2.8}$$

Common choices for the constants are $\delta = 0.1$, $\varrho = 0.001$, $\gamma = 0$, $\alpha_{init}^{(k)} = 1$ for all k (Bertsekas,

We use Algorithm 1 to solve the lasso estimator with least-squares loss given by (2.3). Without loss of generality, we assume the penalty factors (v_j) are all equal to 1.

Descent Direction

For simplicity, we remove the iteration counter (k) from the derivation below.
For $\Theta_j^{(k)} \in \{\beta_1, \ldots, \beta_p\}$, let

$$d_H(\Theta_j) = \arg\min_d G(d) \tag{2.9}$$

where

$$G(d) = \nabla f(\Theta_j)d + \frac{1}{2}d^2 H_{jj} + \lambda|\Theta_j + d|$$

Since $G(d)$ is not differentiable at $-\Theta_j$, we calculate the subdifferential $\partial G(d)$ and search for d with $0 \in \partial G(d)$:

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} + \lambda u \tag{2.10}$$

where

$$u = \begin{cases} 1 & \text{if} \quad d > -\Theta_j \\ -1 & \text{if} \quad d < -\Theta_j \\ [-1, 1] & \text{if} \quad d = \Theta_j \end{cases} \tag{2.11}$$

We consider each of the three cases in (2.10) below

1. $d > -\Theta_j$

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} + \lambda = 0$$
$$d = \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}}$$

Since $\lambda > 0$ and $H_{jj} > 0$, we have

$$\frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}} > \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} = d \stackrel{\text{def}}{>} -\Theta_j$$

The solution can be written compactly as

$$d = \text{mid} \left\{ \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}, -\Theta_j, \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} \right\}$$

where $\text{mid} \{a, b, c\}$ denotes the median (mid-point) of a, b, c.

2. $d < -\Theta_j$

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} - \lambda = 0$$
$$d = \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}$$

Since $\lambda > 0$ and $H_{jj} > 0$, we have

$$\frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} < \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}} = d \stackrel{\text{def}}{<} -\Theta_j$$

Again, the solution can be written compactly as

$$d = \text{mid} \left\{ \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}, -\Theta_j, \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} \right\}$$

3. $d_j = -\Theta_j$

There exists $u \in [-1, 1]$ such that

$$\partial G(d) = \nabla f(\Theta_j) + dH_{jj} + \lambda u = 0$$
$$d = \frac{-(\nabla f(\Theta_j) + \lambda u)}{H_{jj}}$$

For $-1 \leq u \leq 1$, $\lambda > 0$ and $H_{jj} > 0$ we have

$$\frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} \leq d \overset{\text{def}}{=} -\Theta_j \leq \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}$$

The solution can again be written compactly as

$$d = \text{mid}\left\{ \frac{-(\nabla f(\Theta_j) - \lambda)}{H_{jj}}, -\Theta_j, \frac{-(\nabla f(\Theta_j) + \lambda)}{H_{jj}} \right\}$$

We see all three cases lead to the same solution for (2.9). Therefore the descent direction for $\Theta_j^{(k)} \in \{\beta_1, \ldots, \beta_p\}$ for the ℓ_1 penalty is given by

$$d = \text{mid}\left\{ \frac{-(\nabla f(\beta_j) - \lambda)}{H_{jj}}, -\beta_j, \frac{-(\nabla f(\beta_j) + \lambda)}{H_{jj}} \right\} \tag{2.12}$$

Solution for the β parameter

If the Hessian $\nabla^2 f(\boldsymbol{\Theta}^{(k)}) > 0$ then $H^{(k)}$ defined in (2.5) is equal to $\nabla^2 f(\boldsymbol{\Theta}^{(k)})$. Using $\alpha_{init} = 1$, the largest element of $\left\{ \alpha_{init}^{(k)} \delta^r \right\}_{r=0,1,2,\ldots}$ satisfying the Armijo Rule inequality is reached for $\alpha^{(k)} = \alpha_{init}^{(k)} \delta^0 = 1$. The Armijo rule update for the $\boldsymbol{\beta}$ parameter is then given by

$$\beta_j^{(k+1)} \leftarrow \beta_j^{(k)} + d^{(k)}, \qquad j = 1, \ldots, p \tag{2.13}$$

Substituting the descent direction given by (2.12) into (2.13) we get

$$\beta_j^{(k+1)} = \text{mid}\left\{ \beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) - \lambda)}{H_{jj}}, 0, \beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) + \lambda)}{H_{jj}} \right\} \tag{2.14}$$

15

We can further simplify this expression. First, we can re-write the loss function in (2.3) as

$$g(\boldsymbol{\beta}^{(k)}) = \frac{1}{2} \sum_{i=1}^{n} w_i \left(y_i - \sum_{\ell \neq j} X_{i\ell} \beta_\ell^{(k)} - X_{ij} \beta_j^{(k)} \right)^2 \tag{2.15}$$

The gradient and Hessian are given by

$$\nabla f(\beta_j^{(k)}) := \frac{\partial}{\partial \beta_j^{(k)}} g(\boldsymbol{\beta}^{(k)}) = -\sum_{i=1}^{n} w_i X_{ij} \left(y_i - \sum_{\ell \neq j} X_{i\ell} \beta_\ell^{(k)} - X_{ij} \beta_j^{(k)} \right) \tag{2.16}$$

$$H_{jj} := \frac{\partial^2}{\partial \beta_j^{(k)2}} g(\boldsymbol{\beta}^{(k)}) = \sum_{i=1}^{n} w_i X_{ij}^2 \tag{2.17}$$

Substituting (2.16) and (2.17) into $\beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) - \lambda)}{H_{jj}}$

$$\beta_j^{(k)} + \frac{\sum_{i=1}^{n} w_i X_{ij} \left(y_i - \sum_{\ell \neq j} X_{i\ell} \beta_\ell^{(k)} - X_{ij} \beta_j^{(k)} \right) + \lambda}{\sum_{i=1}^{n} w_i X_{ij}^2}$$

$$= \beta_j^{(k)} + \frac{\sum_{i=1}^{n} w_i X_{ij} \left(y_i - \sum_{\ell \neq j} X_{i\ell} \beta_\ell^{(k)} \right) + \lambda}{\sum_{i=1}^{n} w_i X_{ij}^2} - \frac{\sum_{i=1}^{n} w_i X_{ij}^2 \beta_j^{(k)}}{\sum_{i=1}^{n} w_i X_{ij}^2}$$

$$= \frac{\sum_{i=1}^{n} w_i X_{ij} \left(y_i - \sum_{\ell \neq j} X_{i\ell} \beta_\ell^{(k)} \right) + \lambda}{\sum_{i=1}^{n} w_i X_{ij}^2} \tag{2.18}$$

Similarly, substituting (2.16) and (2.17) in $\beta_j^{(k)} + \frac{-(\nabla f(\beta_j^{(k)}) + \lambda)}{H_{jj}}$ we get

$$\frac{\sum_{i=1}^{n} w_i X_{ij} \left(y_i - \sum_{\ell \neq j} X_{i\ell} \beta_\ell^{(k)} \right) - \lambda}{\sum_{i=1}^{n} w_i X_{ij}^2} \tag{2.19}$$

Finally, substituting (2.18) and (2.19) into (2.14) we get

$$
\begin{aligned}
\beta_j^{(k+1)} &= \text{mid} \left\{ \frac{\sum_{i=1}^n w_i X_{ij} \left(y_i - \sum_{\ell \neq j} X_{i\ell} \beta_\ell^{(k)} \right) - \lambda}{\sum_{i=1}^n w_i X_{ij}^2}, 0, \frac{\sum_{i=1}^n w_i X_{ij} \left(y_i - \sum_{\ell \neq j} X_{i\ell} \beta_\ell^{(k)} \right) + \lambda}{\sum_{i=1}^n w_i X_{ij}^2} \right\} \\
&= \frac{\mathcal{S}_\lambda \left(\sum_{i=1}^n w_i X_{ij} \left(y_i - \sum_{\ell \neq j} X_{i\ell} \beta_\ell^{(k)} \right) \right)}{\sum_{i=1}^n w_i X_{ij}^2}
\end{aligned} \tag{2.20}
$$

Where $\mathcal{S}_\lambda(x)$ is the soft-thresholding operator

$$
\mathcal{S}_\lambda(x) = \text{sign}(x)(|x| - \lambda)_+
$$

$\text{sign}(x)$ is the signum function

$$
\text{sign}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}
$$

and $(x)_+ = \max(x, 0)$. Since there is a closed form solution, which can be computed very quickly for each parameter update (given by (2.20)), the CGD algorithm is an attractive approach for solving the lasso estimator.

2.2.2 Lambda sequence

In general, the solution to (2.3) is computed over a decreasing sequence of values for the tuning parameter λ, beginning with the smallest value λ_{max} for which the entire coefficient vector $\widehat{\boldsymbol{\beta}} = \mathbf{0}_p$ (Friedman et al., 2010). To determine λ_{max}, we turn to the Karush-Kuhn-

Tucker (KKT) optimality conditions for (2.3). These conditions can be written as

$$\frac{1}{v_j} \sum_{i=1}^{n} w_i X_{ij} \left(y_i - \sum_{j=1}^{p} X_{ij} \hat{\beta}_j \right) = \lambda \gamma_j,$$

$$\gamma_j \in \begin{cases} \text{sign}(\hat{\beta}_j) & \text{if} \quad \hat{\beta}_j \neq 0 \\ [-1, 1] & \text{if} \quad \hat{\beta}_j = 0 \end{cases}, \qquad \text{for } j = 1, \ldots, p \tag{2.21}$$

where γ_j is the subgradient of the function $f(x) = |x|$ evaluated at $x = \hat{\beta}_j$. From (2.21), we can solve for the smallest value of λ such that the entire vector $(\hat{\beta}_1, \ldots, \hat{\beta}_p)$ is 0. This is given by

$$\lambda_{max} = \max_j \left\{ \left| \frac{1}{v_j} \sum_{i=1}^{n} w_i X_{ij} y_i \right| \right\}, \quad j = 1, \ldots, p \tag{2.22}$$

Following Friedman et al. (2010), we can choose $\tau \lambda_{max}$ to be the smallest value of tuning parameters λ_{min}, and construct a sequence of K values decreasing from λ_{max} to λ_{min} on the log scale. The defaults are set to $K = 100$, $\tau = 0.01$ if $n < p$ and $\tau = 0.001$ if $n \geq p$. The optimal value of λ can be chosen using 5-fold or 10-fold cross-validation. For least-squares loss, this corresponds to choosing the λ which minimizes the mean squared error.

2.2.3 Warm starts

The way in which we have derived the sequence of tuning parameters using the KKT conditions (Section 2.2.2), allows us to exploit warm starts which has been shown to lead to computational speedups (Friedman et al., 2010). That is, the solution $\widehat{\Theta}$ for λ_k is used as the initial value $\Theta^{(0)}$ for λ_{k+1}.

2.2.4   Adaptive lasso

It has been shown that the lasso estimator can produce biased estimates for large coefficients and give inconsistent variable selection results at the optimal λ for prediction, i.e., many noise features are included in the prediction model (Zou, 2006). To overcome the bias problems of the lasso, Zou (2006) proposed the adaptive lasso which allows a different amount of shrinkage for each regression coefficient using adaptive weights. Adaptive weighting has been shown to construct oracle procedures (Fan & Li, 2001), i.e., asymptotically, it performs as well as if the true model were given in advance. The adaptive lasso can be described as a two-stage procedure:

1. Calculate the initial regression estimates $\widehat{\boldsymbol{\beta}}_{init}$ from (2.3)

2. Refit (2.3) using penalty factors v_j equal to $1 / \left| \hat{\beta}_{init,j} \right|$ for $j = 1, \ldots, p$.

As we can see from the weights, the adaptive lasso will shrink larger coefficients less which leads to consistent variable selection results under weaker conditions than the lasso (Bühlmann & Van De Geer, 2011). We detail the adaptive lasso procedure in Algorithm 2.

Algorithm 2: Adaptive lasso algorithm

---
1. For a decreasing sequence $\lambda = \lambda_{max}, \ldots, \lambda_{min}$, fit the lasso with $v_j = 1$ for $j = 1, \ldots, p$
2. Use cross-validation or a data splitting procedure to determine the optimal value for the tuning parameter: $\lambda^{[opt]} \in \{\lambda_{max}, \ldots, \lambda_{min}\}$
3. Let $\hat{\beta}_{init,j}^{[opt]}$ for $j = 1, \ldots, p$ be the coefficient estimates corresponding to the model at $\lambda^{[opt]}$
4. Set the weights to be $v_j = \left(\left| \hat{\beta}_{init,j}^{[opt]} \right| \right)^{-1}$ for $j = 1, \ldots, p$
5. Refit the lasso with the weights defined in step 4), and use cross-validation or a data splitting procedure to choose the optimal value of λ

2.3   Group Lasso

One main drawback of the lasso is that it ignores the grouping structure of the design matrix. When given a predetermined grouping of non-overlapping variables, we might want
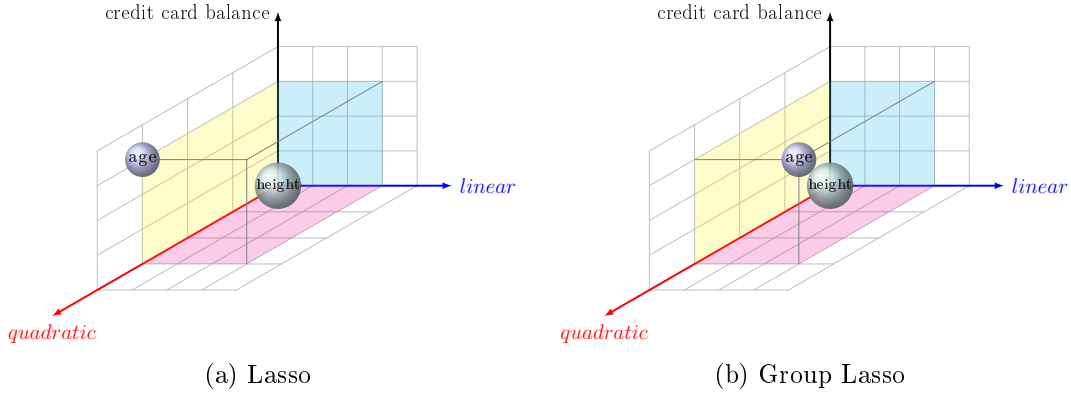
(a) Lasso   (b) Group Lasso

Figure 2.1: The selected components from Model (2.23) for (a) the lasso and (b) the group lasso. The origin of the xyz axis is zero which means both the linear and quadratic terms are 0. In this toy example, the lasso selects only the quadratic term for age while the group lasso selects both linear and quadratic terms.

the coefficients of all members of the group to be either zero or non-zero. For example, when dealing with categorical predictors where each factor is expressed through a set of indicator variables, removing an irrelevant factor is equivalent to setting the coefficients of all the indicator variables to 0. In a generalized additive model (Hastie & Tibshirani, 1987), where each variable is projected on to a set of basis function, e.g. $f_j(X_j) = \sum_{\ell=1}^{m_j} \psi_{j\ell}(X_j)\beta_{j\ell}$, we would want all $\{\beta_{j\ell}\}_{\ell=1}^{m_j}$ to be either zero or non-zero. This key difference between the lasso and group lasso penalty is illustrated in Figure 2.1. Suppose we want to predict an individual's credit card balance from their age and height using the following additive model:

$$\text{credit card balance} = \beta_0 + \beta_{11}\text{age} + \beta_{12}\text{age}^2 + \beta_{21}\text{height} + \beta_{22}\text{height}^2 + \varepsilon \qquad (2.23)$$

In Figures 2.1a and 2.1b we see that both the lasso and group lasso set the linear and quadratic terms for height $(\hat{\beta}_{21}, \hat{\beta}_{22})$ to 0. However, the lasso estimates only a nonzero quadratic term for age $(\hat{\beta}_{11} = 0, \hat{\beta}_{12} \neq 0)$ while the group lasso estimates both linear and quadratic terms to be nonzero $(\hat{\beta}_{11} \neq 0, \hat{\beta}_{12} \neq 0)$. We now provide details on the group lasso estimator.

Assume that the predictors in the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ belong to K groups and define the cardinality of index set I_k to be p_k. These groups are known *a priori* such that $(1, 2, \ldots, p) = \bigcup_{k=1}^{K} I_k$, and are also non-overlapping, i.e., $I_k \bigcap I_{k'} = \emptyset$ for $k \neq k'$. Therefore, group k contains p_k predictors corresponding to $\mathbf{X}^{(k)}$, i.e., the columns of the design matrix X_j for $j \in I_k$, and $1 \leq k \leq K$. The intercept belongs to its own group, i.e., $I_1 = \{1\}$. The group lasso partitions the variable coefficients into K groups $\boldsymbol{\beta} = ([\boldsymbol{\beta}^{(1)}]^{\intercal}, [\boldsymbol{\beta}^{(2)}]^{\intercal}, \cdots, [\boldsymbol{\beta}^{(K)}]^{\intercal})^{\intercal}$, where $\boldsymbol{\beta}^{(k)}$ denotes the segment of $\boldsymbol{\beta}$ corresponding to group k. For least-squares loss, the group lasso estimator (Yuan & Lin, 2006) is given by:

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{(\beta_0, \boldsymbol{\beta})}{\arg\min} \frac{1}{2} \sum_{i=1}^{n} w_i (y_i - \beta_0 - (\mathbf{X}\boldsymbol{\beta})_i)^2 + \lambda \sum_{k=1}^{K} v_k \|\boldsymbol{\beta}^{(k)}\|_2 \qquad (2.24)$$

where $\left\|\boldsymbol{\beta}^{(k)}\right\|_2 = \sqrt{\sum_{j \in I_k} \beta_j^2}$ and $\lambda > 0$ is the tuning parameter. As in the lasso estimator (2.3), there are both observation weights w_i, and penalty factors $v_k \geq 0$ which control the relative strength of the terms within the group lasso penalty. These penalty factors are often set to $\sqrt{p_k}$ (Yuan & Lin, 2006). Note that the same penalty factor is applied to all the coefficients in a group. Solving the group lasso estimator is more challenging than the lasso since there is no closed form solution for (2.24). In the next section, we detail a majorization-minimization (MM) type algorithm (Lange et al., 2000; Y. Yang & Zou, 2015) used to solve (2.24).

2.3.1   Groupwise majorization descent algorithm

This description of the groupwise majorization descent (GMD) algorithm used to solve (2.24) follows mainly from Y. Yang & Zou (2015). The main difference here is that we consider a more general loss function of the form

$$L(\boldsymbol{\beta} \mid \mathbf{D}) = \frac{1}{2} [\boldsymbol{y} - \widehat{\boldsymbol{y}}]^{\top} \mathbf{W} [\boldsymbol{y} - \widehat{\boldsymbol{y}}] \qquad (2.25)$$

where $\widehat{\boldsymbol{y}} = \hat{\beta}_0 + \mathbf{X}\widehat{\boldsymbol{\beta}}$, $\mathbf{D}$ is the working data $\{\boldsymbol{y}, \mathbf{X}\}$, and $\mathbf{W}$ is an $n \times n$ nonsingular and known weight matrix. This weight matrix can be used when the elements of $\boldsymbol{y}$ are correlated as is done in generalized least squares. The original proposal in Y. Yang & Zou (2015) is a special case of (2.25) where $\mathbf{W}$ is a diagonal matrix with entries equal to w_i. The loss function (2.25) satisfies the quadratic majorization (QM) condition, since $L(\boldsymbol{\beta} \mid \mathbf{D})$ is differentiable as a function of $\boldsymbol{\beta}$, i.e., $\nabla L(\boldsymbol{\beta}|\mathbf{D}) = -(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top \mathbf{W}\mathbf{X}$, and there exists a $p \times p$ matrix $\mathbf{H} = \mathbf{X}^\top \mathbf{W}\mathbf{X}$ which only depends on the data $\mathbf{D}$, such that for all $\boldsymbol{\beta}, \boldsymbol{\beta}^*$,

$$L(\boldsymbol{\beta} \mid \mathbf{D}) \leq L(\boldsymbol{\beta}^* \mid \mathbf{D}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \nabla L(\boldsymbol{\beta}^*|\mathbf{D}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top \mathbf{H}(\boldsymbol{\beta} - \boldsymbol{\beta}^*). \qquad (2.26)$$

We can exploit the fact that the loss function (2.25) satisfies the QM condition to majorize the loss function in (2.24) by (2.26), and that the penalty term $\sum_{k=1}^{K} v_k \|\boldsymbol{\beta}^{(k)}\|_2$ in (2.24) is separable with respect to the indices of the variables $k = 1, \ldots, K$ to performs groupwise updates for each $\boldsymbol{\beta}^{(k)}$.

Let $\widetilde{\boldsymbol{\beta}}$ denote the current solution of $\boldsymbol{\beta}$ and define $\mathbf{H}^{(k)}$ as the sub-matrix of $\mathbf{H}$ corresponding to group k. For example, if group 2 is $\{2, 4\}$ then $\mathbf{H}^{(2)}$ is a 2×2 matrix with

$$\mathbf{H}^{(2)} = \begin{bmatrix} h_{2,2} & h_{2,4} \\ h_{4,2} & h_{4,4} \end{bmatrix},$$

where $h_{i,j}$ is the i, jth entry of the $\mathbf{H}$ matrix. Write $\boldsymbol{\beta}$ such that $\boldsymbol{\beta}^{(k')} = \widetilde{\boldsymbol{\beta}}^{(k')}$ for $k' \neq k$. Given $\boldsymbol{\beta}^{(k')} = \widetilde{\boldsymbol{\beta}}^{(k')}$ for $k' \neq k$, the estimator for $\boldsymbol{\beta}^{(k)}$ is given by

$$\widehat{\boldsymbol{\beta}}^{(k)}(\lambda) = \arg\min_{\boldsymbol{\beta}^{(k)}} L(\boldsymbol{\beta} \mid \mathbf{D}) + \lambda v_k \|\boldsymbol{\beta}^{(k)}\|_2. \qquad (2.27)$$

Using $\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}} = (\underbrace{0, \ldots, 0}_{k-1}, \boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)}, \underbrace{0, \ldots, 0}_{K-k})$ and the majorization defined in (2.26) we can

write

$$L(\boldsymbol{\beta} \mid \mathbf{D}) \leq L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) - (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top U^{(k)} + \frac{1}{2}(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top \mathbf{H}^{(k)}(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)}). \quad (2.28)$$

where

$$U^{(k)} = \frac{\partial}{\partial \boldsymbol{\beta}^{(k)}} L(\boldsymbol{\beta} \mid \mathbf{D}) = -(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top \mathbf{W} \mathbf{X}^{(k)}, \quad (2.29)$$

$$\mathbf{H}^{(k)} = \frac{\partial^2}{\partial \boldsymbol{\beta}^{(k)} \partial \boldsymbol{\beta}^{(k)\top}} L(\boldsymbol{\beta} \mid \mathbf{D}) = (\mathbf{X}^{(k)})^\top \mathbf{W} \mathbf{X}^{(k)}. \quad (2.30)$$

The upper bound in (2.28) can be majorized even further to avoid expensive matrix multiplications and storing $\mathbf{H}^{(k)}$ in memory. Let η_k be the largest eigenvalue of $\mathbf{H}^{(k)}$. We set $\gamma_k = (1 + \varepsilon^*)\eta_k$, where $\varepsilon^* = 10^{-6}$ and substitute γ_k for $\mathbf{H}^{(k)}$ in (2.28):

$$L(\boldsymbol{\beta} \mid \mathbf{D}) \leq L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) - (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top U_{(k)} + \frac{1}{2}\gamma_k(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)}). \quad (2.31)$$

It is important to note that the inequality strictly holds unless when $\boldsymbol{\beta}^{(k)} = \widetilde{\boldsymbol{\beta}}^{(k)}$. Substituting the upper bound in (2.31) for $L(\boldsymbol{\beta} \mid \mathbf{D})$ in (2.27) we get the following estimator for $\boldsymbol{\beta}^{(k)}$

$$\arg\min_{\boldsymbol{\beta}^{(k)}} L(\widetilde{\boldsymbol{\beta}} \mid \mathbf{D}) - (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top U^{(k)} + \frac{1}{2}\gamma_k(\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)})^\top (\boldsymbol{\beta}^{(k)} - \widetilde{\boldsymbol{\beta}}^{(k)}) + \lambda v_k \|\boldsymbol{\beta}^{(k)}\|_2. \quad (2.32)$$

Let $\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new})$ be the solution to (2.32) which has a simple closed-from expression:

$$\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) = \frac{1}{\gamma_k}\left(U^{(k)} + \gamma_k\widetilde{\boldsymbol{\beta}}^{(k)}\right)\left(1 - \frac{\lambda v_k}{\|U^{(k)} + \gamma_k\widetilde{\boldsymbol{\beta}}^{(k)}\|_2}\right)_+. \quad (2.33)$$

Algorithm 3 summarizes the details of GMD for the group lasso with least-squares loss function given by (2.25). The GMD algorithm has the strict descent property and converges to the right answer. The proof follows directly from Section 3.1 of Y. Yang & Zou (2015).
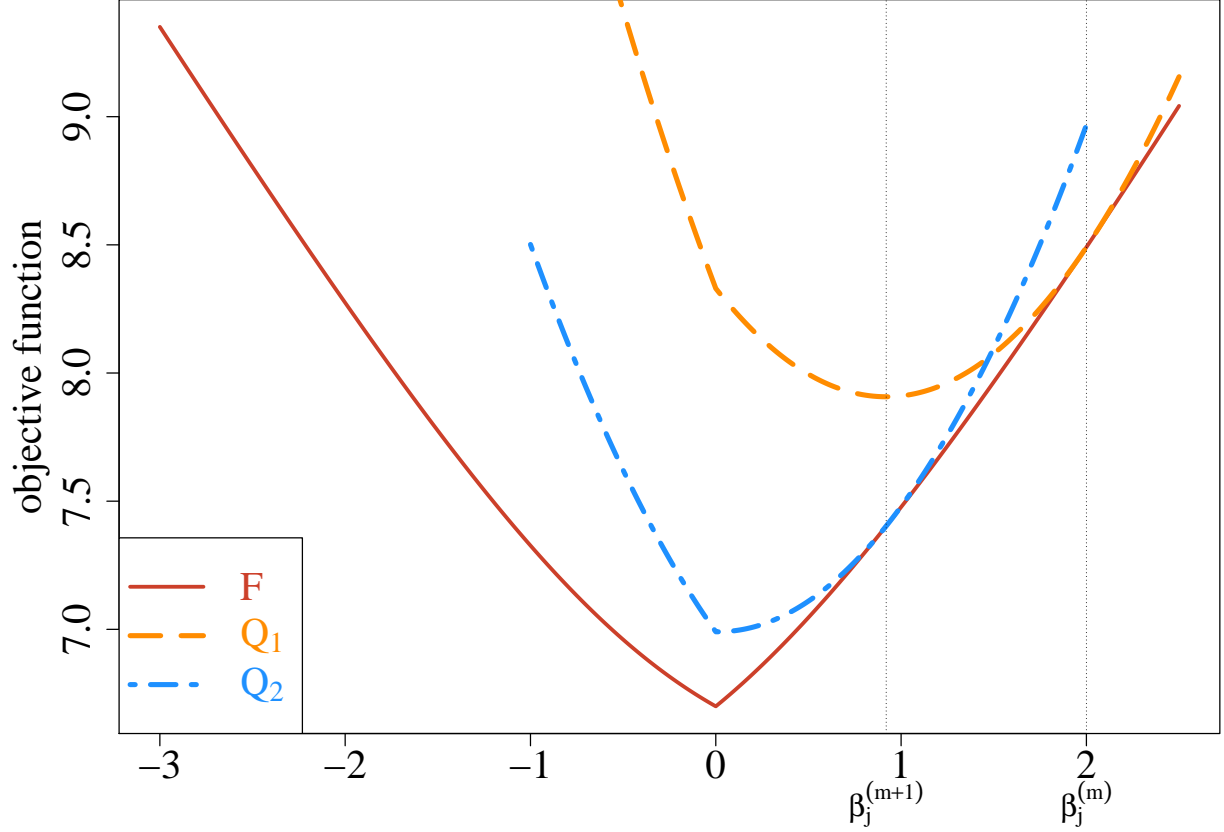
23

Figure 2.2: Illustration of the quadratic majorization technique for updating the coefficient β_j from the m step to the $(m+1)$ step. F is the objective function for which there is no closed form solution. Q_1 and Q_2 are the surrogate functions for which we can find the minimum.

In Figure 2.2 we provide an illustration of the quadratic majorization technique for updating a parameter β_j. The solution lies at the minimum of the F curve but we cannot solve for this directly since there is no closed form solution. Instead we majorize F using Q_1 which is a function consisting of the quadratic approximation of F plus the penalty term evaluated at $\beta_j^{(m)}$. The minimum of Q_1, for which there is a closed form solution, corresponds to the next iteration $\beta_j^{(m+1)}$. We then majorize again using Q_2 and solve for the minimum. This process is repeated until convergence.

Algorithm 3: The GMD algorithm for group lasso with least-squares loss function given by (2.25).

---

1. For $k = 1, \ldots, K$, compute γ_k, the largest eigenvalue of $\mathbf{H}^{(k)} = (\mathbf{X}^{(k)})^\top \mathbf{W} \mathbf{X}^{(k)}$
2. Initialize $\widetilde{\boldsymbol{\beta}}$.
3. Repeat the following cyclic groupwise updates until convergence:
   — for $k = 1, \ldots, K$, do step (3.1)–(3.3)
   3.1 Compute $U(\widetilde{\boldsymbol{\beta}}) = -\nabla L(\widetilde{\boldsymbol{\beta}}|\mathbf{D}) = -(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top \mathbf{W} \mathbf{X}^{(k)}$
   3.2 Compute $\widetilde{\boldsymbol{\beta}}^{(k)}(\text{new}) = \frac{1}{\gamma_k}\left(U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\right)\left(1 - \frac{\lambda v_k}{\|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2}\right)_+$.
   3.3 Set $\widetilde{\boldsymbol{\beta}}^{(k)} = \widetilde{\boldsymbol{\beta}}^{(k)}(\text{new})$.

2.3.2   Lambda sequence

Similar to Section 2.2.2, we compute the solution to (2.24) over a decreasing sequence of values for the tuning parameter λ starting with λ_{max}. From the update formula (2.33) we have that for all k

$$
\begin{cases}
\widetilde{\boldsymbol{\beta}}^{(k)} = \frac{1}{\gamma_k}\left(U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\right)\left(1 - \frac{\lambda v_k}{\|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2}\right) & \text{if } \|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2 > \lambda v_k \\
\widetilde{\boldsymbol{\beta}}^{(k)} = \mathbf{0} & \text{if } \|U^{(k)} + \gamma_k \widetilde{\boldsymbol{\beta}}^{(k)}\|_2 \leq \lambda v_k \; .
\end{cases}
$$

We can then directly obtain the KKT conditions for $k = 1, \ldots, K$:

$$
-U^{(k)} + \lambda v_k \cdot \frac{\widetilde{\boldsymbol{\beta}}^{(k)}}{\|\widetilde{\boldsymbol{\beta}}^{(k)}\|_2} = \mathbf{0} \qquad \text{if } \widetilde{\boldsymbol{\beta}}^{(k)} \neq \mathbf{0},
$$
$$
\left\|U^{(k)}\right\|_2 \leq \lambda w_k \qquad \text{if } \widetilde{\boldsymbol{\beta}}^{(k)} = \mathbf{0} \; .
$$

(2.34)

Using (2.34) we can solve for the smallest value of λ such that the entire vector $\left\{\widehat{\boldsymbol{\beta}}^{(k)}\right\}_{k=1}^{K}$ is 0. This is given by

$$
\lambda_{max} = \max_k \frac{1}{v_k}\left\|U^{(k)}\right\|_2, \quad k = 1, \ldots, K, \; v_k \neq 0
$$

(2.35)

2.3.3 Warm starts and adaptive group lasso

Warm starts can also be implemented for the group lasso as described in Section 2.2.3 for the lasso. Furthermore, the adaptive group lasso can be computed using Algorithm 2; the main difference is that the coefficient estimates in Step 1 are obtained from a group lasso fit and the weights for group k are given by $v_k = \left(\left\| \widehat{\boldsymbol{\beta}}^{(k)} \right\|_2 \right)^{-1}$.

2.4 Penalized interaction models

In this section, we introduce penalized regression methods in the context of interaction models. We consider a regression model for an n-length outcome variable $\boldsymbol{y}$ which follows an exponential family. Let $E = (e_1, \ldots, e_n)$ be the binary environment or exposure vector and $\mathbf{X}$ the $n \times p$ matrix of high-dimensional data. Consider the regression model with main effects and their interactions with E:

$$g(\boldsymbol{\mu}) = \beta_0 + \underbrace{\beta_1 X_1 + \cdots + \beta_p X_p + \beta_E E}_{\text{main effects}} + \underbrace{\alpha_{1E}(X_1 E) + \cdots + \alpha_{pE}(X_p E)}_{\text{interactions}} \tag{2.36}$$

where $g(\cdot)$ is a known link function and $\boldsymbol{\mu} = \mathsf{E}[\boldsymbol{y}|\mathbf{X}, E, \boldsymbol{\beta}, \boldsymbol{\alpha}]$. Due to the large number of parameters to estimate with respect to the number of observations, we might consider adding the lasso penalty introduced in Section 2.2. A natural extension of the lasso penalty applied to the interaction model (2.36) is

$$\underset{\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}}{\arg \min} \frac{1}{2} \|Y - g(\boldsymbol{\mu})\|_2^2 + \lambda \left(\|\boldsymbol{\beta}\|_1 + \|\boldsymbol{\alpha}\|_1 \right) , \tag{2.37}$$

where $\|Y - g(\boldsymbol{\mu})\|_2^2 = \sum_i (y_i - g(\mu_i))^2$, $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j|$, $\|\boldsymbol{\alpha}\|_1 = \sum_j |\alpha_j|$ and $\lambda \geq 0$ is the tuning parameter that can set some of the coefficients to zero when sufficiently large. A limitation of (2.37) is that the selected model may have main effects that are 0 but the corresponding interaction terms are not. This is due to the penalty which treats both main

effects and interactions equally. While there may exist situations where this can occur, statisticians have long argued that large main effects are more likely to lead to detectable interactions than small ones (Cox, 1984) and therefore, main effects should enter the model before interactions. This is known as the strong heredity principle (Chipman, 1996):

$$\hat{\alpha}_{jE} \neq 0 \qquad \Rightarrow \qquad \hat{\beta}_j \neq 0 \quad \text{and} \quad \hat{\beta}_E \neq 0 \qquad (2.38)$$

In words, the interaction term will only have a non-zero estimate if its corresponding main effects are estimated to be non-zero. One benefit brought by hierarchy is that the number of measured variables can be reduced, referred to as practical sparsity (Bien et al., 2013; She & Jiang, 2014). For example, a model involving $X_1, E, X_1 \cdot E$ is more parsimonious than a model involving $X_1, E, X_2 \cdot E$, because in the first model a researcher would only have to measure two variables compared to three in the second model.

There have been several proposals for modeling interactions with the strong heredity constraint in the penalization literature including Composite Absolute Penalties (CAP) (Zhao et al., 2009), Variable selection using Adaptive Nonlinear Interaction Structures in High dimensions (VANISH) (Radchenko & James, 2010), Strong Hierarchical Lasso (hierNet) (Bien et al., 2013), Group-Lasso Interaction Network (glinternet) (Lim & Hastie, 2015), Group Regularized Estimation under Structural Hierarchy (GRESH) (She & Jiang, 2014) and a Framework for Modeling Interactions with a Convex Penalty (FAMILY) (Haris et al., 2014). A popular approach to achieve this structured sparsity is via the group lasso penalty where each group contains both main effects and their corresponding interactions. For example, consider a two-way interaction model of the form

$$\begin{aligned} y = & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ & + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \varepsilon \end{aligned}$$

A grouping structure that would satisfy the strong heredity property is:

group	1	2	3	4			5			6		
parameter	β_1	β_2	β_3	β_1	β_2	β_{12}	β_1	β_3	β_{13}	β_2	β_3	β_{23}

A limitation of this approach is that its impossible to select one variable without selecting all the groups containing it (Jacob et al., 2009). This implies that if a main effect has been selected, all the interaction terms with that main effect will necessarily be selected as well which can be an issue when the truth contains no interactions.

In Chapter 3 of the thesis, we introduce a new structured sparsity called `sail` which satisfies the strong heredity property while overcoming this issue, i.e., a selected main effect does not automatically imply that the corresponding interactions will also be selected. Furthermore, our method can accommodate non-linear interaction effects on the response. The context of Chapter 3 is summarized in Table 2.1.

Table 2.1: Overview of methods with structured sparsity for penalized regression models

Type	Method	Software
Linear	`CAP` (Zhao et al., 2009)	✗
	`SHIM` (Choi et al., 2010)	✗
	`hiernet` (Bien et al., 2013)	`hierNet(x, y)`
	`GRESH` (She & Jiang, 2014)	✗
	`FAMILY` (Haris et al., 2016)	`FAMILY(x, z, y)`
	`glinternet` (Lim & Hastie, 2015)	`glinternet(x, y)`
	`RAMP` (Hao et al., 2018)	`RAMP(x, y)`
	`LassoBacktracking` (Shah, 2016)	`LassoBT(x, y)`
Non-linear	`VANISH` (Radchenko & James, 2010)	✗
	`sail` (Chapter 3 of the thesis)	`sail(x, y, e)`

2.5   Linear mixed-effects models

Linear mixed-effects models (LMMs) ([Laird & Ware, 1982](#)) are a class of statistical models used to account for correlations induced by a natural grouping of the observations such as students in schools or repeated measurements on the same individual. By introducing a subject specific random effect, LMMs can simultaneously model the mean and covariance structure to produce valid standard errors. In this section we briefly introduce LMMs as a preface to Chapter 4 of the thesis which is about penalized LMMs.

Let $i = 1, \ldots, N$ be the grouping index, $j = 1, \ldots, n_i$ the observation index within a group and $N_T = \sum_{i=1}^N n_i$ the total number of observations. For each group let $\boldsymbol{y}_i = (y_1, \ldots, y_{n_i})$ be the observed vector of responses, $\mathbf{X}_i$ an $n_i \times (p+1)$ design matrix (with the column of 1s for the intercept), $\boldsymbol{b}_i$ a group-specific random effect vector of length n_i and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})$ the individual error terms. Denote the stacked vectors $\mathbf{Y} = (\boldsymbol{y}_i, \ldots, \boldsymbol{y}_N)^T \in \mathbb{R}^{N_T \times 1}$, $\boldsymbol{b} = (\boldsymbol{b}_i, \ldots, \boldsymbol{b}_N)^T \in \mathbb{R}^{N_T \times 1}$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_i, \ldots, \boldsymbol{\varepsilon}_N)^T \in \mathbb{R}^{N_T \times 1}$, and the stacked matrix $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)^T \in \mathbb{R}^{N_T \times (p+1)}$. Furthermore, let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T \in \mathbb{R}^{(p+1) \times 1}$ be a vector of fixed effects regression coefficients corresponding to $\mathbf{X}$. We consider the following LMM with a single random effect:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{b} + \boldsymbol{\varepsilon} \tag{2.39}$$

where the random effect $\boldsymbol{b}$ and the error variance $\boldsymbol{\varepsilon}$ are assigned the distributions

$$\boldsymbol{b} \sim \mathcal{N}(0, \eta\sigma^2\boldsymbol{\Phi}) \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, (1-\eta)\sigma^2\mathbf{I}) \tag{2.40}$$

Here, $\boldsymbol{\Phi}_{N_T \times N_T}$ is a known positive semi-definite and symmetric covariance matrix, $\mathbf{I}_{N_T \times N_T}$ is the identity matrix and parameters σ^2 and $\eta \in [0, 1]$ determine how the variance is divided

between $\boldsymbol{b}$ and $\boldsymbol{\varepsilon}$. The joint density of $\mathbf{Y}$ is multivariate normal:

$$\mathbf{Y}|(\boldsymbol{\beta}, \eta, \sigma^2) \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \eta\sigma^2\boldsymbol{\Phi} + (1 - \eta)\sigma^2\mathbf{I}) \tag{2.41}$$

In high-dimensional contexts, the classic LMM is no longer applicable for the same reasons why the linear model does not apply, i.e., singularity of $\mathbf{X}^T\mathbf{X}$. However, naively applying a penalized method such as the lasso to clustered data is also invalid. Indeed, ignoring the grouping structure can lead to incorrect inference due to biased standard errors of parameters, and can also violate the normality of residuals assumption. In Chapter 4 of this thesis, we develop a general framework for penalized mixed models that simultaneously performs variable selection while accounting for the correlations induced by the natural clustering of the observations.

References

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. (2012). Structured sparsity through convex optimization. *Statistical Science*, *27*(4), 450–468.

8

Bertsekas, D. P. (1999). *Nonlinear programming.* Athena scientific Belmont.

12

Bhatnagar, S. R., Yang, Y., Khundrakpam, B., Evans, A. C., Blanchette, M., Bouchard, L., & Greenwood, C. M. (2018). An analytic approach for interpretable predictive models in high-dimensional data in the presence of interactions with exposures. *Genetic epidemiology*, *42*(3), 233–249.

9

Bien, J., Taylor, J., Tibshirani, R., et al. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, *41*(3), 1111–1141.

27, 28

Bühlmann, P., Rütimann, P., van de Geer, S., & Zhang, C.-H. (2013). Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, *143*(11), 1835–1858.

8, 9

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.
10, 19

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, *24*(1), 17–36.
27

Choi, N. H., Li, W., & Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, *105*(489), 354–364.
28

Cox, D. R. (1984). Interaction. *International Statistical Review/Revue Internationale de Statistique*, 1–24.
27

Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS genetics*, *9*(3), e1003348.
6

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, *32*(2), 407–499.
11

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National science review*, *1*(2), 293–314.
2

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), 1348–1360.
7, 19

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*(2), 302–332.

11

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics Springer, Berlin.

2

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), 1.

9, 11, 17, 18

Hao, N., Feng, Y., & Zhang, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 1–11.

28

Haris, A., Witten, D., & Simon, N. (2014). Convex modeling of interactions with strong heredity. *arXiv preprint arXiv:1410.3517*.

27

Haris, A., Witten, D., & Simon, N. (2016). Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, *25*(4), 981–1004.

28

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*(398), 371–386.

20

Hastie, T., Tibshirani, R., Botstein, D., & Brown, P. (2001). Supervised harvesting of expression trees. *Genome Biology*, *2*(1), 1–0003.

9

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthog-
onal problems. *Technometrics*, *12*(1), 55–67.
7

Jacob, L., Obozinski, G., & Vert, J.-P. (2009). Group lasso with overlap and graph lasso.
In *Proceedings of the 26th annual international conference on machine learning* (pp. 433–
440).
28

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., . . . others (2008).
Kegg for linking genomes to life and the environment. *Nucleic acids research*, *36*(suppl
1), D480–D484.
8

Kendall, M. (1957). *A course in multivariate analysis*. London: Griffin.
9

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*,
963–974.
29

Lange, K., Hunter, D., & Yang, I. (2000). Optimization transfer using sur- rogate objective
functions (with discussion). *Journal of Computational and Graphical Statistics*, *9*, 1-20.
21

Leek, J. T., & Storey, J. D. (2008). A general framework for multiple testing dependence.
Proceedings of the National Academy of Sciences, *105*(48), 18718–18723.
6

Lim, M., & Hastie, T. (2015). Learning interactions via hierarchical group-lasso regulariza-
tion. *Journal of Computational and Graphical Statistics*, *24*(3), 627–654.
27, 28

Lin, X., Lee, S., Christiani, D. C., & Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*, kxt006.
6

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., ... others (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.
6

Meier, L., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 53–71.
11

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, *52*(1), 374–393.
7

Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *Journal of Statistical Software*, *53*(9), 1–18. Retrieved from http://www.jstatsoft.org/v53/i09/
9

Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, *20*(3), 389–403.
11

Park, M. Y., Hastie, T., & Tibshirani, R. (2007). Averaged gene expressions for regression. *Biostatistics*, *8*(2), 212–227.
9

Radchenko, P., & James, G. M. (2010). Variable selection using adaptive nonlinear inter-

action structures in high dimensions. *Journal of the American Statistical Association*, *105*(492), 1541–1553.

27, 28

Reid, S., Tibshirani, R., & Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, 35–67.

2

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature*, *461*(7261), 218–223.

6

Schelldorfer, J., Bühlmann, P., DE, G., & VAN, S. (2011). Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, *38*(2), 197–214.

11

Shah, R. D. (2016). Modelling interactions in high-dimensional data with backtracking. *Journal of Machine Learning Research*, *17*(207), 1–31.

28

She, Y., & Jiang, H. (2014). Group regularized estimation under structural hierarchy. *arXiv preprint arXiv:1411.4691*.

27, 28

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

7, 10

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

7

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, *109*(3), 475–494.
11

Tseng, P., et al. (1988). Coordinate ascent for maximizing nondifferentiable concave functions. 11

Tseng, P., & Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, *117*(1), 387–423.
11, 12

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., ... Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, *15*(1), 273–289.
8

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., ... others (2010). Common snps explain a large proportion of the heritability for human height. *Nature genetics*, *42*(7), 565.
6

Yang, Y., & Zou, H. (2014). gglasso: Group lasso penalized learning using a unified bmd algorithm. Retrieved from http://CRAN.R-project.org/package=gglasso (R package version 1.3)
9

Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, *25*(6), 1129–1141.

21, 22, 23

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

8, 21

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 894–942.

7

Zhao, P., Rocha, G., & Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 3468–3497.

27, 28

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

7, 10, 19

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

8