FITTING SPARSE HIERARCHICAL MODELS: APPLICATIONS TO FACTORIAL DESIGNS

by

Majid Nabipoor Sanjebad

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences

University of Alberta

Abstract

We study penalized fitting strategies aimed at sparse model selection of models satisfying certain hierarchical restrictions, in linear models arising from factorial experiments. After discussing various merits of existing approaches, we propose a modification and generalization of the approach of Bien, Taylor and Tibshirani, capable of handling also models with factors with possibly more than two levels. The approach is based on the modified constraint used in conjunction with the group LASSO. The effect of the modified constraint on the selection of main effects and pair interactions is explored. We characterize the solution for both quadratic and logistic loss and give an unbiased Stein-type estimate for the degrees of freedom, the quantity required as the key component for the selection among competing models in regularization. We compare the derived estimates of the degrees of freedom with the existing ones from the literature.

We also study properties of certain alternative approaches: for the so-called standardized group LASSO of Simon and Tibshirani, we show first that it remains unchanged under the transformation of Zhao et al., aimed at unifying group weights, and then we characterize the solution of the newly standardized group LASSO. Based on this characterization, we again derive the unbiased estimate of the degrees of freedom. We establish such an estimate of the degrees of freedom also for the overlapped group LASSO of Obozinski et al.

We after show that the derived estimates of the degrees of freedom converge, when the tuning parameter converges to zero, to the (true) degrees of freedom of the corresponding

constrained least-squares estimator. We investigate certain particular properties of sparse fitting procedures in factorial designs. We establish the connection, for balanced designs, between penalized estimation and traditional constrained least-squares estimators. We also propose methods of selecting the regularization parameter selection based on AIC and BIC. Finally, we show how replications in factorial designs affect the selection process of standardized group LASSO.

To Tahereh

Acknowledgement

I would like to express my sincere gratitude to my supervisor Dr. Mizera for his guidance, critical comments, inspirations, and supportive attitudes throughout my studies. During my Ph.D. study, Dr. Mizera has always given me his strong support and invaluable advice. Without his patience and generous support, as well as his massive knowledge during the thesis writing period, this thesis would not have come as far as it did.

I would like to express my appreciation to Dr. Osornio-Vargas for his supportive attitude and generous support during the last year of my study. Also, I would like to express my thanks to Dr. Hillen for his valuable comments and supportive attitude during my Ph.D. study. I wish to express my appreciation to Dr. Wiens for his constructive suggestions and valuable comments. I wish to thank Dr. Kong for reading my thesis.

Finally, none of this would have been possible without the enduring love and faith of my parents and my wife.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

New scientific problems arising in recent decades brought a need for the analysis of so-called high-dimensional data sets. They are called so because in those data sets the number of predictors p is large and often exceeds the sample size n. For instance, the DNA micro-array data set, analyzed by Hastie et al. [9], consists of 6830 genes of human tumours of 64 patients; that is $n = 64$ and $p = 6830$. In such a situation, the columns of the design matrix X are not independent, and $X^T X$ is singular. Therefore, the traditional theory of least squares regression is not applicable; among other things, it is not possible to calculate p-values and select predictors. Even if p does not exceed n, but is large, using least squares regression and calculating all p-values can be problematic.

High-dimensional data are often analyzed by penalization. The general scheme of penalized regression is

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(Loss + \lambda \, Penalty \right),$$

where $Loss$ quantifies the prediction error, $Penalty$ expresses the desired condition on the fit, and λ determines the weight or importance of the penalty term.

One of the first penalized regression models was ridge regression. Consider a regression

model with an outcome Y and design matrix $X_{n\times p}$, where n is the number of observations and p is the number of predictors. For the regression model,

$$Y = X\beta + \varepsilon,$$

the ridge regression estimate $\hat{\beta}$ is

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \sum_{i=1}^{p} \beta_i^2,$$

where $\lambda > 0$. Solving this problem yields $\hat{\beta} = \left(X^T X + \lambda I\right)^{-1} X^T Y$. Now, even if $X^T X$ is singular, $X^T X + \lambda I$ is invertible.

Ridge regression solves the collinearity problem among predictors, but it has a tendency to result in small nonzero estimates for the components of β with true values equal to zero. Tibshirani [29] proposed LASSO to shrink and select variables in high-dimensional data sets. The LASSO estimate is defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \sum_{i=1}^{p} |\beta_i|.$$

LASSO typically produces a vector of coefficients β with many zero components. It not only addresses collinearity as the ridge regression does, but also selects variables by assigning zero to predictors with little or no importance; the nonzero components yield the model with relatively few nonzero parameters. Such models are called *sparse*.

LASSO provides a good statistical interpretation when the number of levels of existing factors is no more than two. Consider a factorial design with two categorical factors each with three levels. The linear model consists of six dummy variables with coefficients $(\beta_1, \beta_2, \beta_3)$ for the first factor and $(\beta_4, \beta_5, \beta_6)$ for the second factor. It is possible that it selects only levels corresponding to β_2 and β_6 because LASSO treats all dummy variables equally.

Yuan and Lin [18] proposed group LASSO for linear models with factors or group-wise predictors. The group LASSO estimate is defined as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \sum_{g=1}^{p} \left(\sqrt{l_g \sum_{i=1}^{l_g} \beta_i^2} \right),$$

where $g = 1, \ldots, p$ refers to factor indices, and l_g determines the number of levels of g-th factor. Group LASSO works well for disjoint grouped parameters because it selects or drops out groups of variables properly. However, it fails for overlapped groups, which appear in the penalty term when the model includes interactions and obeys hierarchy rules.

In the statistical literature, the fitted models often follow some hierarchical principles. Peixoto [21] used hierarchy in polynomial regressions in such a way that a higher order term is in the model only when lower order terms are in the model. Hamada [35] discussed hierarchy as a heredity principle. Nelder [19] called hierarchy as marginality. Heredity or marginality states that the presence of an interaction in the model is allowed only by presence of, some or all, related main effects. In this thesis we focus on hierarchical linear models applicable to high-dimensional factorial designs.

1.2 Overview of the thesis

There are two main approaches to guarantee hierarchy in fitted linear models: grouping of main effects and interactions in the penalty term; or using a hierarchy constraint.

The first approach makes a proper grouping or a set of groups of main effects and interactions in the penalty term. Consider a two way factorial design with factors A and B, each

with 3 levels, and interaction AB. The linear model is

$$y_{abk} = \mu + \alpha_a + \beta_b + (\alpha\beta)_{ab} + \varepsilon_{abk},$$

where $a = 1, 2, 3$ and $b = 1, 2, 3$. The proper grouping for satisfying hierarchy in the fitting of this linear model is $\{\{A\}, \{B\}, \{A, B, AB\}\}$; see Chapter 2 for more details. Therefore, the corresponding group LASSO penalty is

$$\sqrt{3\sum_a \alpha_a^2} + \sqrt{3\sum_b \beta_b^2} + \sqrt{15\left(\sum_a \alpha_a^2 + \sum_b \beta_b^2 + \sum_a \sum_b (\alpha\beta)_{ab}^2\right)}.$$

Suppose the first and second group of coefficients overlap with the third group; In such a situation, group LASSO could fail to select groups properly. The overlapped group LASSO proposed by Obozinsky et al. [34] is capable of yielding a hierarchical linear model for this case.

The second approach uses a new constraint rather than grouping of main effects and interactions in the penalty term. Its idea is derived from Cox [3], who stipulated that large main effects are more likely to lead to appreciable interactions. Based on this principle, Bien, Taylor, and Tibshirani [27] proposed a constraint to induce hierarchy in linear models. Consider a linear model with two factors A and B, each with two levels, and interaction AB. The linear model by baseline constraint and related dummy variables is as follows

$$y_i = \mu + \beta_A D_{Ai} + \beta_B D_{Bi} + \beta_{AB} D_{ABi} + \varepsilon,$$

where $D_{Ai} = D_A(y_i) = I(y_i \in 2^{\text{nd}} \text{ level of A})$ and $D_{ABi} = D_{Ai} * D_{Bi}$. The hierarchy constraints are then

$$|\beta_{AB}| \le |\beta_A| \quad \text{and} \quad |\beta_{AB}| \le |\beta_B|. \tag{1.2.1}$$

These constraints together with LASSO results in a hierarchical sparse model.

The advantages and disadvantages of both approaches are discussed in Chapter 2, which reviews and compares the existing proposals, to justify our preferred choice between them. In Section 2.2, we provide high-dimensional factorial examples to motivate and illustrate the application of our results.

The main results of this thesis are given in Chapter 3. We focus on the second approach, the approach based on constraints. We propose a modified generalized version of (2.5.7), to satisfy hierarchy in high-dimensional factorial designs. We investigate the effect of this constraint on estimates in Theorem 1. The proposed procedure combines group LASSO and LASSO penalties. Therefore, the solution lies in the set of $\mathcal{S}_{\text{LASSO}} \cap \mathcal{S}_{\text{group LASSO}}$ where $\mathcal{S}$ is the support of the related procedure. We characterize the solution in Theorem 2; and based on this characterization, we calculate the degrees of freedom for the proposed procedure in Theorem 3.

Chapter 4 is devoted to alternative approaches: the standardized group LASSO [28] and the overlapped group LASSO. Standardized group LASSO is a specific case of group LASSO with an orthonormalized design matrix. At first, we propose a normalization to unify group weights, which is investigated in Theorem 4. Then, we characterize the solution in Theorem 5; and calculate the unbiased estimate of degrees of freedom in Theorem 6. Finally, the unbiased estimate of degrees of freedom for the overlapped group LASSO is calculated in Corollary 1.

We apply the modified hierarchy constraint on generalized linear models with a binary response in Chapter 5. The solution is characterized in Theorem 7. The model selection

process of the proposed procedures are investigated in Section 5.5. The calculated estimates of degrees of freedom in this thesis are compared with other known estimates of the degrees of freedom in Section 5.5. Also, we compare the calculated degrees of freedom with the degrees of freedom of the corresponding ANOVA in the extreme case where the tuning parameter $\lambda \to 0$. We show in Theorem 8 and Corollary 2 that they are equivalent.

It seems the only connection investigated so far between classical ANOVA methodology and group LASSO is Lim's [17]. To establish such a connection, we need to make a connection between group LASSO with classical ANOVA concepts, such as factors, balanced design and replication. The parameters in classical ANOVA models are estimated by a constrained least-squares estimator, hereafter called constrained LSE. Chapter 6 further develops Lim's idea [17]; and shows a connection between group LASSO and constrained LSE in Theorem 9. It shows that group LASSO, in the case of balanced design, satisfies sum-to-zero constraints. Section 6.5 investigates the selection process of group LASSO; and shows how replication affects the selection process specifically in the case of balanced designs. Section 6.3 is devoted to the selection of λ. The usual methods for selection of λ are cross-validation, AIC and BIC. However, the calculation of AIC and BIC requires determining the quantity of the degrees of freedom, which are calculated in this thesis.

1.3 Basic concepts and definitions

1.3.1 Norm

A norm on a vector space V is a function $\|\cdot\| : V \longrightarrow \mathbb{R}$ such that for all $a \in \mathbb{R}$ and $u,\ v \in V$ has the following three properties

(i) absolute homogeneity: $\|av\| = |a|\|v\|$

(ii) subadditivity: $\|u + v\| \leq \|u\| + \|v\|$

(iii) zero vector: $\|v\| = 0 \Rightarrow v = 0$.

The following norms are used throughout the thesis. Define $\beta = (\beta_1, \ldots, \beta_p)$ then

- l_1-norm: $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_i|$.
- l_2-norm: $\|\beta\|_2 = \sqrt{\sum_{i=1}^{p} \beta_i^2}$.
- l_p-norm: $\|\beta\|_p = \left[\sum_{i=1}^{p} |\beta_i|^p\right]^{\frac{1}{p}}$ and $p > 1$.
- l_1/l_2-norms: Suppose $g_1, \cdots, g_p$ be the corresponding group of parameters each with size l_{g_i} and $d_{g_i} > 0$ be the weight of group g_i. The corresponding l_1/l_2-norm of $\beta = (\beta_{g_1}, \cdots, \beta_{g_p})$ is defined as

$$\|\beta\|_{1/2} = \sum_{i=1}^{p} \left(d_{g_i} \sqrt{\sum_{j=1}^{l_{g_i}} \beta_j^2} \right).$$

- Induced matrix norm: Suppose $\|\cdot\|_*$ is a vector norm; then, its induced matrix norm is defined as (Atkinson and Han [12], page 57)

$$\|A\|_* = \sup \{ \|Ax\|_* \mid x \in \mathbb{R}^n, \ \|x\|_* \leq 1 \} = \sup_{x \in \mathbb{R}^n, \, \|x\|_* \leq 1} \|Ax\|_*.$$

  Note that $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A)$ i.e. the induced l_2-norm of matrix A is its largest singular value or is the square root of the largest eigenvalue of $A^T A$.

1.3.2 Convex optimization

A set C is convex if the line segment between any two points in C lies in C, i.e., for any $x_1, \ x_2 \in C$ and any $\theta \in [0, 1]$

$$\theta x_1 + (1 - \theta) x_2 \in C.$$

A set C is called a cone, if for every $x \in C$ and $\theta \geq 0$, we have $\theta x \in C$. A set C is convex cone, if for any $x_1,\ x_2 \in C$ and $\theta_1,\ \theta_2 \geq 0$ we have

$$\theta_1 x_1 + \theta_2 x_2 \in C.$$

A cone $K \subseteq \mathbb{R}^n$ is called a proper cone if it satisfies the following:

- K is convex.
- K is closed.
- K has nonempty interior.
- K contains no line (equivalently $x \in K$ and $-x \in K \Rightarrow x = 0$).

Suppose $K \subseteq \mathbb{R}^n$ be a proper cone then, the partial ordering in $\mathbb{R}^n$ is defined as

$$x \preceq_K y \iff y - x \in K.$$

A function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex, if $\mathrm{dom} f$ is a convex set; and if for all $x, y \in \mathrm{dom} f$, and $\theta \in [0, 1]$, we have

$$f(\theta x + (1-\theta)y)) \leq \theta f(x) + (1-\theta) f(y).$$

Consider the optimization problem of the form

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0 \quad i = 1, \ldots, m \\ & h_i(x) = 0 \quad i = 1, \ldots, p, \end{aligned} \tag{1.3.1}$$

with variable $x \in \mathbb{R}^n$. Assume the domain $\mathcal{D} = \bigcap_{i=0}^{m} \mathrm{dom} f_i \cap \bigcap_{i=1}^{p} \mathrm{dom} h_i$ is nonempty, and denote the optimal value of (1.3.1) by p^*. The *Lagrangian* $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x),$$

with dom$L = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^p$. The Lagrange dual function $g : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is defined as

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right).$$

Note that the dual function yields lower bounds on the optimal value p^*:

$$g(\lambda, \nu) \leq p^*.$$

The dual function is concave, as it is the pointwise infimum of a family of affine functions of (λ, ν). The Lagrange dual problem is

$$\begin{aligned} \text{maximize} \quad & g(\lambda, \nu) \\ \text{subject to} \quad & \lambda \succeq 0, \end{aligned}$$

and the optimal value of the problem is denoted by d^*; therefore $d^* \leq p^*$. If strong duality holds, i.e., if $d^* = p^*$, then the optimal dual gap is zero. Slater's theorem states that if the problem is convex and there exists a strictly feasible point, i.e., $\exists x \in \mathcal{D}$ such that

$$f_i(x) < 0, \quad i = 1, \ldots, m, \quad h_i(x) = 0, \quad i = 1, \ldots, p,$$

then strong duality holds. Note that strong duality holds even with weaker condition, i.e., only some of inequality constraints are strict.

Suppose strong duality holds. Let x^* and (λ^*, ν^*) be the primal and dual optimal point. Then

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, \nu^*) \\ &= \inf_x \left[f_0(x) + \sum_{i=1}^{m} \lambda^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \right] \\ &\leq f_0(x^*) + \sum_{i=1}^{m} \lambda^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \leq f_0(x^*). \end{aligned}$$

This shows that

$$\sum_{i=1}^{m} \lambda^* f_i(x^*) = 0,$$

the condition called complementary slackness.

Suppose $f_i,\ i = 1,\dots,m$, are convex; $h_i,\ = 1,\dots,p$, are affine, and x^*, λ^*, ν^* be any points that satisfy the KKT conditions

$$\begin{aligned} f_i(x^*) &\le 0, \quad i = 1,\dots,m \\ h_i(x^*) &= 0, \quad i = 1,\dots,p \\ \lambda_i^* &\ge 0, \quad i = 1,\dots,m \\ \lambda_i^* f_i(x^*) &= 0, \quad i = 1,\dots,m \\ \nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) &= 0. \end{aligned}$$

Then x^* and (λ^*, ν^*) are primal and dual optimal, with zero duality gap.

1.3.3 Subdifferential

Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, be a convex function. Then $v \in \mathbb{R}^n$ is a subgradient of f at point $x_0 \in \text{dom} f$ if

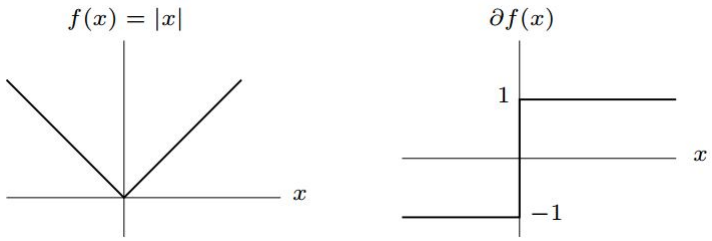
$$f(x) \ge f(x_0) + v^T(x - x_0) \quad \forall x \in \text{dom} f.$$

The subdifferential $\partial f(x_0)$ of f at x_0 is the set of all subgradients

$$\partial f(x_0) = \left\{ v \in \mathbb{R}^n \mid v^T(x - x_0) \le f(x) - f(x_0),\ \forall x \in \text{dom} f \right\}.$$

Suppose that $f(x) = |x|$, $x \in \mathbb{R}$, then

$$\partial f(x) = \frac{x}{|x|} \quad \text{if } x \neq 0, \qquad \partial f(x) = \{ v \in \mathbb{R} \mid |v| \le 1 \} = [-1, 1] \quad \text{if } x = 0.$$



Let f be the Euclidean norm $f(x) = \|x\|_2$ where $x \in \mathbb{R}^n$ then

$$\partial f(x) = \frac{x}{\|x\|_2} \quad \text{if } x \neq 0, \qquad \partial f(x) = \{v \in \mathbb{R}^n \mid \|v\|_2 \leq 1\} \quad \text{if } x = 0.$$

1.3.4 The degrees of freedom

In elementary textbooks, degrees of freedom refer to particular parameters in some distributions such as t, F and chi-square. In the classical ANOVA, the degrees of freedom of sum of squares of treatments and error are needed for calculating F-statistics. This leads to the definition of degrees of freedom for sum of squares; the definition which is a specific number of levels of factors. This concept generalizes to multiple regression involving p covariates; in such a case, the degrees of freedom of fit is p.

Suppose the vector of response $Y \in \mathbb{R}^n$ is normally distributed. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be the prediction rule, $\widehat{Y} = g(Y)$; the risk of g is then, Efron [6],

$$\text{Risk}(g) = \text{E}\,\|g(Y) - \mu\|_2^2 = \text{E}\,\|g(Y) - Y\|_2^2 - n\sigma^2 + 2\sum_{i=1}^{n} \text{cov}(g_i(Y), Y_i).$$

The last term in the right side of the equation can be taken for the following definition of degrees of freedom of g, suggested by Efron [6]:

$$\text{df}(g) = \frac{1}{\sigma^2}\sum_{i=1}^{n} \text{cov}(g_i(Y), Y_i). \tag{1.3.2}$$

We need μ for calculating of $\text{cov}(g_i(Y), Y_i)$ which is the parameter of interest in regression analysis. Also note that $E[g(Y)]$ depends on μ [30]. Therefore, this definition may not be suitable for practical calculations. Stein [24] proposed an unbiased estimate for $\text{df}(g)$ as

$$\text{df}(g) = E[(\triangledown.g)(Y)] \quad \Longrightarrow \quad \widehat{\text{df}}(g) = (\triangledown.g)(Y) = \sum_{i=1}^{n} \frac{\partial g_i(Y)}{\partial Y_i} \tag{1.3.3}$$

where g is continuous and almost everywhere differentiable. Based on this definition the Stein's unbiased risk estimation, SURE, is

$$\widehat{\text{Risk}}(g) = \|g(Y) - Y\|_2^2 - n\sigma^2 + 2\sigma^2 \widehat{\text{df}}(g). \tag{1.3.4}$$

The prediction rule in multiple regression is $g(Y) = X\hat{\beta}(Y)$ therefore

$$\frac{\partial g(Y)}{\partial Y} = X\frac{\partial \hat{\beta}(Y)}{\partial Y} = X\frac{\partial (X^TX)^{-1}X^TY}{\partial Y} = X\left(X^TX\right)^{-1}X^T.$$

This implies that

$$\widehat{\text{df}}(g) = (\triangledown.g)(Y) = \text{tr}(X\left(X^TX\right)^{-1}X^T) = \text{tr}(I_p) = p.$$

It shows that the degrees of freedom in least squares regression are equal to the number of predictors in the model.

1.3.5 The implicit function theorem

The following statement of implicit function theorem can be found in Yu [37].

Let $f: \mathbb{R}^n \longrightarrow \mathbb{R}^m$ with $n > m$. We decompose

$$\mathbb{R}^n = \mathbb{R}^{n-m} \times \mathbb{R}^m$$

and denote the first $n-m$ coordinates by vector x and the rest m coordinates by y. Assume

- f is differentiable and has continuous partial derivatives, (it can be considered in an interval);
- $f(x_0, y_0) = 0$;
- Jacobian $\frac{\partial f}{\partial y}(x_0, y_0)$ is invertible.

Then, there are open sets $U \subseteq \mathbb{R}^{n-m}$, $V \subseteq \mathbb{R}^m$ satisfying $x_0 \subseteq U, y_0 \subseteq V$ and

- For every $x \in U$ the equation $f(x, y) = 0$ has one unique solution $y = g(x) \in V$;
- $g(x_0) = y_0$;
- g is differentiable with continuous partial derivatives;
- For $x \in U$,

$$\frac{\partial g}{\partial x} = -\left(\frac{\partial f}{\partial y}\right)^{-1}\left(\frac{\partial f}{\partial x}\right).$$

Chapter 2

Hierarchy in Linear Models

2.1 Introduction

This chapter reviews existing approaches in groupwise regularization and their properties, to justify the preferred choice for constructing hierarchical sparse models. Unlike the following chapters, this chapter contains no technical results. However, we believe that it is important for understanding certain motivations, which determine our subsequent focus in this thesis.

There are two forms of hierarchy in linear models, weak and strong. Strong hierarchy refers to a rule in which the presence of an interaction is allowed only by presence of all related main effects. Weak hierarchy, on the other hand, requires only one of the related main effects. There are different opinions in this regard: Hamada et al. [35] suggest both weak and strong hierarchy for linear models; Nelder [19], on the contrary, suggests strong hierarchy as a default rule. Fitting linear models satisfying such rules is of interest in high-dimensional data sets [17], [36], [27] and [1].

As already mentioned in the previous chapter, there are two main approaches to achieve hierarchy in fitted linear models. The first approach (Zhao et al. [36] and Lim [17]) uses a specific geometric property of l_1/l_2-norms to satisfy hierarchy. The second approach uses

a constraint to achieve hierarchy which is inspired by Cox [3]. Bien, Taylor and Tibshirani [27] proposed a constraint, which together with LASSO results in a hierarchical sparse model. However, this constraint is restricted to factors at most with two levels.

This chapter starts with two real-data examples in section 2.2 and it is followed in Section 2.3 by definition of interaction in factorial designs. Section 2.4 establishes strong hierarchy as a default rule in this thesis by Nelder's argument. Section 2.5 reviews existing approaches, downward and upward grouping structures, Lim's theorem and the constraint of Bien et al. [27]. The modification of the constraint of Bien et al. [27] is given in Section 2.6; and we justify our preferred choice between two approaches for achieving hierarchy in Section 2.7.

2.2 Examples

2.2.1 A high-dimensional example

The data set of the genome-wide association study [17] deals with 26797 Single Nucleotide Polymorphism (SNP) markers. Each of those markers can be interpreted as 3-level categorical random variables. The data set contains 3500 training examples; apparently $n << p$ in this case. Even if we consider only main effects, classical ANOVA would not be applicable here. If we consider also pair interactions, the model would contain around 3.23×10^9 parameters. It is clear that working with such huge data requires its own method.

The number of sum-to-zero constraints for model with pair interactions are around 2.15×10^9. Fortunately, based on Theorem 9 and related results, the sum-to-zero constraints would be dropped out from group LASSO. Lim [17] designed the R package *GLINTERNET* to apply group LASSO on this data set with pair interactions. It selects the main

effects called SNP6-305 and denseSNP6-6873 together with pair interaction denseSNP6-6881 × denseSNP6-6882. The selected model is clearly not hierarchical. It is similar to our result in Table 6.4, where we will see that the standardized group LASSO drops out one of the effective main effects while it picks up the related pair interaction. In the genome-wide association study, Lim [17] showed with an F-test that the main effect denseSNP6-6882 is significant while it is not picked up by *GLINTERNET*. Thus, we still appear to need a method picking up a model satisfying the hierarchy principle automatically.

2.2.2 A toy example

We consider a two-factor layout from [35] and [11], in which the main effects and the interaction are statistically significant at the 95% confidence level. The design is balanced in order to facilitate the comparison of the group LASSO estimate with the constrained LSE.

A manufacturer found unwanted differences in the torque values of a lock nut that it made. Torque is the work, force×distance, required to tighten the nut. Consequently, the manufacturer conducted an experiment to determine which factors affected the torque values. The type of plating process was isolated as the most probable factor to impact torque, especially using no plating versus using plating. Another factor is the test medium, that is, whether the locknut is threaded onto a bolt or a mandrel. A mandrel is like a bolt but harder. Thus, the two experimental factors were

- type of plating, whose three levels were Cadmium and Wax denoted as C&W, Heat Threaded or no plating denoted as HT, and Phosphate and Oil denoted as P&O
- test medium, whose levels were mandrel and bolt.

The industry standard is 45-foot-pound maximum when the locknut is first threaded onto its mating partner as measured by a manual torque wrench. Table 2.1 shows torque data for the bolt experiment.

	C&W	HT	P&O
Bolt	20,16,17,18,15, 16,19,14,15,24	26,40,28,38,38, 30,26,38,45,38	25,40,30,17,16, 45,49,33,30,20
Mandrel	24,18,17,17,15, 23,14,18,12,11	32,22,30,35,32, 28,27,28,30,30	10,13,17,16,15, 14,11,14,15,16

Table 2.1: Torque data, bolt experiment, Wu [11].

2.3 Interactions in ANOVA models

It is common in additive models that main effects are inadequate for predicting the response variable. For instance, consider a two-factor layout where the effect of a factor on the response variable is associated with the levels of the other factor. In such a situation, the additive model cannot explain the response variable properly and we expect the presence of an interaction in the model. The other names for interaction in statistical literature are joint effect, cross-term, and compound term.

We consider pair interactions as the componentwise product of two factors, in the same way as Bien et al. [27], and Lim [17]. Let X_g denote the submatrix corresponding to the g-th factor with each row containing only a single 1 and other components are zero; such a matrix is called an indicator matrix (Lim [17]). We use indices $g \in \mathcal{G} = \{1, \ldots, p\}$ to refer to factors, where p is the number of factors. The cross-term matrix $X_{g:h}$ is the componentwise product of X_g and X_h, and corresponds to the interaction of the factors g and h.

For instance, consider two factors with two and three levels, respectively. A generic row of the matrix $X_{g:h}$ is given by

$$(a \quad b) * (c \quad d \quad e) = (ac \quad ad \quad ae \quad bc \quad bd \quad be), \tag{2.3.1}$$

where (a, b) and (c, d, e) are the corresponding generic rows of X_g and X_h.

Study of pair interactions in the case of high-dimensional data, $p >> n$, is challenging since there are $\binom{p}{2}$ pair interactions, the number that increases dramatically with p.

2.4 Hierarchy principles

"*Hierarchy*" in model construction and fitting means that only certain models are considered acceptable from the statistical point of view. In ANOVA this requirement may lead to quite complex conditions; in this thesis, we adopt a somewhat simplified view.

If a model with interaction is accompanied by all related main effects then we call it a model with strong hierarchy. We speak about weak hierarchy if it contains only one of the related main effects. Consider a model

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2. \tag{2.4.1}$$

The hierarchy rules for this model are as follows:

$$\begin{aligned} \text{Strong hierarchy}: \quad & \beta_{12} \neq 0 \implies \beta_1 \neq 0 \quad \text{and} \quad \beta_2 \neq 0 \\ \text{Weak hierarchy}: \quad & \beta_{12} \neq 0 \implies \beta_1 \neq 0 \quad \text{or} \quad \beta_2 \neq 0. \end{aligned}$$

Hierarchy is investigated by Hamada [35] and Chipman [2] who call it the heredity principle. Nelder [19], when studying strong hierarchy, speaks about the marginality principle. In the marginality principle, strong hierarchy is the default rule, since weak hierarchy occurs

actually in the so-called slope-ratio assays and the pure interaction model shows a surface with a saddle point at the origin. With this reasoning, pure interaction is too restrictive to be used practically; this establishes the hierarchy principle. However, the restrictive conditions for implication of weak hierarchy in practical situations lead us to consider the strong hierarchy as a default rule.

2.5 Existing approaches in hierarchical fitting

2.5.1 Group LASSO

In a factorial design, each factor in the related linear model corresponds to a group of dummy variables representing its levels. In such models, it is of interest to select or drop an entire factor instead of its particular level or levels. LASSO may not do it since LASSO treats all regressors equally. It may select only one level of a factor and drop other levels of that factor. Group LASSO is a solution to this problem.

The essence of the LASSO penalty is the use of the l_1-norm. The right panel of Figure 2.1 shows the unit ball of this penalty. When the surface of loss intersects with the surface of this ball, it is typically in the extreme points; the vector of parameter estimates thus gets many zeros, resulting in a sparse model. Ridge regression uses the l_2-norm penalty and the unit ball of this penalty is the standard Euclidean, round ball. It yields typically a full regression model; that is, all components of the vector of parameter estimates are nonzero. Yuan and Lin [18] used these two properties of l_1 and l_2-norms and proposed the group LASSO penalty,

$$\Omega_{group\ LASSO}(\beta) = \sum_{g \epsilon \mathcal{G}} d_g \|\beta_g\|_2,$$

where d_g is a groupwise weight. This penalty uses an l_1/l_2-norm where the l_1-norm favours sparsity while the l_2-norm favours the groupwise structure. For instance, consider $\beta = (\beta_1, \beta_2, \beta_3)$ with groups $\{\beta_1, \beta_2\}$ and $\{\beta_3\}$; then

$$\Omega_{LASSO}(\beta) = \sum_{i=1}^{3} |\beta_i| \quad \text{and} \quad \Omega_{group\ LASSO}(\beta) = \|(\beta_1, \beta_2)\|_2 + |\beta_3|.$$

The unit ball for this penalty is illustrated in the left panel of Figure 2.1. Compared to LASSO, the shape of the unit ball of group LASSO is in favour of selecting only $\{\beta_3\}$ on top (bottom) of the ball or selecting group $\{\beta_1, \beta_2\}$ as the circle in the middle of the ball. Note that groups $\{\beta_1, \beta_2\}$ and $\{\beta_3\}$ are disjoint and group LASSO works well for disjoint groups.

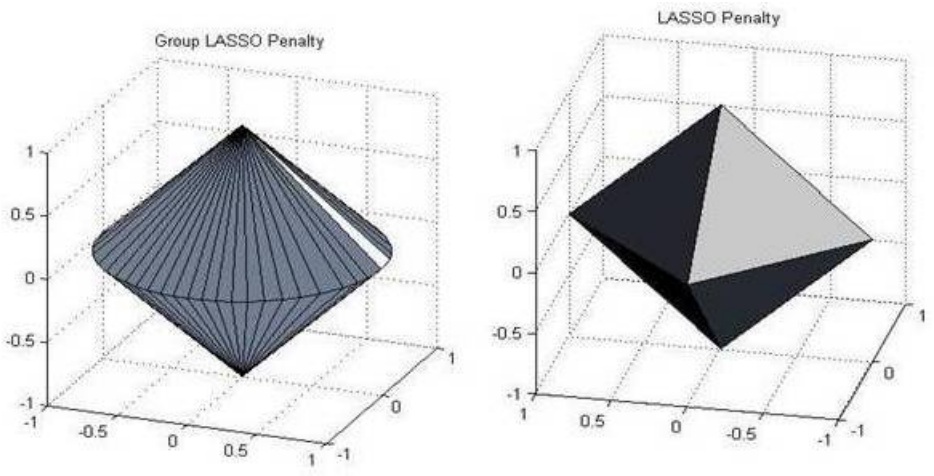


Figure 2.1: The unit balls for group LASSO and LASSO penalties.

2.5.2 CAP

As already mentioned, group LASSO uses the l_2-norm to encourage group structure in the model. The l_p-norm, where $p > 1$, can be used for the same purpose. Fu [10] penalized the negative log-likelihood by an l_p-norm and showed that by choosing a proper data specific

p, prediction is better in the sense of the MSE, the mean squared error.

Zhao et al. [36] used this fact and generalized group LASSO by using l_1/l_p-norms in order to get a better prediction in the sense of the MSE. They called it Composite Absolute Penalty, CAP. The penalty in this procedure is as follows:

$$\Omega_{CAP}(\beta) = \sum_{g \in \mathcal{G}} d_g \|\beta_g\|_p,$$

where $p > 1$. Therefore, the l_1-norm promotes sparsity while the l_p-norm encourages group structure in the model, and at the same time improves prediction error. Like group LASSO, CAP works well for disjoint groups.

Zhao et al. [36] used BLASSO algorithm to trace the solution path in the case of disjoint groups. They also tried, for the first time in statistical literature, to achieve hierarchy in linear models by designing a specific penalty. For this purpose, they illustrated hierarchical structure of parameters as a directed graph and tried to represent it as a set of groups. By designing directed graphs, they [36] argued that,

> "*[E]ach node corresponds to a group of variables g_k and set its descendants to be the groups that should only be added to the model after g_k.*"

To understand the aforementioned quote, consider the main effects A and B and interaction AB with a hierarchy such that AB is a descendant of A and B. Therefore, the descendant AB "*should only be added to the model after*" A. With this definition, since there is no descendant for AB, then $\varnothing$ "*should only be added to the model after*" AB. This gives the grand set $\mathcal{G} = \{\{A, AB\}, \{B, AB\}, \{AB\}\}$ which defines a downward algorithm in

grouping. They proposed the following penalty

$$\Omega_{CAP}(\beta) = \sum_{g \in \mathcal{G}} d_g \|(\beta_g, \beta_{all\ descendants\ of\ g})\|_p. \tag{2.5.1}$$

In this penalty, groups overlap since the descendants appear at least in three different groups. This penalty improves hierarchy in the fit but it does not result in a strong hierarchy. This means that some of interactions appear in the model without their main effects. Zhao et al. [36] used the term "hierarchy gap" to determine the number of variables that are missing to achieve a strong hierarchy.

2.5.3 Overlapped group LASSO

Group LASSO or CAP work well when the groups form a partition; if they do not, the problem with the "hierarchy gap", as observed by Zhao et al. [36], occurs.

Consider three overlapped groups in such a way that the first and third groups get zero coefficients by group LASSO. What remains, is not the second group, but rather the covariates in the second group which are not in the first or third groups. Therefore, we are looking for a penalty that selects groups entirely; that is, the support (the indices of nonzero components) of the solution $\hat{\beta}$ is a union of groups.

For this purpose, Obozinski et al. [34] proposed the following penalty

$$\Omega_{overlap}(\beta) = \inf_{\overline{v} \in V_{\mathcal{G}}, \sum_{g \in \mathcal{G}} v_g = \beta} \sum_{g \in \mathcal{G}} d_g \|v_g\|_2,$$

where $v_g,\ g \in \mathcal{G}$ are auxiliary variables; they represent a decomposition of β as their sum. Let us write this penalty in a simple example. Consider main effects A and B each with three levels and pair interaction AB. Also consider the three overlapped groups $g_1 = \{A\}$,

$g_2 = \{B\}$, and $g_3 = \{A, B, AB\}$, then $\mathcal{G} = \{\{A\}, \{B\}, \{A, B, AB\}\}$, hereafter called grand set. Therefore,

$$\Omega_{overlap}(\beta) = \sqrt{3}\|\alpha_A\|_2 + \sqrt{3}\|\alpha_B\|_2 + \sqrt{15}\sqrt{\|\tilde{\alpha}_A\|_2^2 + \|\tilde{\alpha}_B\|_2^2 + \|\alpha_{AB}\|_2^2},$$

where $\beta_A = \alpha_A + \tilde{\alpha}_A$, $\beta_B = \alpha_B + \tilde{\alpha}_B$, $\beta_{AB} = \alpha_{AB}$, and $\beta = \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_{AB} \end{bmatrix}$. That is, $v_{g_1} = \begin{bmatrix} \alpha_A \\ 0 \\ 0 \end{bmatrix}$, $v_{g_2} = \begin{bmatrix} 0 \\ \alpha_B \\ 0 \end{bmatrix}$, $v_{g_3} = \begin{bmatrix} \tilde{\alpha}_A \\ \tilde{\alpha}_B \\ \alpha_{AB} \end{bmatrix}$, and $\beta = \sum_{i=1}^{3} v_{g_i}$.

By applying an l_1/l_2 penalty to the vectors v_g, some of v_g shrink to zero, while the nonzero vectors satisfy $\sum_{g\in\mathcal{G}} v_g = \beta$. Therefore $\beta_i \neq 0$ as long as i belongs to at least one nonzero group. This shows that if a group is not dropped then all of its covariates have nonzero coefficients. In such a situation, overlapped group LASSO selects the groups.

Overlapped group LASSO can be used for achieving strong hierarchy. First, we review its mechanism in estimating parameters; then, with a simple example, we show how to achieve strong hierarchy with overlapped group LASSO. Consider three overlapping groups g_1, g_2, g_3 and let the overlapped group LASSO lead to an estimate such that $\beta_{g_1} = \beta_{g_2} = 0$. As mentioned above, in group LASSO the components of g_2 that are in g_1 or g_3 will get zero coefficients. However, in overlapped group LASSO all the components of g_2 will have nonzero coefficients. With this principle in mind about overlapped group LASSO, we want to design a penalty in such a way that strong hierarchy is guaranteed.

Example 1. *Consider main effects A and B with pair interaction AB. For observing strong hierarchy, if interaction AB is selected, then the main effects A and B should be selected*

too. This states that all these three effects are in a group, i.e. $\{A, B, AB\}$*. Therefore, the grand set* $\mathcal{G} = \{\{A\}, \{B\}, \{A, B, AB\}\} = \{g_1, g_2, g_3\}$ *will result in an additive model or a model with strong hierarchy. To observe this, suppose overlapped group LASSO leads to* $\alpha_A = 0$*; then* $\beta_A = \alpha_A + \tilde{\alpha}_A \neq 0$, $\beta_B = \alpha_B + \tilde{\alpha}_B \neq 0$, *and* $\beta_{AB} = \alpha_{AB} \neq 0$ *which is a model with strong hierarchy. The model is additive when the estimated coefficients of group* $g_3 = \{A, B, AB\}$ *are zero.*

This form of grouping shows an upward grouping, which opposes to the downward grouping. In fact, main effects are in the first line of hierarchy graph and pair interactions are in the second line. Similar to the equation (2.5.1), we define the penalty of overlapped group LASSO with strong hierarchy as follows

$$\Omega_{overlap}(\beta) = \inf_{\overline{v} \in V_{\mathcal{G}}, \sum_{g \in \mathcal{G}} v_g = \beta} \sum_{g \in \mathcal{G}} d_g \|(v_g, v_{all\ parents\ of\ g})\|_2, \tag{2.5.2}$$

where group g refers to the levels of main effects or interactions.

2.5.4 Upward and downward grouping

Zhao et al. [36] used a downward grouping in order to represent a hierarchy. Let us illustrate the effect of downward grouping in a simple example.

Example 2. *Consider main effects* A *and* B *with interaction* AB*. By a downward grouping, the grand set is* $\mathcal{G} = \{\{A, AB\}, \{B, AB\}, \{AB\}\}$ *and clearly the groups overlap. CAP uses a similar mechanism of group LASSO to deal with overlapping groups. If the estimate of group LASSO leads to* $\beta_{\{A,AB\}} = 0$*, then* $\beta_{AB} = 0$, $\beta_A = 0$, *and* $\beta_B \neq 0$*. This gives an additive model. The same is the case for the second group. Now, if* $\beta_{\{AB\}} = 0$*, then* $\beta_A \neq 0$ *and* $\beta_B \neq 0$ *which is again an additive model. The same is the case when two groups are zero.*

This states that if at least one of the groups gets zero estimate, then the model is additive. Note that the interaction is penalized three times in this grouping and, therefore, it is more likely to be dropped from the model. It shows that the model is always additive when a group is dropped. The behaviour of CAP and group LASSO in case of overlapped groups gets more complicated when the number of groups p increases. The simulations of Zhao et al. [36] showed that downward grouping does not guarantee strong hierarchy. Also downward grouping with overlapped group LASSO does not satisfy strong hierarchy. We check this fact by a simple example.

Example 3. *Consider main effects A and B with interaction AB. By a downward grouping, the grand set is $\mathcal{G} = \{\{A, AB\}, \{B, AB\}, \{AB\}\}$. If the estimate of overlapped group LASSO leads to $v_{\{A,AB\}} = 0$, then $\beta_B \neq 0$ and $\beta_{AB} \neq 0$ which is a model with a weak hierarchy rule. The same is the case for the second group. Now, if $v_{\{AB\}} = 0$, then $\beta_A \neq 0$, $\beta_B \neq 0$, and $\beta_{AB} \neq 0$. This satisfies the strong hierarchy rule. Now, if $v_{\{A,AB\}} = 0$ and $v_{\{B,AB\}} = 0$, then $\beta_{AB} \neq 0$ which is a pure interaction. This contradicts hierarchy.*

Two conclusions can be drawn from this example. Firstly, an overlapped group LASSO with a downward grouping penalty does not achieve hierarchy; secondly, if we penalize a group more, then it is more likely to appear in the model.

Now consider overlapped group LASSO with upward grouping, the same as (2.5.2).

Example 4. *Consider main effects A and B with interaction AB. By an upward grouping, the grand set is $\mathcal{G} = \{\{A\}, \{B\}, \{A, B, AB\}\}$. In this setting, the main effects are penalized two times while interaction is penalized one time. Therefore, the main effects are more likely to be in the model than the interaction. If the estimate of overlapped group LASSO leads to $v_{\{A\}} = 0$, then $\beta_A \neq 0$, $\beta_B \neq 0$, and $\beta_{AB} \neq 0$ which is a model with a strong hierarchy rule. The same is the case for the second group. Now, if $v_{\{A,B,AB\}} = 0$, then $\beta_A \neq 0$ and $\beta_B \neq 0$*

which is an additive model. If two groups get zero estimates, then again the final model preserves additivity or strong hierarchy.

In conclusion, downward grouping together with CAP or overlapped group LASSO do not satisfy strong hierarchy. Also upward grouping with CAP or group LASSO do not achieve strong hierarchy rule. Only upward grouping together with overlapped group LASSO results in a model with strong hierarchy or an additive model. There is no weak hierarchy for this case. Therefore upward grouping is the desirable grouping algorithm for the penalty of overlapped group LASSO. Also, we found that if we penalize a group more in group LASSO or CAP, it is more likely to be dropped out from the model. However, if we penalize a group more in overlapped group LASSO, it is more likely to appear in the model. We summarize this conclusion in the following proposition.

Proposition 1. *Overlapped group LASSO with upward grouping results in strong hierarchy. Also when a group is penalized more in overlapped group LASSO, it is more likely to be in the model.*

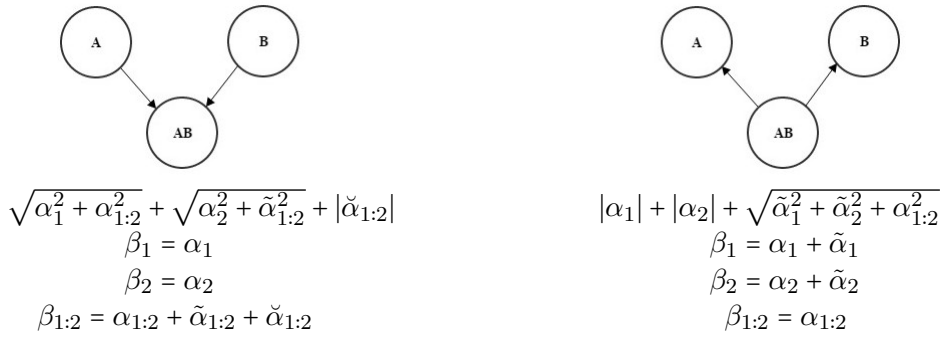


Figure 1.2: Downward and upward overlapped group LASSO penalty.

Now consider overlapped group LASSO with upward grouping where each group consists of only one variable. Let us investigate this case in high-dimensional data by an example.

Example 5. *Consider a linear model with X_1 and X_2 as continuous predictors. By upward grouping, the grand set is $\mathcal{G} = \{\{X_1\}, \{X_2\}, \{X_1, X_2, X_{1:2}\}\}$ where $X_{1:2} = X_1 X_2$. This*

leads us to the following overlapped group LASSO penalty with upward grouping of

$$\Omega_{overlapped}(\beta) = |\alpha_1| + |\alpha_2| + \sqrt{\tilde{\alpha}_1^2 + \tilde{\alpha}_2^2 + \alpha_{1:2}^2},$$

where $\beta_1 = \alpha_1 + \tilde{\alpha}_1$, $\beta_2 = \alpha_2 + \tilde{\alpha}_2$, *and* $\beta_{1:2} = \alpha_{1:2}$. *Note that the main effects are decomposed into two components and each component is penalized separately. If one of these components is not zero then the related main effect will have a nonzero coefficient. If we consider a model with* p *covariates, then each main effect decomposes into* p *components and each component is penalized separately.*

The results show that each main effect decomposes into p components and each component is penalized separately. Now, if a main effect is zero, then that means all of the p related components are zero. As a heuristic reason, suppose that $P(\{\mathrm{C}_i = 0\})$ is the probability of when the component C_i of a particular main effect is zero. Then, $P(\bigcap_{i=1}^{p}\{\mathrm{C}_i = 0\})$ is the probability of when all related components are zero. It gets smaller by increasing the dimension p of data set. Therefore, main effects are more likely to be in the model by increasing p. But, each interaction is penalized only one time and then it is less likely to be in the model. This reveals the weak side of overlapped group LASSO with upward grouping. All in all, representing a hierarchy rule by upward grouping in the case of high-dimensional data may result in misleading models.

2.5.5 Lim's theorem

Lim [17] investigated the overlapped group LASSO in order to induce a strong hierarchy rule in high-dimensional factorial designs. He considered full design matrix with sum-to-zero constraints in an overlapped group LASSO. By adding sum-to-zero constraints to overlapped group LASSO, the optimization problem becomes cumbersome. He simplified this problem to a group LASSO. He showed that a specific form of constrained overlapped

group LASSO is equivalent to a group LASSO. He considered a linear mode with factors X_1 and X_2 with corresponding number of levels L_1 and L_2, respectively, and quantitative response Y. The first order interaction model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{1:2} X_{1:2} + \varepsilon, \tag{2.5.3}$$

where $\beta_1 = (\beta_{11}, \cdots, \beta_{1L_1})^T$, $\beta_2 = (\beta_{21}, \cdots, \beta_{2L_2})^T$, $\beta_{1:2} = (\beta_{1:2,1}, \cdots, \beta_{1:2,L_1L_2})^T$, $\varepsilon \sim N(0, \sigma^2)$ and interaction $X_{1:2}$ is the componentwise product of X_1 and X_2, similar to equation (2.3.1). He defined the constrained overlapped group LASSO as follows:

$$\frac{1}{2}\left\| Y - \alpha_0 \cdot 1 - X_1\alpha_1 - X_2\alpha_2 - \begin{bmatrix} X_1 & X_2 & X_{1:2} \end{bmatrix} \begin{bmatrix} \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \alpha_{1:2} \end{bmatrix} \right\|_2^2$$

$$+\lambda\left(\|\alpha_1\|_2 + \|\alpha_2\|_2 + \sqrt{L_2\|\tilde{\alpha}_1\|_2^2 + L_1\|\tilde{\alpha}_2\|_2^2 + \|\alpha_{1:2}\|_2^2} \right)$$

subject to:

$$\sum_{i=1}^{L_1} \alpha_1^i = 0, \quad \sum_{j=1}^{L_2} \alpha_2^j = 0, \quad \sum_{i=1}^{L_1} \tilde{\alpha}_1^i = 0, \quad \sum_{j=1}^{L_2} \tilde{\alpha}_2^j = 0$$

$$\sum_{i=1}^{L_1} \alpha_{1:2}^{ij} = 0 \ \text{ for fixed } j, \qquad \sum_{j=1}^{L_2} \alpha_{1:2}^{ij} = 0 \ \text{ for fixed } i, \tag{2.5.4}$$

where $\beta_0 = \alpha_0$, $\beta_1 = \alpha_1 + \tilde{\alpha}_1$, $\beta_2 = \alpha_2 + \tilde{\alpha}_2$, and $\beta_{1:2} = \alpha_{1:2}$. First, he showed that $\bar{\tilde{\alpha}}_1 = \bar{\tilde{\alpha}}_2 = 0$ in the overlapped group LASSO without constraints; therefore, the first two constraints are satisfied in the estimates of effects $\hat{\alpha}_1$ and $\hat{\alpha}_2$. In the second step, he proved that by adding the second intercept $\tilde{\alpha}_0$ to the convex problem (2.5.4), the problem would not change since its estimate would be $\hat{\tilde{\alpha}}_0 = 0$. Finally, he considered the following interaction decomposition

$$\begin{aligned} \beta_{1:2}^{ij} &= \beta_{1:2}^{\cdot\cdot} + (\beta_{1:2}^{i\cdot} - \beta_{1:2}^{\cdot\cdot}) + (\beta_{1:2}^{\cdot j} - \beta_{1:2}^{\cdot\cdot}) + (\beta_{1:2}^{ij} - \beta_{1:2}^{i\cdot} - \beta_{1:2}^{\cdot j} + \beta_{1:2}^{\cdot\cdot}) \\ &\equiv \tilde{\alpha}_0 + \tilde{\alpha}_1^i + \tilde{\alpha}_1^j + \tilde{\alpha}_{1:2}^{ij} \end{aligned} \tag{2.5.5}$$

and showed that

$$\|\beta_{1:2}\|_2^2 = L_1 L_2 \|\tilde{\alpha}_0\|_2^2 + L_2 \|\tilde{\alpha}_1\|_2^2 + L_1 \|\tilde{\alpha}_2\|_2^2 + \|\alpha_{1:2}\|_2^2.$$

This reduces the above constrained overlapped group LASSO to the following unconstrained optimization problem

$$\frac{1}{2} \|Y - \alpha_0 \cdot 1 - X_1\alpha_1 - X_2\alpha_2 - X_{1:2}\beta_{1:2}\|_2^2 + \lambda \left(\|\alpha_1\|_2 + \|\alpha_2\|_2 + \|\beta_{1:2}\|_2 \right),$$

which is a group LASSO. In fact Lim's theorem says that we can use a group LASSO to obtain some fits that obey strong hierarchy. Group LASSO can estimate $\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2$ and $\hat{\alpha}_{1:2}$, but $\hat{\tilde{\alpha}}_1$ and $\hat{\tilde{\alpha}}_2$ cannot be estimated directly by a group LASSO and note that the overlapped group LASSO estimates are

$$\begin{aligned}
\hat{\beta}_0 &= \hat{\alpha}_0, \\
\hat{\beta}_1 &= \hat{\alpha}_1 + \hat{\tilde{\alpha}}_1, \\
\hat{\beta}_2 &= \hat{\alpha}_2 + \hat{\tilde{\alpha}}_2, \\
\hat{\beta}_{1:2} &= \hat{\alpha}_{1:2}.
\end{aligned}$$

It is not completely clear how to estimate $\hat{\tilde{\alpha}}_1$ and $\hat{\tilde{\alpha}}_2$. Note that if we consider $\hat{\tilde{\alpha}}_1 = \hat{\beta}^{i\cdot}_{1:2} - \hat{\beta}^{\cdot\cdot}_{1:2}$ from the decomposition of interaction (2.5.5), then $\beta^{i\cdot}_{1:2}$ and $\beta^{\cdot\cdot}_{1:2}$ have to be zero by assumed constraints. Then, $\hat{\tilde{\alpha}}_1 = 0$ and with the same reason, $\hat{\tilde{\alpha}}_2 = 0$. In this situation, the group LASSO may result in pure interaction and this contradicts a strong hierarchy rule. Note that the group LASSO does not satisfy a strong hierarchy rule and so the nonzero estimates of $\hat{\tilde{\alpha}}_1$ and $\hat{\tilde{\alpha}}_2$ are needed.

In the proof of this theorem, each sum-to-zero constraint is decomposed to latent variables. For instance, consider the first constraint on parameters for full design model (2.5.3),

$\sum_{i=1}^{L_1} \beta_1^i = 0$. From (2.5.4), $\beta_1 = \alpha_1 + \tilde{\alpha}_1$, which gives $\sum_{i=1}^{L_1} \beta_1^i = \sum_{i=1}^{L_1} (\alpha_1^i + \tilde{\alpha}_1^i) = 0$. However, this constraint is decomposed to two constraints $\sum_{i=1}^{L_1} \alpha_1^i = 0$ and $\sum_{i=1}^{L_1} \tilde{\alpha}_1^i = 0$. Now, if we consider a high-dimensional data set with p factors, then this constraint will extend to p constraints. This theorem extends the sum-to-zero constraints wherever they are needed. Nelder [20] believes that putting constraints on parameters in linear models is unnecessary because these constraints are on the estimates of parameters not on the parameters themselves. He argues:

> "*It is tempting to match the symmetric constraints, say, on the parameter estimates with corresponding constraints* $\alpha_\cdot = \beta_\cdot = \gamma_{i\cdot} = \gamma_{\cdot j} = 0$ *on the parameters themselves. This temptation should be resisted, ...*"

But this temptation is usually not resisted by statisticians and constraints are used for identifiability. Note that these constraints on estimates of parameters are correct only on balanced designs.

Lim's theorem reveals an inside group balancing in the penalty term. Consider the linear model (2.5.3). The penalty of an overlapped group LASSO with upward grouping is $\sqrt{\|\tilde{\alpha}_1\|_2^2 + \|\tilde{\alpha}_2\|_2^2 + \|\alpha_{1:2}\|_2^2}$. This penalty is replaced with $\sqrt{L_2\|\tilde{\alpha}_1\|_2^2 + L_1\|\tilde{\alpha}_2\|_2^2 + \|\alpha_{1:2}\|_2^2}$ in Lim's theorem. The coefficients L_1 and L_2 appear in the penalty term because of the number of components in each effect. It represents an inside group balancing. In other words, $\tilde{\alpha}_{1:2}$ has L_1L_2 components while $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ have L_1 and L_2 components, respectively. Therefore, coefficients L_1 and L_2 make a balance between main effects and interaction in the group. Usually, groupwise weights d_g are used for balancing among groups not inside groups. This is an interesting form of group weights.

2.5.6 The constraint of Bien et al.

Bien, Taylor and Tibshirani [27] interpreted the hierarchy structure of a linear model as a constraint. The idea is inspired by Cox [3]:

> "*[L]arge component main effects are more likely to lead to appreciable interactions than small components. Also, the interactions corresponding to larger main effects may be in some sense of more practical importance.*"

That means, the interactions related to larger main effects are more likely to be included in the model. Hence, the model focuses on investigating of the interactions with larger main effects. In other words, if a main effect is zero then all related interactions will be zero. In hierarchical LASSO [27], the LASSO penalty promotes sparsity and hierarchy is guaranteed by a constraint, which relies on a statistical principle rather than geometric interpretation of upward grouping.

Bien et al. [27] consider a regression model with continuous outcome Y, predictors $X_1, \ldots, X_p$ and pairwise interactions among predictors. Afterwards, they define the following model

$$Y = \beta_0 + \sum_j \beta_j X_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} X_j X_k + \varepsilon, \tag{2.5.6}$$

where $\varepsilon \sim N(0, \sigma^2)$, $\beta \in \mathbb{R}^p$ and $\Theta \in \mathbb{R}^{p \times p}$ where $\Theta_{jj} = 0$. Note that Θ is a matrix which contains the coefficient of interactions. The coefficient of one half, before the interaction terms, is used because of the notation of interactions as a matrix rather than a vector of length $p(p-1)$. The proposed optimization problem to satisfy the strong hierarchy rule is as follows:

$$\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \Theta \in \mathbb{R}^{p \times p}} L(\beta_0, \beta, \Theta) + \lambda \|\beta\|_1 + \frac{\lambda}{2} \|\Theta\|_1$$

$$\text{s.t.} \qquad \Theta = \Theta^T, \ \|\Theta_j\|_1 \leq |\beta_j| \quad \text{for } j = 1, \ldots, p, \tag{2.5.7}$$

where $\|\Theta_j\|_1 = \sum_{k=1}^{p} \|\Theta_{jk}\|_1$. Note that the objective function is an *all pairs LASSO* [27]. Now, if $\beta_j = 0$, then $\|\Theta_j\|_1 = 0$. This makes all those interactions zero that are related to the jth main effect. The constraint $\|\Theta_j\|_1 - |\beta_j|$ is not convex and, therefore, it may need to be changed to $\|\Theta_j\|_1 \leq \beta^+ + \beta^-$. One may ask why not to use constraint $|\Theta_{jk}| \leq \beta_j^+ + \beta_j^-$? Bien et al. [27] answer:

"*[This] can lead to an overabundance of interactions relative to main effects.*"

The constraint $\Theta = \Theta^T$ means $\Theta_{jk} = \Theta_{kj}$ which results in a strong hierarchy rule, because if $\Theta_{jk} \neq 0$, then $\beta_j \neq 0$ and since $\Theta_{jk} = \Theta_{kj} \neq 0$, then $\beta_k \neq 0$.

The methodology, as proposed, is applicable on linear models with factors having only two levels; the latter become a univariate predictor by dropping out one level.

2.6 Modifying the constraint of Bien et al.

Based on Cox [3], interactions related to larger main effects are more likely to be included in the model. In a *radical* interpretation, interactions related to smaller main effects are less likely to be in the model. Thus, when a main effect has a zero coefficient, then all related interactions have zero coefficients, i.e., $\beta_j = 0 \ \Rightarrow \ \Theta_{jk} = 0$ for $k = 1, \cdots, p$. This leads to the strong hierarchy rule. In a *moderate* interpretation, a pair interaction with only one large main effect is likely to be included in the model which leads to weak hierarchy rule. For the principle "if a main effect is zero then its related interactions are zero"; the constraint $|\Theta_{jk}| \leq |\beta_j|$ is the most simple one to represent it. Bien et al. [27] argue that using this constraint results in "*overabundance of interactions relative to main effects*". Therefore they proposed $\|\Theta_j\|_1 \leq |\beta_j|$ which represents exactly the *radical* interpretation.

In high-dimensional data sets, it is common to face factors which have more than two levels [17]. To deal with such a situation, we consider each factor as a group of its levels in the model and pair interactions are componentwise products of levels of two related factors as defined in (2.3.1). To guarantee hierarchy, we propose the constraint $\|\Theta_g\|_1 \leq \|\beta_g\|_1 \quad \text{for } g = 1, \ldots, p$, where Θ is the matrix of interaction coefficient (3.2.1), Θ_g is the g-th rows of Θ; and $\|\Theta_g\|_1 = \sum_{h=1}^{p} \sum_{i\in g, j\in h} |\Theta_{g:h}^{ij}|$. It reduces to the constraint of Bien et al. [27] in the case of $l_g = 1$, for all $g \in \mathcal{G}$. It follows a similar structure and satisfies the *radical* interpretation of Cox's inspiration. Note that $(\beta_g = 0 \Leftrightarrow \|\beta_g\|_1 = 0)$ and $(\Theta_g = 0 \Leftrightarrow \|\Theta_g\|_1 = 0)$. Hence, the modified constraint gives $(\|\beta_g\|_1 = 0 \Rightarrow \|\Theta_g\|_1 = 0)$. On the other side, when $\hat{\Theta}_{g:h} \neq 0$ then $\|\hat{\Theta}_g\|_1 > 0$ and $\|\hat{\Theta}_h\|_1 > 0$ therefore it gives $\hat{\beta}_g \neq 0$ and $\hat{\beta}_h \neq 0$.

The modified constraint $\|\Theta_g\|_1 \leq \|\beta_g\|_1$ leads the group LASSO to select groups based on the strong hierarchy rule. But in the selected groups, a further inside group selection may occur; this is discussed in Theorem 1. So we change the base inequality to $\|\Theta_g\|_2 \leq \|\beta_g\|_2$ to keep the group structure. Note that $\|\Theta_g\|_2$ is not separable for all $\Theta_{g:h}$, $h \in \mathcal{G}$. Fortunately $\|\Theta_g\|_2 \leq \sum_{h\in\mathcal{G}} \|\Theta_{g:h}\|_2$ and by considering $\sum_{h\in\mathcal{G}} \|\Theta_{g:h}\|_2 \leq \|\beta_g\|_2$, the hierarchy constraint changed to a stronger constraint. By this argument, the following two constraints are available,

- $\sum_{h\in\mathcal{G}} \|\Theta_{g:h}\|_2 \leq \|\beta_g\|_2$
- $\|\Theta_g\|_1 \leq \|\beta_g\|_1$.

The first constraint induces hierarchy, but it is not convex. The second constraint achieves hierarchy; and by a relaxation, it can be made convex. However, it reveals some inside group selection. We adopt the second constraint and develop it in Section 3.2.

2.7 Justifying the choice: discussion

We introduced two frameworks for inducing strong hierarchy in linear models.

(i) upward grouping;

(ii) constraint.

Upward grouping together with the overlapped group LASSO is used to construct hierarchical linear models for factorial designs. Lim [17] used this idea and showed that a specific constrained overlapped group LASSO reduces to a group LASSO. Lim's theorem [17] uses sum-to-zero constraints. They are zero only in the case of balanced designs. There is no balanced design in high-dimensional data and assuming such perfect condition on high-dimensional data requires a clear discussion. The other problem is that Lim's theorem needs to extend these constraints on latent variables. This problem is discussed in Section 2.5. If there are p factors or main effects, then a sum-to-zero constraint on a specific main effect will decomposed into p constraints on p related latent variables.

Lim's approach has some merit, but we will not use upward grouping with overlapped group LASSO because of the three disadvantages outlined below:

- Upward grouping directly relies on a geometric interpretation of the l_1-norm for satisfying hierarchy. Note that hierarchy and sparsity are different concepts with different solutions. For sparsity we minimize $\sum_i f(\beta_i)$ and drop small coefficients. In such a situation, the small coefficients will have a zero value by using the l_1-norm. This is an indirect usage of the l_1-norm. This idea is generalized for group LASSO. It minimizes $\sum_g f(\beta_g)$ and drops groups with small coefficients. In fact, an l_1/l_2-norm is used to drop groups with small coefficients. Again, this is an indirect usage of

the l_1-norm. For achieving hierarchy, we try to make groups of parameters in such a way that if a group gets zero estimate, then the other nonzero groups satisfy strong hierarchy or additivity. Note that we do not minimize $Loss + \sum_g f(\beta_g)$ to get hierarchy; we minimize it for sparsity. Then, hierarchy relies directly on the geometric interpretation of the l_1/l_2-norms.

- Main effects and interactions are not treated equally in an overlapped group LASSO with upward grouping. Each main effect is penalized p times while each pair interaction is penalized only one time. Therefore, main effects are more likely to be selected and interactions are less likely to be in the model. Therefore, it is more likely to lead to additive models.

- Upward grouping with an overlapped group LASSO increases the number of columns in the design matrix. Consider 50 factors each with three levels, together with pair interactions; then, the design matrix has 11175 columns. Now, considering the overlapped group LASSO with upward grouping, the related design matrix has 18525 columns.

On the other hand, as explained in Section 2.4, weak hierarchy is restricted to slope-ratio assays, and such experiments are rare in statistical literature. Therefore strong hierarchy is the default rule. All in all, we adopt and develop the constraint of Bien et al. [27] for factorial designs with a strong hierarchy rule.

Chapter 3

Hierarchical Group LASSO with Quadratic Loss

3.1 Introduction

In this chapter, we study the modification and generalization of the hierarchy constraint of Bien, Taylor, and Tibshirani [27], aimed at achieving hierarchical LASSO fits applicable to factorial designs with factors having possibly more than two levels. Based on the discussion in the previous chapter, we are interested in linear models obeying the strong hierarchy rule.

The effect of the constraint on the main effects and interactions is studied in Theorem 1. The solution of the proposed convex problem is characterized in Theorem 2. The unbiased estimate of degrees of freedom is given in Theorem 3.

3.2 The proposed method

Let us define the linear model with pair interactions for response variable Y and factors $X_1, \ldots, X_g, \ldots, X_p$ as follows:

$$Y = \beta_0 + \sum_{g=1}^{p} X_g \beta_g + \frac{1}{2} \sum_{g \neq h} X_{g:h} \operatorname{vec}(\Theta_{g:h}) + \varepsilon, \tag{3.2.1}$$

where $\varepsilon \sim N(0, \sigma^2 I)$ and $\Theta_{g:h}$ is the interaction coefficient matrix

$$\Theta_{g:h} = \begin{bmatrix} \Theta_{11} & \dots & \Theta_{1l_h} \\ \vdots & \ddots & \vdots \\ \Theta_{l_g 1} & \cdots & \Theta_{l_g l_h} \end{bmatrix}_{l_g \times l_h},$$

where $\beta^T = (\beta_1^T, \dots, \beta_p^T)$, $\Theta = (\Theta_{g:h})_{L\times L}$, $\Theta_{g:g} = 0$, $L = \sum_g l_g$, and $X_{g:h}$ is defined by (2.3.1). The coefficient of one half is the result of our notation for interaction coefficients as a symmetric matrix. In the model (3.2.1) each factor X_g is considered as a group of l_g dummy variables. We want a sparse model such that if a factor is not in the model, then $\beta_g = 0$ and if an interaction is not in the model, then $\Theta_{g:h} = 0$ and $\Theta_{h:g} = 0$. Constructing such sparse models leads us to group LASSO which we define for the model (3.2.1) as

$$\min_{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^L,\Theta\in\mathbb{R}^{L\times L}} \mathcal{L}(\beta_0, \beta, \Theta) + \lambda \sum_g \|\beta_g\|_2 + \frac{\lambda}{2} \sum_{g\neq h} \|\Theta_{g:h}\|_F, \tag{3.2.2}$$

where $\mathcal{L}(\cdot)$ is the quadratic loss. This model only includes main effects and interactions as groups, and hierarchy is not considered yet. To guarantee hierarchy, we propose the constraint $\|\Theta_g\|_1 \le \|\beta_g\|_1 \quad$ for $g = 1, \dots, p$, where Θ_g is the g-th rows of Θ and $\|\Theta_g\|_1 = \sum_{h=1}^p \sum_{i\in g, j\in h} |\Theta_{g:h}^{ij}|$. The resulting optimization problem is

$$\min_{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^L,\Theta\in\mathbb{R}^{L\times L}} \mathcal{L}(\beta_0, \beta, \Theta) + \lambda \sum_g \|\beta_g\|_2 + \frac{\lambda}{2} \sum_{g\neq h} \|\Theta_{g:h}\|_F$$

$$\Theta = \Theta^T, \quad \|\Theta_g\|_1 \le \|\beta_g\|_1 \quad \text{for all } g \in \mathcal{G}. \tag{3.2.3}$$

The added constraint $\|\Theta_g\|_1 - \|\beta_g\|_1$ is not convex because of the negative coefficient of $\|\beta_g\|_1$. Note that $\|\beta_g\|_1$ is equivalent to $\beta_g^+ + \beta_g^-$ where $\beta_g^+ \ge 0$, $\beta_g^- \ge 0$ and $\beta_g^+ * \beta_g^- = 0$. By relaxing $\beta_g^+ * \beta_g^- = 0$, the constraint becomes convex and we have

$$\min_{\beta_0\in\mathbb{R},\beta^\pm\in\mathbb{R}^L,\Theta\in\mathbb{R}^{L\times L}} \mathcal{L}(\beta_0, \beta^+ - \beta^-, \Theta) + \lambda \sum_g \|\beta_g^+ - \beta_g^-\|_2 + \frac{\lambda}{2} \sum_{g\neq h} \|\Theta_{g:h}\|_F,$$

$$\left.\begin{array}{c} \|\Theta_g\|_1 \leq 1^T(\beta_g^+ + \beta_g^-) \\ \beta_g^+ \geq 0, \quad \beta_g^- \geq 0 \end{array}\right\} \quad \text{for all } g \in \mathcal{G}, \tag{3.2.4}$$
$$\Theta = \Theta^T,$$

where $\beta_g \in \mathbb{R}^{l_g}$ and β_g is the difference of the two vectors $\beta_g^+, \beta_g^- \in \mathbb{R}^{l_g}$, i.e. $\beta_g = \beta_g^+ - \beta_g^-$. This relaxation has a specific effect on the hierarchy constraint. In fact, it is possible to have solutions in (3.2.4) such that both $\hat{\beta}_g^+$ and $\hat{\beta}_g^-$ are strictly positive. Hence, this constraint can cover larger interactions than the main effects, in which case both $\hat{\beta}_g^+$ and $\hat{\beta}_g^-$ can be taken to be large.

If we add up the constraint used in (3.2.3) for all groups, we will have $\sum_{g\in\mathcal{G}} \|\Theta_g\|_1 \leq \sum_{g\in\mathcal{G}} \|\beta_g\|_1$. In this situation, interactions are considered two times. Hence, we propose the new constraint as $\frac{1}{2}\|\Theta_g\|_1 \leq \|\beta_g\|_1$. That is, the main effect β_g is greater than half of the total of the interactions of factor g with the other factors. However, this constraint will result in the lack of interactions. Note that Θ_g has l_g rows and β_g has l_g components which shows a balance. Note that Θ_g has $L = \sum_{g=1}^{p} l_g$ columns. If $\beta_g = 0$, then $p-1$ interaction matrices $\Theta_{g:h}$ have to be zero that is, $\beta_g = 0$ results in $l_g L$ zero components which shows an imbalance. Then, we propose a normalization in each interaction. Suppose w_g to be a vector of l_g repeated l_g times and $w = (w_g)_{g\in\mathcal{G}}$. We propose $\frac{1}{2}\|\Theta_g \text{diag}(\frac{1}{w})\|_1 \leq \|\beta_g\|_1$.

We can see that problem (3.2.4) is feasible, since at least zero is a feasible point. Any l_1/l_2-norm is a convex function by definition and since the objective is a positive linear combination of norms, it is convex. The equality and positivity inequality constraints are convex. Finally, we need to show that $\frac{1}{2}\|\Theta_g \text{diag}(\frac{1}{w})\|_1 - 1^T(\beta_g^+ + \beta_g^-)$ is convex. Suppose $(\Theta_1, \beta_1^+, \beta_1^-)$ and $(\Theta_2, \beta_2^+, \beta_2^-)$ are two points such that they satisfy the inequality and let

$\alpha \in [0,1]$. We have

$$\frac{1}{2}\|\alpha\Theta_1 \text{diag}(\frac{1}{w}) + (1-\alpha)\Theta_2 \text{diag}(\frac{1}{w})\|_1 - 1^T\left[\alpha(\beta_1^+ + \beta_1^-) + (1-\alpha)(\beta_2^+ + \beta_2^-)\right] \leq$$

$$\alpha\left[\frac{1}{2}\|\Theta_1 \text{diag}(\frac{1}{w})\|_1 - 1^T(\beta_1^+ + \beta_1^-)\right] + (1-\alpha)\left[\frac{1}{2}\|\Theta_2 \text{diag}(\frac{1}{w})\|_1 - 1^T(\beta_1^+ + \beta_1^-)\right].$$

Therefore, the problem (3.2.4) is a convex optimization problem. For ease of computation, we write $|\Theta|$ in terms of Θ^+ and Θ^-. For notational simplicity, let $\phi = (\beta^+, \beta^-, \text{vec}(\Theta^+), \text{vec}(\Theta^-))$ where $\Theta_{g:g} = 0$ and suppose $\widetilde{X} = (X; -X; Z; -Z)$ where Z is the related matrix of interactions. Then, the strong hierarchical group LASSO can be written as

$$\min_{\beta_0 \in \mathbb{R}, \beta^\pm \in \mathbb{R}^L, \Theta^\pm \in \mathbb{R}^{L\times L}} \frac{1}{2}\left\|Y - \tilde{X}\phi\right\|_2^2 + \lambda \sum_g \|\beta_g^+ - \beta_g^-\|_2 + \lambda \sum_{g\neq h} \|\Theta_{g:h}^+ - \Theta_{g:h}^-\|_F,$$

$$\left.\begin{array}{c} 1^T \text{vec}((\Theta_g^+ + \Theta_g^-)\text{diag}(\frac{1}{w})) \leq 1^T(\beta_g^+ + \beta_g^-) \\ \beta_g^\pm \geq 0, \quad \Theta_g^\pm \geq 0 \end{array}\right\} \quad \text{for all } g \in \mathcal{G}, \tag{3.2.5}$$

$$\Theta^+ - \Theta^- = \Theta^{+T} - \Theta^{-T}.$$

Note that here $\frac{\Theta_g}{2}$ is considered as a new parameter of Θ_g. By decomposing $\|\Theta\|_1$ into Θ^+ and Θ^-, the problem is not changed since $\|\Theta\|_1$ appears in the objective function with a positive coefficient. Therefore $\hat{\Theta}^\pm = \max\left\{\pm\hat{\Theta}, 0\right\}$.

Problem (3.2.5) is a convex problem, but it is not strictly convex to guarantee uniqueness of solution. Tibshirani [30] discussed the case when the number of predictors p exceeds the number of observations; then $\text{rank}(X) < p$ and, thus, there can be multiple minimizers for the LASSO problem. He discussed the conditions which result in a unique solution, specifically when X has entries drawn from a continuous probability distribution. Roth and Fischer [8] studied non-uniqueness in group LASSO and proposed an algorithm to guarantee uniqueness of the solution.

Zou and Hastie [13] found that the LASSO problem gets a unique solution by adding the elastic net term to a LASSO penalty. At first note that the l_2-norm, $\|\cdot\|_2^2$, is strictly convex. Hence, penalizing a group LASSO problem with a tiny fraction of such norm will ensure the uniqueness of the solution. We consider the elastic net penalty $\frac{\varepsilon}{2}(\|\beta^+\|_2^2+\|\beta^-\|_2^2+\|\Theta\|_F^2)$, similar to Bien et al. [27], where ε is a fixed tiny fraction of λ, say, $10^{-8}\lambda$. Note that the case $\hat{\Theta}^+ > 0$ and $\hat{\Theta}^- > 0$ cannot happen. Hence, at least one of the estimates $\hat{\Theta}^+$ and $\hat{\Theta}^-$ must be zero and, therefore, $\|\Theta^+\|_F^2 + \|\Theta^-\|_F^2 = \|\Theta\|_F^2$. The loss function in problem (3.2.5) together with the elastic net penalty, $\frac{\varepsilon}{2}(\|\beta^+\|_2^2 + \|\beta^-\|_2^2 + \|\Theta\|_F^2)$, is equivalent to replacing $\tilde{X}$ and Y by

$$\tilde{X}_\varepsilon = \begin{bmatrix} X & -X & Z & -Z \\ \sqrt{\varepsilon} I_{|X|} & 0 & 0 & 0 \\ 0 & \sqrt{\varepsilon} I_{|X|} & 0 & 0 \\ 0 & 0 & \sqrt{\varepsilon} I_{|Z|} & -\sqrt{\varepsilon} I_{|Z|} \end{bmatrix} \quad \text{and} \quad Y_\varepsilon = \begin{bmatrix} Y \\ 0_{(2|X|+|Z|)\times 1} \end{bmatrix},$$

where $|\cdot|$ refers to cardinality. Therefore, the proposed procedure for the hierarchical group LASSO with a quadratic loss is the problem (3.2.5) where Y and $\tilde{X}$ are replaced with Y_ε and $\tilde{X}_\varepsilon$.

3.3 The effect of the hierarchy constraint

In this section, we investigate the effect of hierarchy constraint on the estimates of main effects β and interactions Θ; we want to see how this constraint affects the selection process of main effects and interactions. We need to find the solution to the convex problem (3.2.5).

Theorem 1. *The solution of the convex problem (3.2.5) satisfies*

- *main effects*

$$\lambda \frac{\hat{\beta}_j^+ - \hat{\beta}_j^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} = \mathscr{S}\left(-X_j^T(Y - \tilde{X}\hat{\phi}), -\hat{\alpha}_g\right) \quad \textit{if} \quad \hat{\beta}_g^+ - \hat{\beta}_g^- \neq 0 \quad \textit{and} \quad j \in g,$$

$$\left\| X_g^T(Y - \tilde{X}\hat{\phi}) - sgn(X_g^T(Y - \tilde{X}\hat{\phi}))\hat{\alpha}_g \right\|_2 \leq \lambda \quad if \quad \hat{\beta}_g^+ - \hat{\beta}_g^- = 0. \tag{3.3.1}$$

- *interactions*

$$\lambda \frac{\hat{\Theta}_{jk}^+ - \hat{\Theta}_{jk}^-}{\|\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-\|_2} = \mathscr{T}\left(X_{jk}^T(Y - \tilde{X}\hat{\phi}), \frac{1}{2}\left(\frac{\hat{\alpha}_g}{l_h} + \frac{\hat{\alpha}_h}{l_g} \right) \right) \quad if \quad \hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^- \neq 0$$

$$and \quad j \in g,\ k \in h,$$

$$\left\| X_{g:h}^T(Y - \tilde{X}\hat{\phi}) + sgn(X_{g:h}^T(Y - \tilde{X}\hat{\phi}))\frac{1}{2}\left(\frac{\hat{\alpha}_g}{l_h} + \frac{\hat{\alpha}_h}{l_g} \right) \right\|_2 \leq \lambda \quad if \quad \hat{\Theta}_{gh}^+ - \hat{\Theta}_{gh}^- = 0. \tag{3.3.2}$$

where $\mathscr{S}$ and $\mathscr{T}$ are thresholding operators as

$$\mathscr{S}(a,b) = \begin{cases} a - b & a > b \\ 0 & -|a| \leq b \\ a + b & a < -b \end{cases}$$

and

$$\mathscr{T}(a,b) = \begin{cases} -a + b & a > b \\ 0 & |a| \leq b \\ -a - b & a < -b. \end{cases}$$

Proof.

Consider the proposed optimization problem (3.2.5). The Lagrangian is

$$\frac{1}{2}\left\| Y - \tilde{X}\phi \right\|_2^2 + \lambda \sum_g \|\beta_g^+ - \beta_g^-\|_2 + \lambda \sum_{g \neq h} \|\Theta_{g:h}^+ - \Theta_{g:h}^-\|_F + \sum_g \alpha_g \left[1_{g.}^T(\Theta_g^+ + \Theta_g^-)\text{diag}(\frac{1}{w}) - 1_g^T(\beta_g^+ + \beta_g^-) \right]$$

$$- \sum_g \left[\gamma_g^{+T}\beta_g^+ + \gamma_g^{-T}\beta_g^- + \mu_g^{+T}\Theta_g^+ + \mu_g^{-T}\Theta_g^- \right] + \left\langle S, \Theta^+ - \Theta^- - \Theta^{+T} + \Theta^{-T} \right\rangle$$

where $\alpha, \gamma^{\pm}, \mu^{\pm}$, and S are dual variables. The KKT conditions for the primal-dual optimal

variable $(\hat{\phi}, \hat{\alpha}, \hat{\gamma}^{\pm}, \hat{\mu}^{\pm}, \hat{S})$ are

$$\pm X_g^T(Y - \tilde{X}\hat{\phi}) = \pm\lambda \frac{\hat{\beta}_g^+ - \hat{\beta}_g^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} - \hat{\alpha}_g 1_g - \hat{\gamma}_g^{\pm}$$

$$\pm X_{g:h}^T(Y - \tilde{X}\hat{\phi}) = \pm\lambda \frac{\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-}{\|\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-\|_2} + \frac{\hat{\alpha}_g}{w_g} 1_g - \hat{\mu}_{gh}^{\pm} + (\hat{S}_{gh} - \hat{S}_{hg}^T)\,\mathrm{sgn}(\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-)$$

$$\hat{\gamma}^{\pm T}\hat{\beta}^{\pm T} = 0, \qquad \hat{\mu}^{\pm T}\hat{\Theta}^{\pm T} = 0, \qquad \hat{\alpha}_g \left[1_{g\cdot}^T(\hat{\Theta}_g^+ + \hat{\Theta}_g^-)\mathrm{diag}(\frac{1}{w}) - 1_g^T(\hat{\beta}_g^+ + \hat{\beta}_g^-) \right] = 0$$

$$\hat{\alpha} \geq 0,\ \ \hat{\gamma}^{\pm} \geq 0,\ \ \hat{\mu}^{\pm} \geq 0,\ \ \hat{\beta}^{\pm} \geq 0,\ \ \hat{\Theta}^{\pm} \geq 0,\ \ 1_{g\cdot}^T(\Theta_g^+ + \Theta_g^-)\mathrm{diag}(\frac{1}{w}) \leq 1_g^T(\beta_g^+ + \beta_g^-),\ \ \hat{\Theta} = \hat{\Theta}^T.$$

First note that $\hat{\alpha}_g$ is a real value, however, $\hat{\gamma}_g$ and $\hat{\mu}_g$ are vectors with the size equal to the number of levels in the group g. We will now investigate the vector $\hat{\beta}_g$ component by component since it is possible that some components are positive and some are negative. There are three cases for main effects:

(i) $\hat{\beta}_j^+ \geq 0,\ \ \hat{\beta}_j^- = 0 \quad \left(\Rightarrow\ \hat{\gamma}_j^+ = 0\right)$ for $j \in g$

$$-X_j^T(Y - \tilde{X}\hat{\phi}) = \lambda \frac{\hat{\beta}_j^+ - \hat{\beta}_j^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} - \hat{\alpha}_g - \hat{\gamma}_j^+ = \lambda \frac{\hat{\beta}_j^+}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} - \hat{\alpha}_g$$

in this case, if $-X_j^T(Y - \tilde{X}\hat{\phi}) \leq -\hat{\alpha}_g$, then $\hat{\beta}_j^+ = 0$; hence

$$\lambda \frac{\hat{\beta}_j^+ - \hat{\beta}_j^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} = \left[-X_j^T(Y - \tilde{X}\hat{\phi}) + \hat{\alpha}_g\right]_+ = \mathscr{S}\left(-X_j^T(Y - \tilde{X}\hat{\phi}), -\hat{\alpha}_g\right).$$

(ii) $\hat{\beta}_j^+ = 0,\ \ \hat{\beta}_j^- \geq 0 \quad \left(\Rightarrow\ \hat{\gamma}_j^- = 0\right)$ for $j \in g$

$$X_j^T(Y - \tilde{X}\hat{\phi}) = -\lambda \frac{\hat{\beta}_j^+ - \hat{\beta}_j^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} - \hat{\alpha}_g - \hat{\gamma}_j^- = \lambda \frac{\hat{\beta}_j^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} - \hat{\alpha}_g$$

in this case, if $X_j^T(Y - \tilde{X}\hat{\phi}) \leq -\hat{\alpha}_g$, then $\hat{\beta}_j^- = 0$; hence

$$\lambda \frac{\hat{\beta}_j^+ - \hat{\beta}_j^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} = \left[-X_j^T(Y - \tilde{X}\hat{\phi}) - \hat{\alpha}_g\right]_- = \mathscr{S}\left(-X_j^T(Y - \tilde{X}\hat{\phi}), -\hat{\alpha}_g\right).$$

(iii) $\hat{\beta}_j^+ > 0,\ \ \hat{\beta}_j^- > 0 \quad \left(\Rightarrow\ \hat{\gamma}_j^{\pm} = 0\right)$ for $j \in g$

$$\mp X_j^T(Y - \tilde{X}\hat{\phi}) = \pm\lambda\frac{\hat{\beta}_j^+ - \hat{\beta}_j^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} - \hat{\alpha}_g - \hat{\gamma}_j^{\pm} = \pm\lambda\frac{\hat{\beta}_j^+ - \hat{\beta}_j^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} - \hat{\alpha}_g \quad \Longrightarrow \quad \hat{\alpha}_g = 0$$

thus, in this case we can write

$$\lambda\frac{\hat{\beta}_j^+ - \hat{\beta}_j^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} = -X_j^T(Y - \tilde{X}\hat{\phi}) = \mathscr{S}\left(-X_j^T(Y - \tilde{X}\hat{\phi}), -\hat{\alpha}_g\right).$$

In the first case, $\hat{\beta}_j^+ = 0$ when $-X_j^T(Y - \tilde{X}\hat{\phi}) \le -\hat{\alpha}_g < 0$; it means $\hat{\beta}_j^+ = 0$ when the correlation $X_j^T(Y - \tilde{X}\hat{\phi})$ is positive. In the second case, $\hat{\beta}_j^- = 0$ when $X_j^T(Y - \tilde{X}\hat{\phi}) \le -\hat{\alpha}_g < 0$; it means $\hat{\beta}_j^- = 0$ when the correlation $X_j^T(Y - \tilde{X}\hat{\phi})$ is negative. Hence, for the main effects one can say that

$$\lambda\frac{\hat{\beta}_j^+ - \hat{\beta}_j^-}{\|\hat{\beta}_g^+ - \hat{\beta}_g^-\|_2} = \mathscr{S}\left(-X_j^T(Y - \tilde{X}\hat{\phi}), -\hat{\alpha}_g\right) \quad \text{if} \quad \hat{\beta}_g^+ - \hat{\beta}_g^- \neq 0 \quad \text{and} \quad j \in g,$$

$$\left\|X_g^T(Y - \tilde{X}\hat{\phi}) - \text{sgn}(X_g^T(Y - \tilde{X}\hat{\phi}))\hat{\alpha}_g\right\|_2 \le \lambda \quad \text{if} \quad \hat{\beta}_g^+ - \hat{\beta}_g^- = 0.$$

In the previous section, it was explained that $\hat{\Theta}_g^{\pm} = \{\pm\hat{\Theta}_g, 0\}$, because $\|\hat{\Theta}_g\|_1$ has a positive coefficient in the Lagrangian. If a component of interaction effect $\hat{\Theta}_{gh}$ satisfies $\hat{\Theta}_{jk}^+ > 0$ and $\hat{\Theta}_{jk}^- > 0$, then by subtracting both of them by a constant value, the loss function will not change. However, the penalty term will result in a strictly lower value. Thus, there are two cases for the interaction effects:

(i) $\hat{\Theta}_{jk}^+ \ge 0,\ \ \hat{\Theta}_{jk}^- = 0 \quad \left(\Rightarrow\ \hat{\mu}_{jk}^+ = 0\right)$ for $\quad j \in g,\ k \in h,$

$$\hat{\Theta}_{kj}^{\pm} = \max\{\pm\hat{\Theta}_{kj}, 0\} \text{ and } \hat{\Theta}_{kj}^+ - \hat{\Theta}_{kj}^- = \hat{\Theta}_{jk}^+ - \hat{\Theta}_{jk}^- = \hat{\Theta}_{jk}^+ \quad \Longrightarrow \quad \hat{\Theta}_{kj}^- = 0,\ \ \hat{\Theta}_{kj}^+ = \hat{\Theta}_{jk}^+,$$

$$-X_{jk}^T(Y - \tilde{X}\hat{\phi}) = \lambda\frac{\hat{\Theta}_{jk}^+}{\|\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-\|_2} + \frac{\hat{\alpha}_g}{l_h} + (\hat{S}_{jk} - \hat{S}_{kj}),$$

$$-X_{kj}^T(Y-\tilde{X}\hat{\phi}) = \lambda\frac{\hat{\Theta}_{kj}^+}{\|\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-\|_2} + \frac{\hat{\alpha}_h}{l_g} + (\hat{S}_{kj} - \hat{S}_{jk}).$$

By summation of the both sides of equations we have

$$\begin{aligned}\lambda\frac{\hat{\Theta}_{jk}^+ - \hat{\Theta}_{jk}^-}{\|\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-\|_2} &= \left[-X_{jk}^T(Y-\tilde{X}\hat{\phi}) - \frac{1}{2}\left(\frac{\hat{\alpha}_g}{l_h} + \frac{\hat{\alpha}_h}{l_g}\right)\right]_+ \\ &= \mathscr{T}\left(X_{jk}^T(Y-\tilde{X}\hat{\phi}), \frac{1}{2}\left(\frac{\hat{\alpha}_g}{l_h} + \frac{\hat{\alpha}_h}{l_g}\right)\right).\end{aligned}$$

(ii) $\hat{\Theta}_{jk}^+ = 0,\ \ \hat{\Theta}_{jk}^- \geq 0 \quad \left(\Rightarrow\ \hat{\mu}_{jk}^- = 0\right)$ for $\quad j \in g,\ k \in h,$

in the same way $\hat{\Theta}_{kj}^+ = 0,\ \ \hat{\Theta}_{kj}^- = \hat{\Theta}_{jk}^-$ and

$$X_{jk}^T(Y-\tilde{X}\hat{\phi}) = -\lambda\frac{\hat{\Theta}_{jk}^+ - \hat{\Theta}_{jk}^-}{\|\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-\|_2} + \frac{\hat{\alpha}_g}{l_h} + (\hat{S}_{jk} - \hat{S}_{kj}),$$

$$X_{kj}^T(Y-\tilde{X}\hat{\phi}) = -\lambda\frac{\hat{\Theta}_{kj}^+ - \hat{\Theta}_{kj}^-}{\|\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-\|_2} + \frac{\hat{\alpha}_h}{l_g} + (\hat{S}_{kj} - \hat{S}_{jk}).$$

By summation of the both sides of equations we have

$$\begin{aligned}\lambda\frac{\hat{\Theta}_{jk}^+ - \hat{\Theta}_{jk}^-}{\|\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-\|_2} &= \left[-X_{jk}^T(Y-\tilde{X}\hat{\phi}) + \frac{1}{2}\left(\frac{\hat{\alpha}_g}{l_h} + \frac{\hat{\alpha}_h}{l_g}\right)\right]_- \\ &= \mathscr{T}\left(X_{jk}^T(Y-\tilde{X}\hat{\phi}), \frac{1}{2}\left(\frac{\hat{\alpha}_g}{l_h} + \frac{\hat{\alpha}_h}{l_g}\right)\right).\end{aligned}$$

Note that $\hat{\mu}^{\pm} \geq 0$, thus, for the interaction effects one can say that

$$\lambda\frac{\hat{\Theta}_{jk}^+ - \hat{\Theta}_{jk}^-}{\|\hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^-\|_2} = \mathscr{T}\left(X_{jk}^T(Y-\tilde{X}\hat{\phi}), \frac{1}{2}\left(\frac{\hat{\alpha}_g}{l_h} + \frac{\hat{\alpha}_h}{l_g}\right)\right) \quad \text{if} \quad \hat{\Theta}_{g:h}^+ - \hat{\Theta}_{g:h}^- \neq 0$$

$$\text{and} \quad j \in g,\ k \in h,$$

$$\left\|X_{g:h}^T(Y-\tilde{X}\hat{\phi}) + \text{sgn}(X_{g:h}^T(Y-\tilde{X}\hat{\phi}))\frac{1}{2}\left(\frac{\hat{\alpha}_g}{l_h} + \frac{\hat{\alpha}_h}{l_g}\right)\right\|_2 \leq \lambda \quad \text{if} \quad \hat{\Theta}_{gh}^+ - \hat{\Theta}_{gh}^- = 0.$$

□

The form of the solution derived in Theorem 1 suggests that one additional model selection takes place within groups because of the l_1-norm in the hierarchy constraint. The hierarchy constraint has an increasing effect on the selection of main effects and a decreasing effect on the selection of interactions.

3.4 Characterization of the solution

Theorem 1 characterizes the solution for main effects and interactions separately, to see how the hierarchy constraint affects the estimates of the main effects and interactions. In fact, main effects are more likely to be included in the fitted model than interactions based on the effect of the hierarchy constraints. To calculate the unbiased estimate of the degrees of freedom, we need to characterize the solution in a different way. The Lagrangian of (3.2.5) based on this new formulation is then

$$\frac{1}{2}\|Y_\epsilon - \widetilde{X}_\varepsilon\phi\|_2^2 + \lambda\sum_g \|D_g\phi_g\|_2 + \lambda\sum_{g\neq h}\|D_{g:h}\phi_{g:h}\|_2 - \mu^T A\phi + \upsilon^T E\phi. \tag{3.4.1}$$

In the Lagrangian, $A\phi$ and $E\phi$ represent the inequality and equality constraints, respectively. We rewrite the optimization problem with Lagrangian (3.4.1) in the form

$$\min_{\phi,u}\frac{1}{2}\|Y_\epsilon - \widetilde{X}_\varepsilon\phi\|_2^2 + \lambda\sum_g u_g + \lambda\sum_{g\neq h} u_{g:h} - \mu^T A\phi + \upsilon^T E\phi$$

where

$$\|D_g\phi_g\|_2 \le u_g \quad \text{and} \quad \|D_{g:h}\phi_{g:h}\|_2 \le u_{g:h}. \tag{3.4.2}$$

We define $\mathcal{H}$ as the new grand set containing all factors and interactions groups. By second-order cone programming or "Lorentz cone" [33]

$$\|D_h\phi_h\|_2 \le u_h \Longleftrightarrow \begin{bmatrix} D_h\phi_h \\ u_h \end{bmatrix} \in \mathcal{C}_{l_h+1} \Longleftrightarrow -\begin{bmatrix} D_h\phi_h \\ u_h \end{bmatrix} \preceq_{\mathcal{C}_{l_h+1}} 0,$$

where

$$\mathcal{C}_{l_h+1} = \left\{ \begin{bmatrix} x \\ t \end{bmatrix} \mid x \in \mathbb{R}^{l_h}, t \in \mathbb{R}, \|x\|_2 \leq t \right\}.$$

For ease of notation define $Y := Y_\varepsilon$ and $\tilde{X} := \tilde{X}_\varepsilon$. The Lagrangian with the dual variables $(\gamma_h, \delta_h) \in \mathbb{R}^{l_h} \times \mathbb{R}$, where $\|\gamma_h\|_2 \leq \delta_h$, is

$$\mathcal{L}(\phi, u, \mu, \upsilon, \gamma, \delta) = \frac{1}{2}\|Y - \widetilde{X}\phi\|_2^2 + \lambda \sum_{h \in \mathcal{H}} u_h - \mu^T A\phi + \upsilon^T E\phi - \sum_{h \in \mathcal{H}} \begin{bmatrix} D_h\phi_h \\ u_h \end{bmatrix}^T \begin{bmatrix} \gamma_h \\ \delta_h \end{bmatrix},$$

where the primal variables are (ϕ, u). Derivatives with respect to the primal variables are

$$\begin{aligned} \nabla_\phi \mathcal{L}(\phi, u, \mu, \upsilon, \gamma, \delta) &= -\widetilde{X}^T(Y - \widetilde{X}\phi) - A^T\mu + E^T\upsilon - D^T\gamma \\ \nabla_u \mathcal{L}(\phi, u, \mu, \upsilon, \gamma, \delta) &= \lambda 1 - \delta. \end{aligned} \tag{3.4.3}$$

From the second derivative, we have $\lambda = \delta_h$, thus, $\|\gamma_h\|_2 \leq \lambda$. Based on Slater's theorem, strong duality holds if there exists a strictly feasible point. In inequality constraints of (3.2.5) the strict inequality holds when $\beta^\pm > 0$. It holds with respect to relaxation to the l_1-norm penalty and, therefore, strong duality holds. Hence, the complementary slackness holds and

$$\gamma_h^T D_h\phi_h + \delta_h u_h = 0.$$

Note that $\|D_h\phi_h\|_2 \leq u_h$, $\|\gamma_h\|_2 \leq \delta_h$ and by Cauchy-Schwartz inequality $|\gamma_h^T D_h\phi_h| \leq \|\gamma_h\|_2\|D_h\phi_h\|_2$. Therefore

$$\gamma_h^T D_h\phi_h + \delta_h u_h \geq -\|\gamma_h\|_2\|D_h\phi_h\|_2 + \delta_h u_h \geq 0.$$

Hence, the complementary slackness holds if and only if the following conditions hold [16]

- $\|\gamma_h\|_2 < \delta_h \Rightarrow \|D_h\phi_h\|_2 = u_h = 0$
- $\|D_h\phi_h\|_2 < u_h \Rightarrow \|\gamma_h\|_2 = \delta_h = 0$

- $\|D_h\phi_h\|_2 = u_h, \ \ \|\gamma_h\|_2 = \delta_h \ \Rightarrow \ \gamma_h^T D_h\phi_h = -\delta_h u_h.$

The third condition with respect to (3.4.3) can be rewritten as

$$\gamma_h^T D_h\phi_h + \lambda\|D_h\phi_h\|_2 = 0. \tag{3.4.4}$$

The complementary slackness (3.4.4) holds if and only if [33]

- $D_h\phi_h = 0$
- $D_h\phi_h \neq 0$, $\|\gamma_h\|_2 = \lambda$ and $\gamma_h = -\lambda\frac{D_h\phi_h}{\|D_h\phi_h\|_2}$.

The KKT conditions for $(\hat{\phi}(y), (\hat{\mu}(y), \hat{v}(y), \hat{\gamma}(y)))$ to be an optimal primal-dual pair are as follows:

$$\begin{aligned}
\widetilde{X}^T(Y - \widetilde{X}\hat{\phi}) &= -A^T\hat{\mu} + E^T\hat{v} - D^T\hat{\gamma},\\
\hat{\mu}_i^T(A\hat{\phi})_i &= 0,\\
\hat{\gamma}_h^T D_h\hat{\phi}_h + \lambda\|D_h\hat{\phi}_h\|_2 &= 0,\\
\hat{\mu} &\geq 0,\\
A\hat{\phi} &\geq 0,\\
E\hat{\phi} &= 0,\\
\|\hat{\gamma}_h\|_2 &\leq \lambda.
\end{aligned} \tag{3.4.5}$$

Let us define the boundary set $\mathcal{A}(\hat{\phi})$ as

$$\mathcal{A}(\hat{\phi}) = \left\{i : A_i\hat{\phi} = 0\right\}. \tag{3.4.6}$$

By the KKT condition, $\hat{\mu}_i^T(A\hat{\phi})_i = 0$, we have $\hat{\mu}_i = 0$ iff $(A\hat{\phi})_i > 0$ and $\hat{\mu}_i > 0$ iff $(A\hat{\phi})_i = 0$ because $\hat{\mu} \geq 0$ and $A\hat{\phi} \geq 0$. In terms of the boundary set $\mathcal{A}(\hat{\phi})$, the first KKT condition

becomes

$$\widetilde{X}^T(Y - \widetilde{X}\hat{\phi}) = -\left(A_{\mathcal{A}(\hat{\phi})}\right)^T \hat{\mu}_{\mathcal{A}(\hat{\phi})} + E^T\hat{v} - D^T\hat{\gamma}. \tag{3.4.7}$$

The number of inequality constraints is $3p + 2p^2$, thus $\hat{\mu} \in \mathbb{R}^{3p+2p^2}$ and the set of $\mathcal{A}(\hat{\phi})$ refers to constraints with the values of zero. The matrix A has $3p + 2p^2$ rows as the number of inequality constraints and $2p+2p^2$ columns as the dimension of parameter of interest $\hat{\phi}$. The matrix $A_{\mathcal{A}(\hat{\phi})}$ refers to rows, or inequality constraints, of A where $A_i\hat{\phi} = 0$; the subscript is used to refer to corresponding rows. We need to use brackets for distinct matrices, A or A^T. Hence, $(A_i)^T$ means that we first select rows and then we transpose the new matrix, but $(A^T)_i$ means that at first the matrix is transposed and then the rows are selected. This means that, the columns of A are selected.

Let us define the support of $\hat{\phi}$ as

$$\mathcal{S}(\hat{\phi}) = \left\{h \in \mathcal{H} : D_h\hat{\phi}_h \neq 0\right\}. \tag{3.4.8}$$

Note that $\widetilde{X}_{\mathcal{S}}\hat{\phi}_{\mathcal{S}} = \widetilde{X}\hat{\phi}$ since $D_h\hat{\phi}_h = 0$ for $h \notin \mathcal{S}$. The equation (3.4.7) in terms of the support $\mathcal{S}$ is

$$\left(\widetilde{X}^T\right)_{\mathcal{S}}(Y - \widetilde{X}_{\mathcal{S}}\hat{\phi}_{\mathcal{S}}) = -\left(\left(A_{\mathcal{A}(\hat{\phi})}\right)^T\right)_{\mathcal{S}}\hat{\mu}_{\mathcal{A}(\hat{\phi})} + \left(E^T\right)_{\mathcal{S}}\hat{v} - (D_{\mathcal{S}})^T\hat{\gamma}_{\mathcal{S}}. \tag{3.4.9}$$

It is easier notionally and computationally to refer $\mathcal{S}$ as a subset of parameters in vector $(\hat{\beta}, \hat{\Theta})$. Then one can define the operator $O = \begin{bmatrix} O^+ \\ O^- \end{bmatrix}$ such that

$$O^+ = \begin{bmatrix} I_{|X|} & 0_{|X|} & 0 & 0_{|Z|} \\ 0 & 0 & I_{|Z|} & 0 \end{bmatrix} \quad \text{and} \quad O^- = \begin{bmatrix} 0_{|X|} & I_{|X|} & 0_{|Z|} & 0 \\ 0 & 0 & 0 & I_{|Z|} \end{bmatrix}$$

and, therefore, $O_{\mathcal{S}}\hat{\phi} = \begin{bmatrix} O_{\mathcal{S}}^{+} \\ O_{\mathcal{S}}^{-} \end{bmatrix}\hat{\phi} = \hat{\phi}_{\mathcal{S}}$ will refer to the corresponding subset of the vector $\hat{\phi}$. By this definition, one can write the equation (3.4.9) from the equation (3.4.7); we only need to show that $(D^T)_{\mathcal{S}}\,\hat{\gamma} = (D_{\mathcal{S}})^T\,\hat{\gamma}_{\mathcal{S}}$. In order to show this equality, define the matrix D as

$$D = \begin{bmatrix} I_{|\beta|} & -I_{|\beta|} & 0 & 0 \\ 0 & 0 & I_{|\Theta|} & -I_{|\Theta|} \end{bmatrix}.$$

Vector $\hat{\gamma}$ is in $\mathbb{R}^{p+p^2}$ and it can be shown as two parts of main effects and interactions as $\hat{\gamma} = \begin{bmatrix} \hat{\gamma}_{\beta} \\ \hat{\gamma}_{\Theta} \end{bmatrix}$. Therefore,

$$O_{\mathcal{S}}D^T\hat{\gamma} = O_{\mathcal{S}}\begin{bmatrix} \hat{\gamma}_{\beta} \\ -\hat{\gamma}_{\beta} \\ \hat{\gamma}_{\Theta} \\ -\hat{\gamma}_{\Theta} \end{bmatrix} = \begin{bmatrix} \hat{\gamma}_{\mathcal{S}(\beta)} \\ -\hat{\gamma}_{\mathcal{S}(\beta)} \\ \hat{\gamma}_{\mathcal{S}(\Theta)} \\ -\hat{\gamma}_{\mathcal{S}(\Theta)} \end{bmatrix} = (D_{\mathcal{S}})^T\begin{bmatrix} \hat{\gamma}_{\mathcal{S}(\beta)} \\ \hat{\gamma}_{\mathcal{S}(\Theta)} \end{bmatrix} = (D_{\mathcal{S}})^T\,\hat{\gamma}_{\mathcal{S}},$$

where

$$D_{\mathcal{S}} = \begin{bmatrix} I_{\beta_{\mathcal{S}}} & -I_{\beta_{\mathcal{S}}} & 0 & 0 \\ 0 & 0 & I_{\Theta_{\mathcal{S}}} & -I_{\Theta_{\mathcal{S}}} \end{bmatrix}. \tag{3.4.10}$$

Define the set $\mathcal{A}_{\mathcal{S}}(\hat{\phi})$ as

$$\mathcal{A}_{\mathcal{S}}(\hat{\phi}) = \left\{(i,h) \mid \; A_i\hat{\phi} = 0;\; D_h\hat{\phi}_h \neq 0\right\}, \tag{3.4.11}$$

where i refers to rows of matrix A and h refers to the columns of A. This means that $\mathcal{A}(\hat{\phi})$ refers to a subset of inequality constraints while $\mathcal{S}(\hat{\phi})$ refers to a subset of parameters of interest in the vector $\hat{\phi}$; therefore, the intersection of these two subsets is meaningless. In fact $A_{\mathcal{A}_{\mathcal{S}}(\hat{\phi})}$ refers to a submatrix of A with rows of $i \in \mathcal{A}_{(}\hat{\phi})$ and columns of $h \in \mathcal{S}(\hat{\phi})$.

Then, with some neglect in notation, one can write the equation (3.4.9) as

$$\widetilde{X}_{\mathcal{S}}^{T}(Y - \widetilde{X}_{\mathcal{S}}\hat{\phi}_{\mathcal{S}}) = -A_{\mathcal{A}_{\mathcal{S}}(\hat{\phi})}^{T}\hat{\mu}_{\mathcal{A}(\hat{\phi})} + E_{\mathcal{S}}^{T}\hat{v} - D_{\mathcal{S}}^{T}\hat{\gamma}_{s}. \tag{3.4.12}$$

It is important to note that when $h \in \mathcal{S}$, $\hat{\gamma}_h = -\lambda\frac{D_h\hat{\phi}_h}{\|D_h\hat{\phi}_h\|_2}$; however, in the case $h \notin \mathcal{S}$, one can say that $\hat{\gamma}_h$ is a vector in the subdifferential containing all the vectors $\hat{\gamma}_h$ such that $\left\|\frac{\hat{\gamma}_h}{\lambda}\right\|_2 \leq 1$. Theorem 2 investigates these two cases. Let us define the subdifferential of $\|D_h\phi_h\|_2$ before stating Theorem 2. In the equation (3.4.1), the term $\Omega(\phi) = \lambda\sum_h \|D_h\phi_h\|_2$ is not differentiable at $D_h\phi_h = 0$. We state the following lemma to investigate the subdifferential of $\Omega(\phi)$.

Lemma 1. *Let x be a vector in $\mathbb{R}^{2p}$ and $D_{p\times 2p}$ be a matrix, then the subdifferential of $\|Dx\|_2$ is*

$$\partial\|Dx\|_2 = \left\{D^T u : u \in \mathbb{R}^p, \|u\|_2 \leq 1\right\}.$$

Proof.

Let $f : \mathbb{R}^p \to \mathbb{R}$ be a convex function. Then the subdifferential $\partial f(y_0)$ is given by

$$u \in \partial f(y_0) \iff f(y) - f(y_0) \geq u \cdot (y - y_0) \quad \text{for all} \quad y \in \mathbb{R}^p$$

where $\cdot$ is inner product. Let $g(x) = f(Dx)$ then

$$\begin{aligned} w \in \partial g(x_0) &\iff g(x) - g(x_0) \geq w \cdot (x - x_0) \quad \text{for all} \quad x \in \mathbb{R}^{2p} \\ &\iff f(Dx) - f(Dx_0) \geq w \cdot (x - x_0) \quad \text{for all} \quad x \in \mathbb{R}^{2p}. \end{aligned}$$

Now, suppose that $w = D^T v$, then $f(Dx) - f(Dx_0) \geq v \cdot (Dx - Dx_0)$; hence,

$$\partial g(x_0) = D^T \partial f(y_0),$$

where $y_0 = Dx_0$. This means that for the subdifferential of $g(x) = f(Dx)$, we can just calculate the subdifferential of f, evaluate it at Dx_0 and multiply the set by D^T from the

left.

In the case that $f(y) = \|y\|_2$, $x_0 = 0 \in \mathbb{R}^{2p}$ and $y_0 = 0 \in \mathbb{R}^p$ we get

$$\partial f(y_0) = \{u \in \mathbb{R}^p : \|u\| \leq 1\}.$$

Therefore,

$$\begin{aligned}\partial g(x_0) &= D^T \{u \in \mathbb{R}^p : \|u\| \leq 1\} \\ &= \left\{D^T u : u \in \mathbb{R}^p, \|u\| \leq 1\right\}. \quad \square\end{aligned}$$

Thus, for the case $D_h \phi_h = 0$,

$$\partial_{\phi_h} \Omega(\phi) = \left\{\lambda D_h^T w_h \in \mathbb{R}^{|h|} : \|w_h\|_2 \leq 1\right\}.$$

Now, let $v_h = \lambda w_h$; hence,

$$\partial_{\phi_h} \Omega(\phi) = \left\{D_h^T v_h \in \mathbb{R}^{|h|} : \|\frac{v_h}{\lambda}\|_2 \leq 1\right\}.$$

Note that $\|\gamma_h\|_2 \leq \lambda$, i.e. $\|\frac{\gamma_h}{\lambda}\|_2 \leq 1$ and $\gamma_h, v_h \in \mathbb{R}^{|h|}$, then the above subdifferential is equivalent to

$$\partial_{\phi_h} \Omega(\phi) = \left\{D_h^T \gamma_h \in \mathbb{R}^{|h|} : \|\frac{\gamma_h}{\lambda}\|_2 \leq 1\right\}.$$

Theorem 2. *Let the support $\mathcal{S}(\hat{\phi})$ be defined as in (3.4.8). Therefore, the optimal solution of (3.2.5), $\hat{\phi}$, satisfies*

$$\begin{aligned} P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T (Y - \widetilde{X}_{\mathcal{S}} P_{\mathcal{S}} \hat{\phi}_{\mathcal{S}}) &= \lambda D_{\mathcal{S}}^T \mathcal{N}(\hat{\phi}_{\mathcal{S}}) && \text{where } h \in \mathcal{S}(\hat{\phi}), \\ \left\| P_h \overline{X}_h^T (Y - \overline{X}_{\mathcal{S}} \hat{b}_{\mathcal{S}}) \right\|_2 &\leq \lambda && \text{where } h \notin \mathcal{S}(\hat{\phi}). \end{aligned}$$

Proof.

1. In the case where $h \in \mathcal{S}$:

In the equation (3.4.12) for solution $\hat{\phi}_{\mathcal{S}}(y)$, we have $A_{\mathcal{A}_{\mathcal{S}}(\hat{\phi})}\hat{\phi}_{\mathcal{S}} = 0$ and $E_{\mathcal{S}}\hat{\phi}_{\mathcal{S}} = 0$ because $\hat{\phi}_{-\mathcal{S}} = 0$. This means that $\hat{\phi}_{\mathcal{S}} \in \text{null}(A_{\mathcal{A}_{\mathcal{S}}(\hat{\phi})}) \bigcap \text{null}(E_{\mathcal{S}})$. Suppose that $\widetilde{A}_{\mathcal{S}} = \begin{bmatrix} -A_{\mathcal{A}_{\mathcal{S}}(\hat{\phi})} \\ E_{\mathcal{S}} \end{bmatrix}$ and define the projection matrix $P_{\mathcal{S}} = I_{\mathcal{S}} - \widetilde{A}_{\mathcal{S}}^T(\widetilde{A}_{\mathcal{S}}\widetilde{A}_{\mathcal{S}}^T)^+\widetilde{A}_{\mathcal{S}}$, where A^+ stands for Moore-Penrose pseudo-inverse of A; thus, $P_{\mathcal{S}}\widetilde{A}_{\mathcal{S}}^T = 0$ and $P_{\mathcal{S}}\hat{\phi}_{\mathcal{S}} = \hat{\phi}_{\mathcal{S}}$ because $\widetilde{A}_{\mathcal{S}}\hat{\phi}_{\mathcal{S}} = 0$. By multiplying the projection matrix $P_{\mathcal{S}}$ from the left side into the equation (3.4.12), we have

$$(\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}})^T(Y - \widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}\hat{\phi}_{\mathcal{S}}) = -P_{\mathcal{S}}D_{\mathcal{S}}^T\hat{\gamma}_s. \tag{3.4.13}$$

This equation deals with $\widetilde{X}P$ instead of $\widetilde{X}$ and it is similar to the result of Tibshirani et al. [30] and Bien et al. [27]. We know that when $D_h\hat{\phi}_h \neq 0$, $\hat{\gamma}_h = -\lambda\frac{D_h\hat{\phi}_h}{\|D_h\hat{\phi}_h\|_2}$ and, therefore,

$$P_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^T(Y - \widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}\hat{\phi}_{\mathcal{S}}) = \lambda P_{\mathcal{S}}D_{\mathcal{S}}^T\mathcal{N}(\hat{\phi}_{\mathcal{S}}),$$

where $\mathcal{N}(\phi)$ is a normalization operator such that

$$\mathcal{N}(\phi) = u \quad \text{where} \quad u_h = \frac{D_h\phi_h}{\|D_h\phi_h\|_2}.$$

It is easy to verify that $\lambda P_{\mathcal{S}}D_{\mathcal{S}}^T\mathcal{N}(\hat{\phi}_{\mathcal{S}}) = \lambda D_{\mathcal{S}}^T\mathcal{N}(\hat{\phi}_{\mathcal{S}})$ because by considering (3.4.10) we have

$$P_{\mathcal{S}}D_{\mathcal{S}}^TD_{\mathcal{S}}\hat{\phi}_{\mathcal{S}} = P_{\mathcal{S}}\left(I - \begin{bmatrix} 0 & I_{\beta_{\mathcal{S}}} & 0 & 0 \\ I_{\beta_{\mathcal{S}}} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{\Theta_{\mathcal{S}}} \\ 0 & 0 & I_{\Theta_{\mathcal{S}}} & 0 \end{bmatrix}\right)\hat{\phi}_{\mathcal{S}} = P_{\mathcal{S}}\hat{\phi}_{\mathcal{S}} - P_{\mathcal{S}}\begin{bmatrix} \hat{\beta}_{\mathcal{S}}^- \\ \hat{\beta}_{\mathcal{S}}^+ \\ \hat{\Theta}_{\mathcal{S}}^- \\ \hat{\Theta}_{\mathcal{S}}^+ \end{bmatrix} = \hat{\phi}_{\mathcal{S}} - \begin{bmatrix} \hat{\beta}_{\mathcal{S}}^- \\ \hat{\beta}_{\mathcal{S}}^+ \\ \hat{\Theta}_{\mathcal{S}}^- \\ \hat{\Theta}_{\mathcal{S}}^+ \end{bmatrix} = D_{\mathcal{S}}^TD_{\mathcal{S}}\hat{\phi}_{\mathcal{S}}.$$

Note that

$$P_{\mathcal{S}}\begin{bmatrix}\hat{\beta}^{-}_{\mathcal{S}}\\ \hat{\beta}^{+}_{\mathcal{S}}\\ \hat{\Theta}^{-}_{\mathcal{S}}\\ \hat{\Theta}^{+}_{\mathcal{S}}\end{bmatrix} := P_{\mathcal{S}}\hat{\phi}^{*}_{\mathcal{S}} = \hat{\phi}^{*}_{\mathcal{S}} - \widetilde{A}^{T}_{\mathcal{S}}(\widetilde{A}_{\mathcal{S}}\widetilde{A}^{T}_{\mathcal{S}})^{+}\widetilde{A}_{\mathcal{S}}\hat{\phi}^{*}_{\mathcal{S}} = \hat{\phi}^{*}_{\mathcal{S}},$$

because $\widetilde{A}_{\mathcal{S}}\hat{\phi}^{*}_{\mathcal{S}} = 0$ since $\widetilde{A}_{\mathcal{S}}\hat{\phi}_{\mathcal{S}} = 0$. This means that the constraints of $\widetilde{A}_{\mathcal{S}}$ are satisfied with respect to $\hat{\phi}^{*}_{\mathcal{S}}$ since the inequality and equality constraints do not change by changing the order of the positive and negative parts of the estimate, i.e. $\hat{\beta}^{+}, \hat{\beta}^{-}, \hat{\Theta}^{+}$, and $\hat{\Theta}^{-}$. Therefore,

$$P_{\mathcal{S}}\widetilde{X}^{T}_{\mathcal{S}}(Y - \widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}\hat{\phi}_{\mathcal{S}}) = \lambda D^{T}_{\mathcal{S}}\mathcal{N}(\hat{\phi}_{\mathcal{S}}). \tag{3.4.14}$$

2. In the case where $h \notin \mathcal{S}$:

We know that $DD^{T} = 2I$. Let $\overline{X} = (X, Z)$ and $b = (\beta, \Theta)$, thus, $\widetilde{X} = \overline{X}D$. One can write (3.4.7) as

$$\overline{X}^{T}(Y - \overline{X}\hat{b}) = -\frac{1}{2}DA^{T}_{\mathcal{A}(\hat{\phi})}\hat{\mu}_{\mathcal{A}(\hat{\phi})} + \frac{1}{2}DE^{T}\hat{\upsilon} - \hat{\gamma}.$$

Both sides of the equality are vectors and, therefore, they are componentwise equal and equality holds for each group as follows:

$$\overline{X}^{T}_{h}(Y - \overline{X}\hat{b}) = -\frac{1}{2}D_{h}A^{T}_{\mathcal{A}(\hat{\phi})}\hat{\mu}_{\mathcal{A}(\hat{\phi})} + \frac{1}{2}D_{h}E^{T}\hat{\upsilon} - \hat{\gamma}_{h}$$

where the index h stands for rows of the related matrix. Suppose that $\widetilde{A}_{h} = \begin{bmatrix} A_{\mathcal{A}(\hat{\phi})}D^{T}_{h} \\ ED^{T}_{h}\end{bmatrix}$ and similar as before, define $P_{h} = I - \widetilde{A}^{T}_{h}(\widetilde{A}_{h}\widetilde{A}^{T}_{h})^{-1}\widetilde{A}_{h}$ and

$$P_{h}\overline{X}^{T}_{h}(Y - \overline{X}_{\mathcal{S}}\hat{b}_{\mathcal{S}}) = -P_{h}\hat{\gamma}_{h}.$$

Note that $\hat{\gamma}_h$ is the vector in the subdifferential which contains all the vectors in $\mathbb{R}^{l_h}$ such that $\|\frac{\hat{\gamma}_h}{\lambda}\|_2 \leq 1$. Therefore,

$$\|P_h \overline{X}_h^T (Y - \overline{X}_{\mathbb{S}} \hat{b}_{\mathbb{S}})\|_2 \leq \lambda \sup_{\|\frac{\hat{\gamma}_h}{\lambda}\|_2 \leq 1} \left\| P_h \frac{\hat{\gamma}_h}{\lambda} \right\|_2 = \lambda \|P_h\|_2 ,$$

where $\|P_h\|_2$ is the induced l_2-norm of matrix P_h. Let δ be eigenvalue and υ the eigenvector of P_h. Then $P_h\upsilon = \delta\upsilon$ and $P_h(P_h\upsilon) = P_h(\delta\upsilon) = \delta(\delta\upsilon) = \delta^2\upsilon$ and $P_h(P_h\upsilon) = P_h\upsilon = \delta\upsilon$; thus, $\delta^2 = \delta$ and $\delta = \{0, 1\}$. This means that $\|P_h\|_2 = \sqrt{\max\{0,1\}} = 1$, and, therefore,

$$\|P_h \overline{X}_h^T (Y - \overline{X}_{\mathbb{S}} \hat{b}_{\mathbb{S}})\|_2 \leq \lambda. \quad \square$$

Theorem 2 provides the characterization for nonzero groups which will be used for calculating of the degrees of freedom. Also it provides a selection process as a mechanism to drop out zero groups which is useful for the computational algorithm.

3.5 The matrix $\widetilde{A}$

First note that $\widetilde{A}_{\mathbb{S}} = \begin{bmatrix} A_{\mathcal{A}_{\mathbb{S}}(\hat{\phi})} \\ E_{\mathbb{S}} \end{bmatrix}$ where A is the matrix which refers to the inequality constraints $A\phi \geq 0$ in (3.2.5) and E is the matrix which refers to the equality constraints $E\phi = 0$ in (3.2.5). The matrix A in (3.4.1) has a negative sign that means $Ax > 0$ so all inequalities in matrix $\widetilde{A}$ will be greater than zero. Each row of matrix $\widetilde{A}$ corresponds to a constraint and its columns corresponds to $(\beta^+, \beta^-, \Theta^+, \Theta^-)$. The inequality constraints in (3.2.5) are

C_1: $\|\hat{\Theta}_g\|_1 \leq 1^T(\hat{\beta}_g^+ + \hat{\beta}_g^-)$,

C_2: $\hat{\beta}_j^+ > 0$,

C_3: $\hat{\beta}_j^- > 0$,

C_4: $\hat{\Theta}_{jk}^+ > 0$,

C_5: $\hat{\Theta}_{jk}^- 0$,

the equality constraints are

C_6: $\hat{\Theta}^{+}_{g:g} = 0$,

C_7: $\hat{\Theta}^{-}_{g:g} = 0$,

C_8: $\hat{\Theta}^{+}_{jk} - \hat{\Theta}^{-}_{jk} = \hat{\Theta}^{+}_{kj} - \hat{\Theta}^{-}_{kj}$.

Note that the number of constraints depends on the subscripts of groups or components. The first constraint is applied on group g which results in a single constraint. The constraints C_6 and C_7 are applied on the matrix $\hat{\Theta}^{\pm}_{g:g} = 0$ which results in a number of constraints equals to the number of components of the matrix $\hat{\Theta}^{\pm}_{g:g}$. Let 1_p be the row vector of ones, and e_j be the row vector of zeros with a single 1 in j-th position. The vector e_g is constructed similarly. Similarly to matrix $\widetilde{D}$ in [27], we define the matrix $\widetilde{A}$ as follows:

R1.	For all factors g	$(e_g$	e_g	$-e_g \otimes 1_p$	$-e_g \otimes 1_p$ $)$
R2.	For all levels j of factors g	$(e_j$	0	0	0 $)$
R3.	For all levels j of factors	$(0$	e_j	0	0 $)$
R4.	For all levels jk of interactions	$(0$	0	$e_j \otimes e_k$	0 $)$
R5.	For all levels jk of interactions	$(0$	0	0	$e_j \otimes e_k$ $)$
R6.	For all levels j of factors	$(0$	0	$e_j \otimes e_k$	0 $)$
R7.	For all levels j of factors	$(0$	0	0	$e_j \otimes e_k$ $)$
R8.	For all levels $j < k$	$(0$	0	$e_j \otimes e_k - e_k \otimes e_j$	$-e_j \otimes e_k + e_k \otimes e_j$ $)$.

3.6 The degrees of freedom

In Chapter 1, we discussed the definition of degrees of freedom from Efron [6] and the unbiased estimate of degrees of freedom of Stein [24]. The definition of Efron [6] may not be applicable in practical situations since it depends on μ, the parameter of interest of regression analysis. Stein's unbiased estimate of degrees of freedom (1.3.3) does not depend on μ and is computationally efficient. It states that the estimate of degrees of freedom is the

divergence of the fit; thus, we need to calculate $\frac{\partial \hat{\phi}}{\partial y}(Y)$ from the equation (3.4.7). To this end, we need to determine the function $\hat{\phi}(y)$ implicitly from the KKT condition (3.4.7). The use of the implicit function theorem for this task is hampered by the fact that the KKT condition (3.4.7) is not everywhere differentiable. We follow the notation of Vaiter et al. [4]. Let $\bar{\mathcal{S}}$ be the support of a vector ϕ. For any group $h \notin \bar{\mathcal{S}}$, define the boundary as

$$\mathcal{B}_{\bar{\mathcal{S}},h} = \left\{ y \in \mathbb{R}^n \mid \exists \phi : \frac{\|P_h \bar{X}_h^T r\|_2}{\|P_h\|_2} = \lambda \quad \text{and} \quad P_{\bar{\mathcal{S}}} \widetilde{X}_{\bar{\mathcal{S}}} r = \lambda D^T \mathcal{N}(\phi_{\bar{\mathcal{S}}}) \right\},$$

where $r = y - \widetilde{X}\phi = y - \bar{X}\beta$ and let

$$\mathcal{B} = \bigcup_{\bar{\mathcal{S}} \in \mathcal{H}} \bigcup_{h \notin \bar{\mathcal{S}}} \mathcal{B}_{\bar{\mathcal{S}},h}. \tag{3.6.1}$$

Vaiter et al. [4] showed that the Lebesgue measure of the boundary $\mathcal{B}$ is zero on $\mathbb{R}^n$. Let $Y \notin \mathcal{B}_{\mathcal{S}}$ and $\mathcal{S}$ be the support of $\hat{\phi}$ as defined in (3.4.8); then define the following mapping

$$\Gamma(\phi_{\mathcal{S}}(y), y) = P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T \widetilde{X}_{\mathcal{S}} P_{\mathcal{S}} \phi_{\mathcal{S}} - P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T y + \lambda D_{\mathcal{S}}^T \mathcal{N}(\phi_{\mathcal{S}}) \tag{3.6.2}$$

Note that based on Theorem 2 the optimal solution satisfies $\Gamma(\hat{\phi}_{\mathcal{S}}(Y), Y) = 0$. The implicit function theorem requires the invertibility of the Jacobian of the KKT condition. The following assumption on $X_{\mathcal{S}}$ is needed for this purpose.

Assumption 1. *Suppose that $\mathcal{S}$ is the support. We assume that for all solutions $b_{\mathcal{S}}$, the fit $\widehat{Y} = \overline{X}_{\mathcal{S}} b_{\mathcal{S}} \neq \underline{0}$.*

This assumption is feasible since in a typical case, the fit cannot be constant zero. It is needed for the following lemma which investigates invertibility of the Jacobian.

Lemma 2. *Suppose that $\phi_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$, $\lambda > 0$ and Assumption 1 holds. Thus $\partial_{\phi_{\mathcal{S}}} \Gamma(\phi_{\mathcal{S}}(y), y)$ is invertible.*

Proof.

First, note that $\forall h \in \mathcal{S},\ D_h\hat{\phi}_h \neq 0$, therefore,

$$\partial_{\phi_{\mathcal{S}}}\Gamma(\phi_{\mathcal{S}}(y),y) = P_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^T\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}} + \lambda \text{diag}\left(D_h^T\left[\frac{I - \frac{D_h\phi_h\phi_h^TD_h^T}{\|D_h\phi_h\|_2^2}}{\|D_h\phi_h\|_2}\right]D_h\right)_{h\in\mathcal{S}}. \tag{3.6.3}$$

We know that $P_{\mathcal{S}} = P_{\mathcal{S}}^T$ so $P_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^T\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}$ is symmetric positive semi-definite because $x^TP_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^T\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}x = (\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}x)^T(\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}x) = \|\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}x\|_2^2 \geq 0$. On the other hand, $I - \frac{D_h\phi_h\phi_h^TD_h^T}{\|D_h\phi_h\|_2^2} = I - \beta_h(\beta_h^T\beta_h)^{-1}\beta_h^T = \text{Proj}_{\beta_h^\perp}$. Therefore, $B_h = \frac{I-\frac{D_h\phi_h\phi_h^TD_h^T}{\|D_h\phi_h\|_2^2}}{\|D_h\phi_h\|_2} = \frac{\text{Proj}_{(D_h\phi_h)^\perp}}{\|D_h\phi_h\|_2}$ is symmetric positive semi-definite because

$$\langle x, \text{Proj}_{(D_h\phi_h)^\perp}x\rangle = \|\text{Proj}_{(D_h\phi_h)^\perp}x\|_2^2 \geq 0$$

Hence, diag $\left(D_h^TB_hD_h\right)_{h\in\mathcal{S}}$ is symmetric positive semi-definite. Note that $\|\text{Proj}_{(D_h\phi_h)^\perp}x\|_2^2 = 0$ if and only if $x = D_h\phi_h$. It means $\ker(B_{\mathcal{S}}) = \{D_{\mathcal{S}}\phi_{\mathcal{S}}\}$, hence, $\ker(D_{\mathcal{S}}^TB_{\mathcal{S}}D_{\mathcal{S}}) = \{D_{\mathcal{S}}^TD_{\mathcal{S}}\phi_{\mathcal{S}}\}$ since $D_{\mathcal{S}}D_{\mathcal{S}}^T = 2I$. If $D_{\mathcal{S}}^TD_{\mathcal{S}}\phi_{\mathcal{S}} \in \ker(P_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^T\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}})$ then

$$\langle D_{\mathcal{S}}^TD_{\mathcal{S}}\phi_{\mathcal{S}}, P_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^T\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}D_{\mathcal{S}}^TD_{\mathcal{S}}\phi_{\mathcal{S}}\rangle = \|\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}D_{\mathcal{S}}^TD_{\mathcal{S}}\phi_{\mathcal{S}}\|_2^2 = 0$$

and

$$\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}D_{\mathcal{S}}^TD_{\mathcal{S}}\phi_{\mathcal{S}} = 2\overline{X}_{\mathcal{S}}b_{\mathcal{S}} = \underline{0}$$

which contradicts Assumption 1. Therefore, $\ker(D_{\mathcal{S}}^TB_{\mathcal{S}}D_{\mathcal{S}}) \bigcap \ker(P_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^T\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}) = \{0\}$ and $\partial_{\phi_{\mathcal{S}}}\Gamma(\phi_{\mathcal{S}}(y),y)$ is invertible. □

Lemma 3. *Let $\hat{\phi}(Y)$ be the solution of the convex problem (3.2.5) with support $\mathcal{S}$, defined in (3.4.8); then there exists unique function $\hat{\phi}(y)$ in a neighbourhood $\widetilde{\mathcal{O}}$ of Y such that*

$$\forall h \notin \mathcal{S}, \quad \partial_y\hat{\phi}_h(Y) = 0 \quad \textit{and} \quad \partial_y\hat{\phi}_{\mathcal{S}}(Y) = \left(\partial_{\hat{\phi}_{\mathcal{S}}}\Gamma(\hat{\phi}_{\mathcal{S}}(Y),Y)\right)^{-1}P_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^T.$$

Proof.

Let $\Gamma(\phi_{\mathcal{S}}(y),y)$ be the mapping defined in (3.6.2) reads on $\left(\mathbb{R}^{|\mathcal{S}|}\setminus U\right)\times\mathbb{R}^n$ where U =

$\left\{u \in \mathbb{R}^{|\mathcal{S}|} \mid \exists h \in \mathcal{S} : u_h = 0\right\}$. Note that $\Gamma(\hat{\phi}_{\mathcal{S}}(Y), Y) = 0$ according to Theorem 2 and $\partial_{\phi_{\mathcal{S}}}\Gamma($ $\phi_{\mathcal{S}}(Y), Y)$ is invertible according to Lemma 2. By the implicit function theorem, there exists a neighborhood $\mathcal{O}$ of Y such that

$$\forall y \in \mathcal{O} \quad \Gamma(\phi_{\mathcal{S}}(y), y) = 0 \quad \text{and} \quad \phi_{\mathcal{S}}(Y) = \left[\hat{\phi}(Y)\right]_{\mathcal{S}}$$

and we extend $\phi_{\mathcal{S}}(y)$ on $\mathcal{S}^c$ as $\phi_{\mathcal{S}^c}(y) = 0$. We assumed in definition (3.6.2) that $Y \notin \mathcal{B}_{\mathcal{S}}$. For $\hat{\phi}(Y)$, we have

$$\forall h \notin \mathcal{S}, \quad \left\|P_h \bar{X}_h^T (Y - \widetilde{X}_{\mathcal{S}} \left[\hat{\phi}(Y)\right]_{\mathcal{S}})\right\|_2 \leq \lambda \left\|P_h\right\|_2$$

where $\widetilde{X}_{\mathcal{S}} \left[\hat{\phi}(Y)\right]_{\mathcal{S}} = \bar{X}_{\mathcal{S}} \left[\hat{b}(Y)\right]_{\mathcal{S}}$. If there exists $h \notin \mathcal{S}$ such that

$$\left\|P_h \bar{X}_h^T (Y - \widetilde{X}_{\mathcal{S}} \left[\hat{\phi}(Y)\right]_{\mathcal{S}})\right\|_2 = \lambda \left\|P_h\right\|_2$$

then $Y \in \mathcal{B}_{\mathcal{S}}$ which contradicts the assumption of $Y \notin \mathcal{B}_{\mathcal{S}}$. Hence

$$\forall h \notin \mathcal{S}, \quad \left\|P_h \bar{X}_h^T (Y - \widetilde{X}_{\mathcal{S}} \left[\hat{\phi}(Y)\right]_{\mathcal{S}})\right\|_2 < \lambda \left\|P_h\right\|_2 .$$

We know that $\phi_{\mathcal{S}}(Y) = \left[\hat{\phi}(Y)\right]_{\mathcal{S}}$ and $\phi_{\mathcal{S}}(y)$ is continuous for every $y \in \mathcal{O}$. Then we can find $\widetilde{\mathcal{O}} \subseteq \mathcal{O}$ including Y such that for every $y \in \widetilde{\mathcal{O}}$, we have

$$\forall h \notin \mathcal{S}, \quad \left\|P_h \bar{X}_h^T (y - \widetilde{X}_{\mathcal{S}} \phi_{\mathcal{S}}(y))\right\|_2 \leq \lambda \left\|P_h\right\|_2 .$$

On the other hand by the definition of the mapping $\phi_{\mathcal{S}}(y)$ for every $y \in \widetilde{\mathcal{O}}$, we have

$$P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T (y - X_{\mathcal{S}} \phi_{\mathcal{S}}(y)) = \lambda D_{\mathcal{S}}^T \mathcal{N}(\phi_{\mathcal{S}}(y)) \quad \text{and} \quad \operatorname{supp}(\phi_{\mathcal{S}}(y)) = \mathcal{S}.$$

Then, from Theorem 2, $\phi(y)$ is a solution for (3.2.3). The solution of (3.2.3) is unique and, therefore, for every $y \in \widetilde{\mathcal{O}}$, $\phi(y) = \hat{\phi}(y)$. This result states that the derived implicit function

equals to the minimizer of the convex problem (3.2.3). Therefore

$$\forall h \notin \mathcal{S},\ \forall y \in \widetilde{\mathcal{O}},\quad \left[\hat{\phi}(y)\right]_h = \phi_h(y) = 0 \quad \Longrightarrow \quad \forall h \notin \mathcal{S},\quad \partial_y \hat{\phi}_h(Y) = 0$$

and with respect to the implicit function theorem

$$\partial_y \hat{\phi}_{\mathcal{S}}(Y) = -\left(\partial_{\hat{\phi}_{\mathcal{S}}} \Gamma(\hat{\phi}_{\mathcal{S}}(Y), Y)\right)^{-1} \partial_y \Gamma(\hat{\phi}_{\mathcal{S}}(Y), Y) = \left(\partial_{\hat{\phi}_{\mathcal{S}}} \Gamma(\hat{\phi}_{\mathcal{S}}(Y), Y)\right)^{-1} P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T. \quad \square$$

Theorem 3. *The unbiased estimate of the degrees of freedom of the hierarchical group LASSO (3.2.5) is*

$$\widehat{df}_\lambda(\widetilde{X}\hat{\phi}(Y)) = tr\left(\widetilde{X}_{\mathcal{S}} P_{\mathcal{S}} \left(P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T \widetilde{X}_{\mathcal{S}} P_{\mathcal{S}} + \lambda U\right)^{-1} P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T\right), \tag{3.6.4}$$

where

$$U = diag\left(D_h^T \left[\frac{I - \frac{D_h \hat{\phi}_h \hat{\phi}_h^T D_h^T}{\|D_h \hat{\phi}_h\|_2^2}}{\|D_h \hat{\phi}_h\|_2}\right] D_h\right)_{h \in \mathcal{S}}. \tag{3.6.5}$$

Proof.

By decomposing $\widetilde{X} = \left(\widetilde{X}_{\mathcal{S}},\ \widetilde{X}_{\mathcal{S}^c}\right)$, the degrees of freedom with respect to Stein's formula are

$$\begin{aligned} df_\lambda(\widetilde{X}\hat{\phi}(Y)) &= E\left[\left(\nabla \cdot \widetilde{X}\hat{\phi}\right)(Y)\right] \\ &= E_{\mathcal{B}}\left[\left(\nabla \cdot \widetilde{X}\hat{\phi}\right)(Y)\right] + E_{\mathcal{B}^c}\left[\left(\nabla \cdot \widetilde{X}\hat{\phi}\right)(Y)\right] \\ &= 0 + E_{\mathcal{B}^c}\left[\left(\nabla \cdot \widetilde{X}\hat{\phi}\right)(Y)\right] \\ &= E_{\mathcal{B}^c}\left[\mathrm{tr}\left(\widetilde{X}_{\mathcal{S}} P_{\mathcal{S}} \left(\partial_{\hat{\phi}_{\mathcal{S}}} \Gamma(\hat{\phi}_{\mathcal{S}}(Y), Y)\right)^{-1} P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T + \widetilde{X}_{\mathcal{S}^c} 0\right)\right], \end{aligned}$$

where $\partial_{\hat{\phi}_{\mathcal{S}}} \Gamma(\hat{\phi}_{\mathcal{S}}(y), y)$ is defined by (3.6.3). Note that the Lebesgue measure of $\mathcal{B}$ is zero and $\hat{\phi}_{\mathcal{S}} = P_{\mathcal{S}} \hat{\phi}_{\mathcal{S}}$. The trace operator is invariant under the transpose; hence,

$$\widehat{df}_\lambda(\widetilde{X}\hat{\phi}(Y)) = \mathrm{tr}\left(\widetilde{X}_{\mathcal{S}} P_{\mathcal{S}} \left(\partial_{\hat{\phi}_{\mathcal{S}}} \Gamma(\hat{\phi}_{\mathcal{S}}(Y), Y)\right)^{-1} P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T\right).$$

Note that by Stein's formula the divergence of fit is an unbiased estimator of degrees of freedom i.e. $E\left[\widehat{df}\right] = df$. Therefore,

$$\widehat{df}_\lambda(\widetilde{X}\hat{\phi}(Y)) = \text{tr}\left(\widetilde{X}_{\mathcal{S}} P_{\mathcal{S}} \left(P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T \widetilde{X}_{\mathcal{S}} P_{\mathcal{S}} + \lambda U\right)^{-1} P_{\mathcal{S}} \widetilde{X}_{\mathcal{S}}^T\right) \tag{3.6.6}$$

where U is defined in (3.6.3). □

This is similar to the degrees of freedom of group LASSO derived by Vaiter et al. [4], equal to

$$\text{tr}\left(\widetilde{X}_{\mathcal{S}} \left(\partial_{\hat{\phi}_{\mathcal{S}}} \Gamma(\hat{\phi}_{\mathcal{S}}(Y), Y)\right)^{-1} \widetilde{X}_{\mathcal{S}}^T\right).$$

The estimate of degrees of freedom for hierarchical LASSO derived by Bien et al. [27] is equal to

$$\text{tr}\left((\widetilde{X}P)(\widetilde{X}P)^+\right) \tag{3.6.7}$$

where $(\cdot)^+$ refers to Moore-Penrose pseudo inverse. In our case, $(\widetilde{X}P)^+$ is replaced with a more complicated matrix of $\left(\partial_{\hat{\phi}_{\mathcal{S}}} \Gamma(\hat{\phi}_{\mathcal{S}}(y), y)\right)^{-1} \widetilde{X}_{\mathcal{S}}^T$ because instead of the LASSO penalty in the hierarchical LASSO, the group LASSO penalty is used in our case.

This is also comparable to the degrees of freedom of elastic net regularization [13]:

$$\widehat{df}_\lambda(\text{elastic net}) = \text{tr}\left(X_{\mathcal{A}} \left(X_{\mathcal{A}}^T X_{\mathcal{A}} + \lambda_2 I\right)^{-1} X_{\mathcal{A}}^T\right), \tag{3.6.8}$$

where $\mathcal{A}$ is the active set. It carries some characteristics from each regularization; for instance, the projection matrix $P_{\mathcal{S}}$ is analogous to the hierarchical LASSO (3.6.7) and the general form is analogous to elastic net regularization.

Chapter 4

Group LASSO with Quadratic Loss: Further Theory

4.1 Introduction

In this chapter, we study a standardized group LASSO as a specific version of group LASSO. We use the transformation $\frac{X_g}{\sqrt{l_g}}$ of Zhao et al. [36] to unify group weights. Our Theorem 4 shows that by using this transformation, the estimate of the standardized group LASSO does not change. The solution of the standardized group LASSO with this transformation is characterized in Theorem 5, and the unbiased estimate of the degrees of freedom is given by Theorem 6. Finally, we calculate the degrees of freedom for the overlapped group LASSO in Corollary 1.

The standard algorithm for group LASSO, proposed by Yuan and Lin [18], assumes that the design matrix in each group is orthonormal, i.e. $X_g^T X_g = I_g$. Simon and Tibshirani [28] noticed that this orthonormalization is neglected in the statistical literature. They showed that by orthonormalizing the design matrix in group LASSO, the problem changes into a new problem. In fact, by orthonormalizing the design matrix of each group, group LASSO penalizes the fit instead of the coefficients. They also showed that the new problem selects groups roughly according to a UMP test.

4.2 Group weights and normalization

Yuan and Lin [18] proposed the group weights $d_g = \sqrt{l_g}$, for the group LASSO penalty. Lim [17] considered $d_g = 1$ for all groups g. Zhao et al. [36] used the transformation $X_g/\sqrt{l_g}$ for making unit group weights. Vaiter et al. [4] used the group LASSO penalty with the unit group weights in order to get rid of some technical difficulties in the calculation of the degrees of freedom; for the same reason, we need a penalty with unit group weights.

Lim [17] argued that the quantity $\|X_g^T(Y-\hat{Y})\|_2$ in the KKT conditions determines whether group g gets zero coefficients or not. He considered $Y - \hat{Y} = \varepsilon$ where $\varepsilon \sim N(0, I)$. In fact, the tuning parameter λ encourages considering such a null model. As a heuristic reason for this null model, note that in the KKT conditions $\|X_g^T(Y - \hat{Y})\|_2 \leq \lambda d_g$ for all g. One can say that d_g is related to X_g and λ is related to $Y - \hat{Y}$; thus, one may consider $\|Y - \hat{Y}\|_2 \propto \lambda$ which leads to the above null model. Hence,

$$d_g^2 = E\|X_g^T\varepsilon\|_2^2 = E\left[\text{tr}(\varepsilon^T X_g X_g^T \varepsilon)\right] = \text{tr}(X_g^T X_g) = \|X_g\|_F^2.$$

Therefore, Lim [17] picked $d_g = \|X_g\|_F$, the Frobenius norm of the matrix X_g. In the case of orthonormal X_g, we have $X_g^T X_g = I_g$, thus $d_g = \sqrt{l_g}$ which is proposed by [18]. However, in Lim's case [17], X_g is an indicator matrix where each row contains a single 1 with zero other components. Therefore, all groups have a Frobenius norm equal to $\sqrt{n}$ and, thus, he considered $d_g = 1$.

We noticed that the transformation $X_g/\sqrt{l_g}$ can be used in a standardized group LASSO. In this section, we will show that by using this transformation in a standardized group LASSO, the estimate of coefficients will not change. In fact, the group weights move from

the penalty term into the loss function.

As already mentioned, group LASSO assumes orthonormality within each group of covariates [18], i.e. $X_g^T X_g = I_g$. This orthonormalization changes group LASSO to a new problem

$$\min_{\beta} \frac{1}{2} \left\| Y - \sum_g X_g \beta_g \right\|_2^2 + \lambda \sum_g \sqrt{l_g} \|X_g \beta_g\|_2. \tag{4.2.1}$$

This new problem penalizes the loss function with the fit of each group $X_g \beta_g$ instead of group coefficients β_g. Our goal is to normalize X_g in such a way that the penalty weights change to $d_g = 1$. If X_g is full column rank, then it can be decomposed by QR factorization into $X_g = U_g R_g$ where $U_g^T U_g = I$ and R_g is an invertible upper triangular square matrix. We rewrite (4.2.1) as

$$\min_{\beta} \frac{1}{2} \left\| Y - \sum_g \frac{U_g}{\sqrt{l_g}} \tilde{\beta}_g \right\|_2^2 + \lambda \sum_g \|U_g \tilde{\beta}_g\|_2, \tag{4.2.2}$$

where

$$\hat{\beta}_g = \frac{1}{\sqrt{l_g}} R_g^{-1} \hat{\tilde{\beta}}_g.$$

In this reformulation, we see that the penalty weight $\sqrt{l_g}$ is transferred from the penalty term into the loss function. This new problem (4.2.2) is equivalent to (4.2.1). To prove this equivalency, define $\tilde{X}_g = \frac{U_g}{\sqrt{l_g}}$ thus, it simplifies to

$$\min_{\beta} \frac{1}{2} \left\| Y - \sum_g \tilde{X}_g \tilde{\beta}_g \right\|_2^2 + \lambda \sum_g \|\tilde{\beta}_g\|_2, \tag{4.2.3}$$

because $\|U_g \tilde{\beta}_g\|_2 = \|\tilde{\beta}_g\|_2$. Note that the group weights are $d_g = \left\|\tilde{X}_g\right\|_F = \left\|\frac{U_g}{\sqrt{l_g}}\right\|_F = 1$. This means that we can normalize each group of covariates as $\tilde{X}_g = \frac{U_g}{\sqrt{l_g}}$ and run a group LASSO without penalty weights. We show that the solution of (4.2.3) is equal to the solution of Simon and Tibshirani [28]; thus, it benefits from all of the advantages of the standardized

group LASSO as established by Simon and Tibshirani [28].

Theorem 4. *The convex problems (4.2.1) and (4.2.3) are equivalent.*

Proof.

The characterization of solution for (4.2.3) is

$$\tilde{X}_g^T(y - \tilde{X}\hat{\tilde{\beta}}) = \lambda \frac{\hat{\tilde{\beta}}_g}{\|\hat{\tilde{\beta}}_g\|_2} \qquad \text{for} \quad \hat{\tilde{\beta}}_g \neq 0,$$

$$\tilde{X}_g^T(y - \tilde{X}\hat{\tilde{\beta}}) = \lambda v_g \qquad \text{where} \quad v_g \in \mathbb{R}^{|g|}, \quad \|v_g\|_2 \leq 1 \qquad \text{for} \quad \hat{\tilde{\beta}}_g = 0.$$

Note that v_g is in the subdifferential. The first condition can be written as

$$S_g = \hat{\tilde{\beta}}_g \left(\frac{1}{l_g} + \frac{\lambda}{\|\hat{\tilde{\beta}}_g\|_2} \right),$$

where $S_g = \tilde{X}_g^T(y - \tilde{X}\hat{\tilde{\beta}}_{-g})$ and $\hat{\tilde{\beta}}_{-g} = (\hat{\tilde{\beta}}_1^T, \cdots, \hat{\tilde{\beta}}_{g-1}^T, 0^T, \cdots, \hat{\tilde{\beta}}_p^T)$. Therefore, $\frac{S_g}{\|S_g\|_2} = \frac{\hat{\tilde{\beta}}_g}{\|\hat{\tilde{\beta}}_g\|_2}$ and this gives

$$\frac{1}{l_g}\hat{\tilde{\beta}}_g = \left(1 - \frac{\lambda}{\|S_g\|_2}\right) S_g. \tag{4.2.4}$$

From second condition, $-S_g + \frac{1}{l_g}\hat{\tilde{\beta}}_g = \lambda v_g$ where $\|v_g\|_2 \leq 1$. Note that $\hat{\tilde{\beta}}_g = 0$ and, therefore, $\|S_g\|_2 \leq \lambda$ which gives

$$(1 - \frac{\lambda}{\|S_g\|_2}) \leq 0. \tag{4.2.5}$$

With respect to (4.2.4) and (4.2.5) one can write

$$\frac{1}{l_g}\hat{\tilde{\beta}}_g = \left(1 - \frac{\lambda}{\|S_g\|_2}\right)_+ S_g.$$

By considering $\hat{\beta}_g = \frac{1}{\sqrt{l_g}} R_g^{-1} \hat{\tilde{\beta}}_g$, $X_g = U_g R_g$, and $U_g^T U_g = I$, we have

$$\hat{\beta}_g = \left(1 - \frac{\lambda\sqrt{l_g}}{\|U_g^T r_{-g}\|_2}\right)_+ (X_g^T X_g)^{-1} X_g^T r_{-g},$$

which is equal to the equation 11 in Simon and Tibshirani [28]. This shows that the problem (4.2.3) is equivalent to

$$\min_{\beta} \frac{1}{2} \left\| Y - \sum_{g} X_g \beta_g \right\|_2^2 + \lambda \sum_{g} \sqrt{l_g} \|X_g \beta_g\|_2. \quad \square \tag{4.2.6}$$

It is worth to mention that we proved this result only for the standardized group LASSO. We are cautious in using it in group LASSO because $d_g = \left\| \frac{X_g}{\sqrt{l_g}} \right\|_F \neq 1$. Breheny and Huang [15] implemented the Simon and Tibshirani standardization [28] in the R package *grpreg*. Here, we follow the basic steps from their work. They considered eigendecomposition of covariance matrix of a group g as

$$\frac{1}{n} X_g^T X_g = Q_g \Lambda_g Q_g^T,$$

where $Q_g^T Q_g = I$ and Λ_g is a diagonal matrix of eigenvalues of $\frac{1}{n} X_g^T X_g$. We define $\tilde{X}_g$ in (4.2.3) as

$$\tilde{X}_g = \frac{1}{\sqrt{l_g}} X_g Q_g \Lambda_g^{-\frac{1}{2}}.$$

Hence, the original coefficient is

$$\hat{\beta}_g = \frac{1}{\sqrt{l_g}} Q_g \Lambda_g^{-\frac{1}{2}} \tilde{\beta}_g.$$

4.3 The characterization of the solution

Consider a group LASSO formulation

$$\min_{\beta} \frac{1}{2} \left\| Y - \sum_{g} \frac{X_g}{\sqrt{l_g}} \beta_g^* \right\|_2^2 + \lambda \sum_{g} \|X_g \beta_g^*\|_2.$$

We rewrite this optimization problem in the form

$$\min_{\beta} \frac{1}{2} \left\| Y - \sum_{g} \frac{X_g}{\sqrt{l_g}} \beta_g^* \right\|_2^2 + \lambda \sum_{g} u_g, \tag{4.3.1}$$

where

$$\|X_g\beta_g^*\|_2 \leq u_g.$$

By second order cone programming theory

$$\|X_g\beta_g^*\|_2 \leq u_g \Longleftrightarrow \begin{bmatrix} X_g\beta_g^* \\ u_g \end{bmatrix} \in \mathcal{C}_{n+1} \Longleftrightarrow -\begin{bmatrix} X_g\beta_g^* \\ u_g \end{bmatrix} \preceq_{\mathcal{C}_{n+1}} 0,$$

where

$$\mathcal{C}_{n+1} = \left\{ \begin{bmatrix} x \\ t \end{bmatrix} \mid x \in \mathbb{R}^n, t \in \mathbb{R}, \|x\|_2 \leq t \right\}.$$

The Lagrangian with the dual variables $(\gamma_g, \delta_g) \in \mathbb{R}^n \times \mathbb{R}$ where $\|\gamma_g\|_2 \leq \delta_g$ is

$$\mathcal{L}(\beta^*, u, \gamma, \delta) = \frac{1}{2}\|Y - \sum_{g\in\mathcal{G}} \frac{X_g}{\sqrt{l_g}}\beta_g^*\|_2^2 + \lambda \sum_{g\in\mathcal{G}} u_g - \sum_{g\in\mathcal{G}} \begin{bmatrix} X_g\beta_g^* \\ u_g \end{bmatrix}^T \begin{bmatrix} \gamma_g \\ \delta_g \end{bmatrix},$$

where the primal variables are (β^*, u). Derivatives with respect to the primal variables are as follows:

$$\begin{aligned} \nabla_{\beta_g^*}\mathcal{L}(\beta^*, u, \gamma, \delta) &= -\frac{X_g^T}{\sqrt{l_g}}(Y - \sum_{g\in\mathcal{G}} \frac{X_g}{\sqrt{l_g}}\beta_g^*) - X_g^T\gamma_g, \\ \nabla_{u_g}\mathcal{L}(\beta^*, u, \gamma, \delta) &= \lambda - \delta_g. \end{aligned} \tag{4.3.2}$$

Equating the second derivative to zero gives $\lambda = \delta_g$; thus $\|\gamma_g\|_2 \leq \lambda$. Suppose that $\beta_g = 0$ and $u_g = \epsilon$ for all $g \in \mathcal{G}$; therefore, the strict inequality in (4.3.1) holds. Hence, strong duality holds and the complementary slackness is

$$\gamma_g^T X_g\beta_g^* + \delta_g u_g = 0.$$

Note that $\|X_g\beta_g^*\|_2 \leq u_g$ and $\|\gamma_g\|_2 \leq \delta_g$. On the other hand, by the Cauchy-Schwartz

inequality, $|\gamma_g^T X_g \beta_g^*| \leq \|\gamma_g\|_2 \|X_g \beta_g^*\|_2$. Hence,

$$\gamma_g^T X_g \beta_g^* + \delta_g u_g \geq -\|\gamma_g\|_2 \|X_g \beta_g^*\|_2 + \delta_g u_g \geq 0.$$

Therefore, complementary slackness holds if and only if the following conditions hold [16]:

- $\|\gamma_g\|_2 < \delta_g \;\Rightarrow\; \|X_g \beta_g^*\|_2 = u_g = 0,$
- $\|X_g \beta_g^*\|_2 < u_g \;\Rightarrow\; \|\gamma_g\|_2 = \delta_g = 0,$
- $\|X_g \beta_g^*\|_2 = u_g, \;\; \|\gamma_g\|_2 = \delta_g \;\Rightarrow\; \gamma_g^T X_g \beta_g^* = -\delta_g u_g.$

The third condition with respect to (4.3.2) is

$$\gamma_g^T X_g \beta_g^* + \lambda \|X_g \beta_g^*\|_2 = 0. \tag{4.3.3}$$

The complementary slackness condition for the second order cone (4.3.3) holds if and only if [33]

- $X_g \beta_g^* = 0,$
- $X_g \beta_g^* \neq 0$, $\|\gamma_g\|_2 = \lambda$ and $\gamma_g = -\lambda \frac{X_g \beta_g^*}{\|X_g \beta_g^*\|_2}$.

The KKT conditions for $(\hat{\beta}^*(y), \hat{\gamma}(y))$ are as follows:

$$\begin{aligned} \frac{X_g^T}{\sqrt{l_g}}(Y - \sum_{g \in \mathcal{G}} \frac{X_g}{\sqrt{l_g}} \hat{\beta}_g^*) &= -X_g^T \hat{\gamma}_g, \\ \hat{\gamma}_g^T X_g \hat{\beta}_g^* + \lambda \|X_g \hat{\beta}_g^*\|_2 &= 0, \\ \|\hat{\gamma}_g\|_2 &\leq \lambda. \end{aligned} \tag{4.3.4}$$

Let us define the support of $\hat{\beta}^*$ as follows:

$$\mathcal{S}(\hat{\beta}^*) = \left\{ g \in \mathcal{G} : X_g \hat{\beta}_g^* \neq 0 \right\}. \tag{4.3.5}$$

Theorem 5. *Let us define the support $\mathcal{S}(\hat{\beta}^*)$ as (4.3.5). The optimal solution of (4.2.6), $\hat{\beta}^*$, satisfies*

$$\frac{X_g^T}{\sqrt{l_g}}(Y - \sum_{g\in\mathcal{S}} \frac{X_g}{\sqrt{l_g}}\hat{\beta}_g^*) = \lambda \frac{X_g^T X_g \hat{\beta}_g^*}{\|X_g\hat{\beta}_g^*\|_2} \qquad \text{where } g \in \mathcal{S}(\hat{\beta}^*),$$

$$\left\| \frac{X_g^T}{\sqrt{l_g}}(Y - \sum_{g\in\mathcal{S}} \frac{X_g}{\sqrt{l_g}}\hat{\beta}_g^*) \right\|_2 \leq \lambda \|X_g\|_2 \qquad \text{where } g \notin \mathcal{S}(\hat{\beta}^*),$$

and $\|X_g\|_2$ is the induced l_2-norm of X_g.

Proof.

1. In the case where $g \in \mathcal{S}$,

Note that $\sum_{g\in\mathcal{G}} \frac{X_g}{\sqrt{l_g}}\hat{\beta}_g^* = \sum_{g\in\mathcal{S}} \frac{X_g}{\sqrt{l_g}}\hat{\beta}_g^*$ and, when $X_g\beta_g^* \neq 0$, then $\gamma_g = -\lambda\frac{X_g\beta_g^*}{\|X_g\beta_g^*\|_2}$. Therefore, the KKT condition in terms of $\mathcal{S}$ is

$$\frac{X_g^T}{\sqrt{l_g}}(Y - \sum_{g\in\mathcal{S}} \frac{X_g}{\sqrt{l_g}}\hat{\beta}_g^*) = \lambda \frac{X_g^T X_g \hat{\beta}_g^*}{\|X_g\hat{\beta}_g^*\|_2} \qquad \text{where } g \in \mathcal{S}(\hat{\beta}^*).$$

2. In the case where $g \notin \mathcal{S}$,

Suppose that $\Omega(\beta^*) = \lambda \sum_{g\in\mathcal{G}} \|X_g\beta_g^*\|_2$. By Lemma 1 for case $X_g\beta_g^* = 0$, we have

$$\partial_{\beta_g^*}\Omega(\beta^*) = \left\{\lambda X_g^T\omega_g : \omega_g \in \mathbb{R}^n, \ \ \|\omega_g\|_2 \leq 1\right\}.$$

Let $v_h = \lambda w_h$; then

$$\partial_{\beta_g^*}\Omega(\beta^*) = \left\{X_g^T v_g : v_g \in \mathbb{R}^n, \ \ \|\frac{v_g}{\lambda}\|_2 \leq 1\right\}.$$

Note that $\|\gamma_g\|_2 \leq \lambda$, i.e. $\|\frac{\gamma_g}{\lambda}\|_2 \leq 1$ and $\gamma_g, v_g \in \mathbb{R}^n$; therefore, the above subdifferential is equivalent to

$$\partial_{\beta_g^*}\Omega(\beta^*) = \left\{X_g^T\gamma_g : \gamma_g \in \mathbb{R}^n, \ \ \|\frac{\gamma_g}{\lambda}\|_2 \leq 1\right\}.$$

Now, consider the KKT condition

$$\frac{X_g^T}{\sqrt{l_g}}(Y - \sum_{g\in\mathcal{S}} \frac{X_g}{\sqrt{l_g}}\hat{\beta}_g^*) = -X_g^T\hat{\gamma}_g,$$

and note that γ_g is in the subdifferential; therefore,

$$\left\| \frac{X_g^T}{\sqrt{l_g}}(Y - \sum_{g\in\mathcal{S}} \frac{X_g}{\sqrt{l_g}}\hat{\beta}_g^*) \right\|_2 \leq \lambda \sup_{\|\frac{\gamma_g}{\lambda}\|_2\leq 1} \left\| X_g \frac{\gamma_g}{\lambda} \right\|_2 = \lambda \|X_g\|_2 ,$$

where $\|X_g\|_2$ is the induced l_2-norm of matrix X_g. □

4.4 The degrees of freedom

In this section, we will calculate the degrees of freedom for the standardized group LASSO.

Theorem 6. *The unbiased estimate of the degrees of freedom for the standardized group LASSO is*

$$\widehat{df}_\lambda(\widetilde{X}_{\mathcal{S}}\hat{\beta}^*(y)) = tr\left(\widetilde{X}_{\mathcal{S}}^T \left(\widetilde{X}_{\mathcal{S}}^T \widetilde{X}_{\mathcal{S}} + \lambda U\right)^{-1} \widetilde{X}_{\mathcal{S}}\right), \tag{4.4.1}$$

where

$$U = diag\left(X_g^T \left[\frac{I - \frac{X_g\hat{\beta}_g^*\hat{\beta}_g^{*T}X_g^T}{\|X_g\hat{\beta}_g^*\|_2^2}}{\|X_g\hat{\beta}_g^*\|_2} \right] X_g^T \right)_{g\in\mathcal{S}}. \tag{4.4.2}$$

Proof.

Suppose $\mathcal{S}$ be the support of $\hat{\beta}^*$ defined in (4.3.5) and define $\widetilde{X}_{\mathcal{S}} = \left(\frac{X_g}{\sqrt{l_g}}\right)_{g\in\mathcal{S}}$. Therefore, the first KKT condition in Theorem 5 can be written as follows:

$$\widetilde{X}_{\mathcal{S}}^T(Y - \widetilde{X}_{\mathcal{S}}\hat{\beta}_{\mathcal{S}}^*) = \lambda X_{\mathcal{S}}^T \mathcal{N}(\hat{\beta}_{\mathcal{S}}^*),$$

where $\mathcal{N}(\hat{\beta}_{\mathcal{S}}^*)$ is a normalization operator where

$$\mathcal{N}_g(\beta) = \frac{X_g\beta_g}{\|X_g\beta_g\|_2}.$$

Define the mapping

$$\Gamma\left(\hat{\beta}^*(y), y\right) = \widetilde{X}_{\mathcal{S}}^T \widetilde{X}_{\mathcal{S}}\hat{\beta}_{\mathcal{S}}^* - \widetilde{X}_{\mathcal{S}}^T Y + \lambda X_{\mathcal{S}}^T \mathcal{N}(\hat{\beta}_{\mathcal{S}}^*);$$

then

$$\partial_y \hat{\beta}^*(y) = \left(\partial_{\hat{\beta}^*} \Gamma\left(\hat{\beta}^*(y), y\right)\right)^{-1} \widetilde{X}_S^T,$$

and thus

$$\widehat{df}_\lambda(\widetilde{X}_S \hat{\beta}^*(y)) = \operatorname{tr}\left(\widetilde{X}_S^T \left(\widetilde{X}_S^T \widetilde{X}_S + \lambda U\right)^{-1} \widetilde{X}_S\right), \tag{4.4.3}$$

where U is defined in (4.4.2). □

Note that the matrix U involves X_g instead of $\widetilde{X}_g$ because the penalty weight $\sqrt{l_g}$ is transferred from the penalty into the loss function. In Figure 4.1, the unbiased estimate of the degrees of freedom for the standardized group LASSO is compared with the actual degrees of freedom

$$df(f) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \operatorname{Cov}\left(Y_i, f(Y_i)\right). \tag{4.4.4}$$

For this comparison, suppose $X \in \mathbb{R}^{100\times 75}$ and $\beta \in \mathbb{R}^{75}$ are fixed for 25 groups, each with 3 levels. Then, we generate $B = 1000$ Monte Carlo replicates of $y^{(b)} = X\beta + \mathrm{N}(0,1)$. The Monte Carlo estimate of $E\left[\hat{df}_\lambda\right]$ and actual df_λ are computed from the generated values of y by using formulas (4.4.3) and (4.4.4). Figure 4.1 shows that the actual degrees of freedom and the Monte Carlo estimate coincide, i.e. $E\left[\hat{df}_\lambda\right] = df_\lambda$. It means that the estimate 4.1 is unbiased, which is the main property of SURE theory.

The error bars in Figure 4.1 are smaller for larger values of tuning parameter λ. The error bars get larger specifically for models that have more than 16 predictors. These large error bars happen because the interval of tuning parameters for these models are very small. The *grpreg* algorithm selects groups based on the value of the tuning parameter λ. First, the algorithm estimates the largest value of λ, then it makes a grid of m points between zero and the estimated λ. The set of grid points gives the solution path. If the number of grid

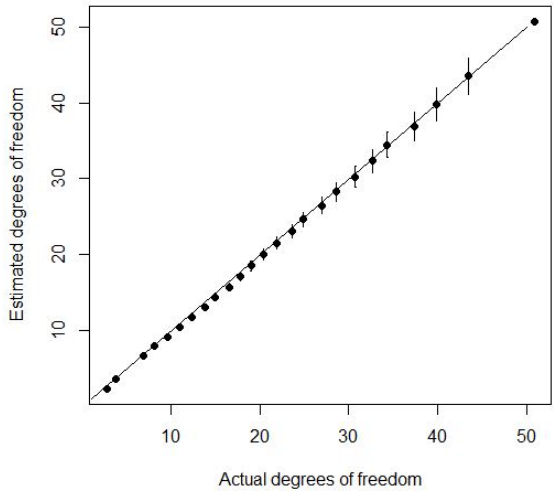


Figure 4.1: Comparing the unbiased estimate of degrees of freedom with the actual degrees of freedom for the standardized group LASSO.

points m is small, then it is possible that the number of selected groups shows a big jump from a point to another point in the solution path. For instance, the model selects 14 groups in one point and in the next point, the model contains 18 groups. It shows that algorithm missed the small jumps of λ that selects 15, 16 and 17 groups. This can be solved by taking a large number of grid points, but this solution is expensive computationally. The alternative is to consider an algorithm similar to LARS. Consider a case that q groups are selected in a very small interval of λ; thus, the degrees of freedom are sensitive to a small change of value in λ, which makes the error bars large. We believe that a LARS type algorithm for group LASSO is useful to make the error bars smaller because it will find the break points in the solution path easier and cheaper. For this simulation, we used *grpreg* algorithm with a grid of 10000 points for the tuning parameter λ. The full model with 25 groups shows a very small error bar because of the large interval of λ for this full model. Also the full model takes a very small λ close to zero and we will discuss the property of degrees of freedom for this point in Section 5.5.

4.5 The degrees of freedom for overlapped group LASSO

As a corollary of Vaiter et al. result [4], we calculate the degrees of freedom for an overlapped group LASSO.

Corollary 1. *The unbiased estimate of degrees of freedom for overlapped group LASSO is*

$$\widehat{df}_{\lambda}(\textit{overlapped group LASSO}) = tr\left(X_{\mathcal{S}}^{*}\left(X_{\mathcal{S}}^{*T}X_{\mathcal{S}}^{*}+\lambda U\right)^{-1}X_{\mathcal{S}}^{*T}\right), \tag{4.5.1}$$

where

$$U = \text{diag}\left[\frac{I-\frac{\hat{\beta}_g^*\hat{\beta}_g^{*T}}{\|\hat{\beta}_g^*\|_2^2}}{\|\hat{\beta}_g^*\|_2}\right]_{g\in\mathcal{S}}. \tag{4.5.2}$$

Proof.

Consider the problem

$$\|Y-X\beta\|_2^2+\lambda\sum_{g\in\mathcal{G}}\|\beta_g\|_2, \tag{4.5.3}$$

where the groups $g\in\mathcal{G}$ overlap. Obozinski [34] implemented this problem as group LASSO by duplicating the covariate matrix X as X^*, where

$$X\longrightarrow X^*=\bigoplus_{g\in\mathcal{G}}(x_i)_{i\in g},$$

and $\bigoplus:\mathbb{R}^P\longrightarrow\mathbb{R}^{\Sigma_{g\in\mathcal{G}}|g|}$ is the duplication operator. Thus, the problem (4.5.3) can be written as

$$\|Y-X^*\beta^*\|_2^2+\lambda\sum_{g\in\mathcal{G}}\|\beta_g^*\|_2,$$

where β^* refers to the coefficients with respect to the coordinates of groups g. Note that $X\hat{\beta}=X^*\hat{\beta}^*$ and by the result of Vaiter et al. [4], the degrees of freedom are as follows:

$$df_{\lambda}(X_{\mathcal{S}}^*\hat{\beta}_{\mathcal{S}}^*(y))=E\left[\nabla\cdot X_{\mathcal{S}}^*\hat{\beta}_{\mathcal{S}}^*\right]$$

$$= E\left[\text{tr}\left(X_{\mathcal{S}}^{*}\left[\partial_{\hat{\beta}_{\mathcal{S}}^{*}}\Gamma(\hat{\beta}_{\mathcal{S}}^{*}(y),y)\right]^{-1}X_{\mathcal{S}}^{*T}\right)\right],$$

where $\partial_{\hat{\beta}_{\mathcal{S}}^{*}}\Gamma(\hat{\beta}_{\mathcal{S}}^{*}(y),y) = X_{\mathcal{S}}^{*T}X_{\mathcal{S}}^{*} + \lambda\text{diag}\left[\frac{I-\frac{\hat{\beta}_g^{*}\hat{\beta}_g^{*T}}{\|\hat{\beta}_g^{*}\|_2^2}}{\|\hat{\beta}_g^{*}\|_2}\right]_{g\in\mathcal{S}}$ and $\mathcal{S} = \left\{g \in \mathcal{G} : \hat{\beta}_g^{*} \neq 0\right\}$. This results in

$$\widehat{df}_{\lambda}(\text{overlapped group LASSO}) = \text{tr}\left(X_{\mathcal{S}}^{*}\left(X_{\mathcal{S}}^{*T}X_{\mathcal{S}}^{*} + \lambda U\right)^{-1}X_{\mathcal{S}}^{*T}\right), \tag{4.5.4}$$

where U is defined in (4.5.2). □

Chapter 5

Hierarchical Group LASSO with the Logistic Loss

5.1 Introduction

In this chapter, we apply the hierarchy constraint on group LASSO with logistic loss. The solution of the proposed convex problem is characterized in Theorem 7. The derived unbiased estimates of the degrees of freedom in the current work are compared with the known unbiased estimates of the degrees of freedom. We show in Theorem 8 that our estimates of the degrees of freedom converge to the rank of the design matrix of selected variables when $\lambda \searrow 0$. Finally, we study the selection process of the proposed procedures.

Meier et al. [32] proposed an algorithm, implemented in the R-package *grplasso*, for group LASSO with the logistic loss and they showed that the group LASSO estimator is statistically consistent. They tried to produce a hierarchical model by a two-stage algorithm. In the first stage, they used a group LASSO penalty for all main effects, and after that they penalized all the main effects and related interactions with ridge penalty in the second stage. Ridge penalty shrinks all effects but prevents selection. It means that a hierarchical model is made only by selecting main effects. There is an issue here. To explain this issue, consider that a group LASSO selects 6 main effects in the first stage; hence, there are 15 interactions.

The ridge penalty produces a model with 6+15=21 effects. Someone may ask: do all of the 15 interactions have to be selected?

5.2 Logistic regression

Consider a dichotomous response variable Y, which divides the population under question into two classes. That is, Y is an indicator variable such that if the response is in the first class, then $Y = 1$; otherwise $Y = 0$. Let X be the matrix of predictors, continuous or discrete. First, note that $P(Y = 1|X = x) = E[Y = 1|X = x]$ and assume that there is a relation through a function p and a parameter β as $P(Y = 1|X = x) = p(x, \beta)$. Thus, the likelihood function is

$$\prod_{i=1}^{n} P(Y = y_i|X = x_i) = \prod_{i=1}^{n} p(x_i, \beta)^{y_i}(1 - p(x_i, \beta))^{1-y_i}.$$

The function p cannot be represented by linear regression since $p \in (0, 1)$ and linear functions are unbounded in both directions. The term $\log(p)$ appears in the log-likelihood, which cannot be represented by linear models, because $\log(p)$ is unbounded in one direction. The logistic function $\log(\frac{p}{1-p})$ is the proper function for this purpose; therefore,

$$\log\left[\frac{p(x)}{1 - p(x)}\right] = \beta_0 + x\beta,$$

and the log-likelihood is

$$\mathcal{L}(\beta_0, \beta) = Y^T(\beta_0 1 + X\beta) - 1^T\left(\log(1 + e^{\beta_0 1 + X\beta})\right),$$

where $\log$ and $\exp$ are taken component-wise.

5.3 The proposed method

For the logistic loss, we will use the same modification and generalization of the constraint of Bien et al. [27] as for the quadratic loss in Chapter 3. The objective is again to make

models which satisfy the strong hierarchy rule. Let us write the linear model with pair interactions

$$\log\left[\frac{P(Y=1|X)}{1-P(Y=1|X)}\right] = \beta_0 + \sum_{g=1}^{p} X_g\beta_g + \frac{1}{2}\sum_{g\neq h} X_{g:h}\,\mathrm{vec}(\Theta_{g:h})$$

where $\Theta_{g:h}$ is the matrix of coefficients of pair interactions, same as the parameter in the equation (3.2.1). The negative log-likelihood is considered as the loss function; thus

$$\min_{\beta_0\in\mathbb{R},\beta\in\mathbb{R}^L,\Theta\in\mathbb{R}^{L\times L}} -\Bigg[Y^T\left(\beta_0 1 + \sum_{g=1}^{p} X_g\beta_g + \frac{1}{2}\sum_{g\neq h} X_{g:h}\,\mathrm{vec}(\Theta_{g:h})\right) - 1^T\left(\log(1+e^{\beta_0 1+\sum_{g=1}^{p} X_g\beta_g+\frac{1}{2}\sum_{g\neq h} X_{g:h}\,\mathrm{vec}(\Theta_{g:h})})\right)\Bigg] + \lambda\sum_g \|\beta_g\|_2 + \frac{\lambda}{2}\sum_{g\neq h}\|\Theta_{g:h}\|_F.$$

The above group LASSO considers main effects and interactions as groups and the induced model will include some main effects and interactions without a hierarchy structure. To guarantee hierarchy, we will use the constraint as constructed in Section 3.2. Hence, we take $1^T\mathrm{vec}((\Theta_g^+ - \Theta_g^-)\mathrm{diag}(\frac{1}{w})) \leq 1^T(\beta_g^+ + \beta_g^-)$ as the modified version of the constraint of Bien et al. [27], where Θ is a symmetric matrix and w is the weight vector defined in Section 3.2. For the simplicity of notation, let $\phi = (\beta_0, \beta^+, \beta^-, \mathrm{vec}(\Theta^+), \mathrm{vec}(\Theta^-))$ where $\Theta_{g:g} = 0 \notin \mathrm{vec}(\Theta)$ and suppose that $\widetilde{X} = (1; X; -X; Z; -Z)$, where Z is the matrix of interactions defined in (2.3.1). Hence, the hierarchical group LASSO with logistic loss is

$$\min_{\beta_0\in\mathbb{R},\beta^\pm\in\mathbb{R}^L,\Theta^\pm\in\mathbb{R}^{L\times L}} -\left[Y^T\widetilde{X}\phi - 1^T\left(\log(1+e^{\widetilde{X}\phi})\right)\right] + \lambda\sum_g\|\beta_g^+ - \beta_g^-\|_2 + \lambda\sum_{g\neq h}\|\Theta_{g:h}^+ - \Theta_{g:h}^-\|_F,$$

$$\left.\begin{array}{c} 1^T\mathrm{vec}((\Theta_g^+ - \Theta_g^-)\mathrm{diag}(\frac{1}{w})) \leq 1^T(\beta_g^+ + \beta_g^-) \\ \beta_g^\pm \geq 0, \quad \Theta_g^\pm \geq 0 \end{array}\right\} \quad \text{for all } g \in \mathcal{G}, \qquad (5.3.1)$$

$$\Theta^+ - \Theta^- = \Theta^{+T} - \Theta^{-T}.$$

Note that here $\frac{\Theta_g}{2}$ is considered as a new parameter instead of Θ_g. We know that the uniqueness of the solution is guaranteed if the problem is strictly convex. Note that the sec-

ond derivative of the logistic loss is $\widetilde{X}^T \operatorname{diag}\left(\frac{e^{\widetilde{X}_i\phi}}{(1+e^{\widetilde{X}_i\phi})^2}\right)_{i=1}^{n} \widetilde{X} = \widetilde{X}^T V \widetilde{X} = \widetilde{X}^T V^{\frac{1}{2}} V^{\frac{1}{2}} \widetilde{X} = U^T U \geq 0$; thus the loss function is convex, and it would be strictly convex if and only if $\widetilde{X}$ is full column rank. If $\widetilde{X}$ is not full column rank, then there exists vector x such that $\widetilde{X}x = 0$ and, therefore, $x^T \widetilde{X}^T V \widetilde{X} x = 0$. As a result, in the high-dimensional case with $n << p$, the solution of (5.3.1) may not be unique.

In Chapter 4, a tiny fraction of an elastic net term made the problem strictly convex. This tiny fraction, ε, controls the effect of the elastic net term. The elastic net term is transferred into the design matrix X which simplifies the optimization problem. The form of the logistic loss prevents the transferring of the elastic net term into the design matrix. There is another approach for uniqueness in which every submatrix X_a, where $|a| \leq n$, is assumed to have full column rank. Therefore, the active set is chosen in such a way that $|\mathcal{A}| \leq n$. For instance [8], [7] and [22] proposed different constraints in view of this approach. Roth and Fischer [8] proposed a simpler constraint and algorithm for group LASSO. We will use and describe this algorithm since it adds only a fixed parameter κ as constraint to the group LASSO penalty and this simplifies the algorithm mathematically and computationally.

In logistic regression, when the dimensionality exceeds the number of observations, the uniqueness of the solution is not guaranteed. In this situation, every $\beta^* = \beta^0 + \xi$ is also a solution where $\xi \in \ker(X)$. Therefore, Roth and Fischer [8] defined the group LASSO as

$$\begin{aligned} &\text{minimize } \mathcal{L}(\beta_0, \beta) \quad \text{s.t.} \quad g(\beta) \geq 0 \\ &\text{where} \quad g(\beta) = \kappa - \sum_g \|\beta_g\|_2, \end{aligned} \tag{5.3.2}$$

where active constraint is required, i.e. $\kappa < \kappa_0$ and $\kappa_0 := \min_{\xi \in \ker(X)} \sum_g \|\beta_g^0 + \xi_g\|_2$. It is worth noting that κ_0 is unique even with several vectors $\xi \in \ker(X)$. Roth and Fischer [8] proved that the solution of (5.3.2) is unique and proposed the following algorithm. Define

$h_j = X_j^T(Y - \frac{e^{X\beta}}{1+e^{X\beta}})$.

A: Initialize set $\mathcal{A} = \{j_0\}$, β_{j_0} arbitrary with $\|\beta_{j_0}\|_2 = \kappa$.

B: Optimize over the current active set $\mathcal{A}$. Define set $\mathcal{A}^+ = \{j \in \mathcal{A} : \|\beta_j\|_2 > 0\}$. Define $\lambda = \max_{j\in\mathcal{A}^+} \|h_j\|_2$. Adjust the active set $\mathcal{A} = \mathcal{A}^+$.

C: Lagrangian violation. For all $j \notin \mathcal{A}$, check if $\|h_j\|_2 \leq \lambda$. If this is the case, we have found a global solution. Otherwise, include the group with the largest violation to $\mathcal{A}$ and go to **B**.

D: Completeness and uniqueness. For all $j \notin \mathcal{A}$, check if $\|h_j\|_2 = \lambda$. If so, there might exist other solutions with identical costs that include these groups in the active set. Otherwise, the active set is complete in the sense that it contains all relevant groups. If $|\mathcal{A}| \leq n$, then the solution is unique.

5.4 Characterization of the solution

We need to rewrite problem (5.3.1) with respect to the uniqueness algorithm of Roth and Fischer [8]. Note that the Lagrangian for group LASSO defined in (5.3.2) is

$$\mathcal{L}(\beta_0, \beta) - \lambda(\kappa - \sum_g \|\beta_g\|_2).$$

Therefore, the convex problem (5.3.1) with respect to this new definition of group LASSO changes to

$$\min_{\beta_0\in\mathbb{R},\beta^\pm\in\mathbb{R}^L,\Theta^\pm\in\mathbb{R}^{L\times L}} -\left[Y^T\widetilde{X}\phi - 1^T\left(\log(1+e^{\widetilde{X}\phi})\right)\right] - \lambda\left[\kappa - \sum_g \|\beta_g^+ - \beta_g^-\|_2 - \sum_{g\neq h} \|\Theta_{g:h}^+ - \Theta_{g:h}^-\|_F\right]$$

$$\left.\begin{array}{c} 1^T\text{vec}((\Theta_g^+ - \Theta_g^-)\text{diag}(\frac{1}{w})) \leq 1^T(\beta_g^+ + \beta_g^-) \\ \beta_g^\pm \geq 0, \quad \Theta_g^\pm \geq 0 \end{array}\right\} \quad \text{for all } g \in \mathcal{G}, \qquad (5.4.1)$$

$$\Theta^{+} - \Theta^{-} = \Theta^{+T} - \Theta^{-T}.$$

Similar to the equation (3.4.1), this convex problem can be rewritten as the minimization of

$$-\left[Y^T \widetilde{X}\phi - 1^T\left(\log(1+e^{\widetilde{X}\phi})\right)\right] - \lambda\left[\kappa - \sum_g \|D_g\phi_g\|_2 - \sum_{g\neq h}\|D_{g:h}\phi_{g:h}\|_2\right] - \mu^T A\phi + \upsilon^T E\phi.$$

The KKT conditions are the same as (3.4.5) except the first condition which changes to

$$\widetilde{X}^T(Y - \frac{e^{\widetilde{X}\hat{\phi}}}{1+e^{\widetilde{X}\hat{\phi}}}) = -A^T\hat{\mu} + E^T\hat{\upsilon} - D^T\hat{\gamma}. \tag{5.4.2}$$

The boundary set $\mathcal{A}(\hat{\phi})$ and the support $\mathcal{S}(\hat{\phi})$ are similar to (3.4.6) and (3.4.8). Now, we are equipped to write the following theorem.

Theorem 7. *Define the support $\mathcal{S}(\hat{\phi})$ as (3.4.8). Therefore, the optimal solution $\hat{\phi}$ of the convex problem (5.4.1) satisfies*

$$P_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^T(Y - \frac{e^{\widetilde{X}_{\mathcal{S}}\hat{\phi}_{\mathcal{S}}}}{1+e^{\widetilde{X}_{\mathcal{S}}\hat{\phi}_{\mathcal{S}}}}) = \lambda D_{\mathcal{S}}^T\mathcal{N}(\hat{\phi}_{\mathcal{S}}) \qquad \textit{where } g \in \mathcal{S}(\hat{\phi}),$$

$$\left\|P_g\overline{X}_g^T(Y - \frac{e^{\overline{X}_{\mathcal{S}}\hat{b}_{\mathcal{S}}}}{1+e^{\overline{X}_{\mathcal{S}}\hat{b}_{\mathcal{S}}}})\right\|_2 \leq \lambda \qquad \textit{where } g \notin \mathcal{S}(\hat{\phi}).$$

Proof.

The proof is analogous to the proof of Theorem 2. □

5.5 Discussion

5.5.1 Selection process

We rewrite the UMP test given in Simon and Tibshirani [28]. Suppose that a least squares linear regression model has fitted on $X = (X_1\ X_2\cdots X_{h-1})$ and we are deciding whether or not to add a new group of variables X_h. Suppose that the variance σ^2 is known and, thus,

the uniformly most powerful test of $H_0 : \beta_h = 0$ is rejected at level α if

$$\|\hat{y}_h - \hat{y}_{h-1}\|_2^2 \geq \sigma^2 \chi^2_{l_h}(1-\alpha), \tag{5.5.1}$$

where $\hat{y}_i$ is the prediction for the linear model on $X = (X_1\ X_2 \cdots X_i)$ [28]. Now consider the characterization of solution for the standardized group LASSO in Theorem 5. The decision for the inclusion of a group X_h is based on the magnitude of

$$\left\| X_h^T \left(Y - \sum_{g \in \mathcal{S}} \frac{X_g}{\sqrt{l_g}} \hat{\beta}_g^* \right) \right\|_2 . \tag{5.5.2}$$

The decision in (5.5.1) is based on the deviation between two fits, but in Theorem 5 it is based on a covariance and they look different. Note that if we consider $\tilde{\beta}$ instead of β^*, as explained in Section 4.1, then the magnitude (5.5.2) will change to

$$\left\| U_h^T \left(Y - \sum_{g \in \mathcal{S}} \frac{U_g}{\sqrt{l_g}} \hat{\tilde{\beta}}_g \right) \right\|_2 = \left\| Y - \hat{Y}_{\mathcal{S}} \right\|_2$$

which is comparable with the left side of (5.5.1). This result is mentioned in Simon and Tibshirani [28]. Suppose that

$$Y = \hat{Y}_{\mathcal{S}} + \varepsilon \qquad \text{where } \varepsilon \sim N(0, \sigma^2 I),$$

hence, for selection process in the standardized group LASSO, derived in Theorem 5, we will have

$$\begin{aligned}
\frac{1}{l_h \|X_h\|_2^2} \|X_h^T \varepsilon\|_2^2 &= \frac{1}{l_h \|X_h\|_2^2} \varepsilon^T X_h X_h^T \varepsilon \\
&= \frac{1}{l_h \max_i(\delta_{h,i})} \sum_{i=1}^{l_h} \delta_{h,i} \varepsilon^T \upsilon_{h,i} \upsilon_{h,i}^T \varepsilon \\
&= \frac{1}{l_h \max_i(\delta_{h,i})} \sum_{i=1}^{l_h} \delta_{h,i} < \upsilon_{h,i}, \varepsilon >^2 \\
&\sim \frac{1}{l_h} \sum_{i=1}^{l_h} \frac{\delta_{h,i}}{\max_i(\delta_{h,i})} \sigma^2 \frac{[\mathrm{N}(0,\sigma^2)]^2}{\sigma^2}
\end{aligned}$$

$$\sim \frac{\sigma^2}{l_h} \sum_{i=1}^{l_h} \delta^*_{h,i} \chi^2_1, \tag{5.5.3}$$

where $\delta_{h,i},\ \upsilon_{h,i}$ are eigenvalues and eigenvectors of $X_h^T X_h$. Standardized group LASSO uses $\frac{1}{l_h}\sum_{i=1}^{l_h} \delta^*_{h,i}\chi^2_1$ instead of $\chi^2_{l_h}$ in the above UMP test. This result is analogous to the result of Simon and Tibshirani [28]. Note that we worked with $\hat{\beta}^*$ directly instead of $\hat{\tilde{\beta}}$ and this gives the authority to show interesting properties in balanced layouts, which will be discussed in Chapter 6. Let us check this for hierarchical group LASSO. Its characterization of the solution is provided in Theorem 2 and we have

$$\begin{aligned} \|P_h X_h^T \varepsilon\|_2^2 &= \varepsilon^T X_h P_h^T P_h X_h^T \varepsilon \\ &= \sum_i \delta_{h,i} \varepsilon^T \upsilon_{h,i}^T \upsilon_{h,i} \varepsilon \\ &\sim \sigma^2 \sum_i \delta_{h,i} \chi^2_1, \end{aligned} \tag{5.5.4}$$

where $\delta_{h,i},\ \upsilon_{h,i}$ are eigenvalues and eigenvectors of $(P_h X_h^T)^T (P_h X_h^T)$.

5.5.2 The degrees of freedom

Stein's Unbiased Risk Estimation, SURE, theory requires normal response Y and, in logistic regression, the response Y is binary. Therefore, it is not possible to compute the degrees of freedom for the hierarchical logistic group LASSO. In Table 5.1, the forms of known estimates of degrees of freedom with the computed estimates of degrees of freedom are compared. This table shows how the degrees of freedom gradually change from simpler formulas to more complex formulas. In Section 1.3, we saw that the degrees of freedom of linear multiple regression are

$$\widehat{\mathrm{df}} = \mathrm{tr}\left((X^T X)(X^T X)^{-1}\right) = \mathrm{rank}(X).$$

Known unbiased estimates of degrees of freedom	
LASSO	$\text{tr}\,(X_{\mathcal{S}}(X_{\mathcal{S}})^{+}) = \text{rank}(X_{\mathcal{S}})$
Hierarchical LASSO	$\text{tr}\left((\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}})(\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}})^{+}\right) = \text{rank}(\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}})$
Elastic net	$\text{tr}\left(X_{\mathcal{S}}(X_{\mathcal{S}}^{T}X_{\mathcal{S}} + \lambda_2 I)^{-1}X_{\mathcal{S}}^{T}\right)$
Group LASSO	$\text{tr}\left(X_{\mathcal{S}}(X_{\mathcal{S}}^{T}X_{\mathcal{S}} + \lambda U)^{-1}X_{\mathcal{S}}^{T}\right)$ where $U = \text{diag}\left[\frac{I - \frac{\hat{\beta}_g\hat{\beta}_g^{T}}{\|\hat{\beta}_g\|_2^2}}{\|\hat{\beta}_g\|_2}\right]_{g\in\mathcal{S}}$
Calculated unbiased estimates of degrees of freedom	
Standardized group LASSO	$\text{tr}\left(\widetilde{X}_{\mathcal{S}}(\widetilde{X}_{\mathcal{S}}^{T}\widetilde{X}_{\mathcal{S}} + \lambda U)^{-1}\widetilde{X}_{\mathcal{S}}^{T}\right)$ where $U = \text{diag}\left(X_g^{T}\left[\frac{I - \frac{X_g\hat{\beta}_g^{\star}\hat{\beta}_g^{\star T}X_g^{T}}{\|X_g\hat{\beta}_g^{\star}\|_2^2}}{\|X_g\hat{\beta}_g^{\star}\|_2}\right]X_g^{T}\right)_{g\in\mathcal{S}}$
Overlapped group LASSO	$\text{tr}\left(X_{\mathcal{S}}^{\star}(X_{\mathcal{S}}^{\star T}X_{\mathcal{S}}^{\star} + \lambda U)^{-1}X_{\mathcal{S}}^{\star T}\right)$ where $U = \text{diag}\left[\frac{I - \frac{\hat{\beta}_g^{\star}\hat{\beta}_g^{\star T}}{\|\hat{\beta}_g^{\star}\|_2^2}}{\|\hat{\beta}_g^{\star}\|_2}\right]_{g\in\mathcal{S}}$
Hierarchical group LASSO	$\text{tr}\left(\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}}(P_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^{T}\widetilde{X}_{\mathcal{S}}P_{\mathcal{S}} + \lambda U)^{-1}P_{\mathcal{S}}\widetilde{X}_{\mathcal{S}}^{T}\right)$ where $U = \text{diag}\left(D_g^{T}\left[\frac{I - \frac{D_g\hat{\phi}_g\hat{\phi}_g^{T}D_g^{T}}{\|D_g\hat{\phi}_g\|_2^2}}{\|D_g\hat{\phi}_g\|_2}\right]D_g^{T}\right)_{g\in\mathcal{S}}$

Table 5.1: Comparing the known unbiased estimates of degrees of freedom with the results of this thesis.

Tibshirani et al. [25] showed that the degrees of freedom for LASSO are $\text{tr}\left((X_{\mathcal{S}})(X_{\mathcal{S}})^{+}\right) = \text{rank}(X_{\mathcal{S}})$. Also Bien et al. [27] showed the same property for hierarchical LASSO, which is $\text{tr}\left((\tilde{X}_{\mathcal{S}}P_{\mathcal{S}})(\tilde{X}_{\mathcal{S}}P_{\mathcal{S}})^{+}\right) = \text{rank}(\tilde{X}_{\mathcal{S}}P_{\mathcal{S}})$. Here we see that linear regression takes the rank of X, which is the complete design matrix. However, LASSO represents the rank of a design matrix which contains only selected variables. Hierarchical LASSO takes the rank of the projected design matrix for selected variables. This projection matrix reflects the hierarchy conditions.

For elastic net, when $\lambda_1, \lambda_2 \longrightarrow 0^{+}$, the degrees of freedom converge to those of the least squares regression and, therefore, its degrees of freedom should converge to the degrees of

freedom of linear regression. Hence,

$$\begin{aligned}\lim_{\lambda_1,\lambda_2 \searrow 0} \operatorname{tr}\left(X_{\mathcal{S}}(X_{\mathcal{S}}^T X_{\mathcal{S}} + \lambda_2 I)^{-1} X_{\mathcal{S}}^T\right) &= \operatorname{tr}\left(X_{\mathcal{S}} \lim_{\lambda_2 \searrow 0}\left[(X_{\mathcal{S}}^T X_{\mathcal{S}} + \lambda_2 I)^{-1} X_{\mathcal{S}}^T\right]\right) \\ &= \operatorname{tr}\left(X_{\mathcal{S}}(X_{\mathcal{S}})^+\right) \\ &= \operatorname{rank}\left(X_{\mathcal{S}}\right). \end{aligned} \tag{5.5.5}$$

Note that $\lim_{\varepsilon \searrow 0}\left[(A^T A + \varepsilon I)^{-1} A^T\right] = A^+$, where A^+ is the Moore-Penrose pseudo inverse. This property shows that the formula works well in the extreme case. However, does this property hold for group LASSO type procedures as well? Note that the estimate of the degrees of freedom in groupwise regularization contains matrix U instead of identity matrix I in elastic net regularization. First, we give the following lemma and then we will show this property for estimates of degrees of freedom in groupwise regularized procedures.

Lemma 4. *Suppose that U is a symmetric positive definite matrix. Then*

$$\lim_{\varepsilon \searrow 0}\left[(A^T A + \varepsilon U)^{-1} A^T\right] = V^{-1}(AV^{-1})^+$$

where $U = V^T V$.

Proof.

The matrix U is symmetric positive definite. Therefore $U = V^T V$ where V is invertible and

$$\begin{aligned}\lim_{\varepsilon \searrow 0}\left[(A^T A + \varepsilon U)^{-1} A^T\right] &= \lim_{\varepsilon \searrow 0}\left[(A^T A + \varepsilon V^T V)^{-1} A^T\right] \\ &= \lim_{\varepsilon \searrow 0}\left[\left[V^T(V^{-T} A^T A V^{-1} + \varepsilon I)V\right]^{-1} A^T\right] \\ &= \lim_{\varepsilon \searrow 0}\left[V^{-1}((AV^{-1})^T(AV^{-1}) + \varepsilon I)^{-1}(AV^{-1})^T\right] \\ &= V^{-1}\lim_{\varepsilon \searrow 0}\left[\left((AV^{-1})^T(AV^{-1}) + \varepsilon I\right)^{-1}(AV^{-1})^T\right] \\ &= V^{-1}(AV^{-1})^+. \quad \square \end{aligned}$$

Note that the matrix U is a function of $\hat{\beta}$ and, thus, it depends on λ. Note that in the case of

balanced design, when $\lambda \searrow 0$, then $\hat{\beta}^{GL} \to \hat{\beta}^{LSE}$, thus, we can take $\hat{\beta}^{LSE}$ and construct U. This gives a constant matrix U.

Theorem 8. *The unbiased estimate of degrees of freedom of group LASSO in the extreme case of* $\lambda \searrow 0$ *is*

$$\lim_{\lambda \searrow 0} df_{\lambda}\,(\textit{group LASSO}) \;=\; \textit{rank}\,(X_{\mathcal{S}})\,.$$

Proof.

The matrix U is symmetric positive semi-definite for group LASSO [4]. By some neglect, we can consider it as a symmetric positive definite matrix since $U+\varepsilon I$ is symmetric positive definite for a positive small value of ε. Then, $U = V^T V$ where V is invertible. Later on, we will see that the final answer is independent of the matrix V. We know that $\hat{\beta}_g \neq 0$ for all $g \in \mathcal{S}$ then U is symmetric positive definite and we have

$$\begin{aligned}
\lim_{\lambda \searrow 0} \operatorname{tr}\left(X_{\mathcal{S}}(X_{\mathcal{S}}^T X_{\mathcal{S}} + \lambda U)^{-1} X_{\mathcal{S}}^T\right) &= \operatorname{tr}\left(X_{\mathcal{S}} \lim_{\lambda \searrow 0}\left[(X_{\mathcal{S}}^T X_{\mathcal{S}} + \lambda U)^{-1} X_{\mathcal{S}}^T\right]\right) \\
&= \operatorname{tr}\left((X_{\mathcal{S}} V^{-1})(X_{\mathcal{S}} V^{-1})^{+}\right) \\
&= \operatorname{rank}\left(X_{\mathcal{S}} V^{-1}\right) \\
&= \operatorname{rank}\left(X_{\mathcal{S}}\right), \qquad (5.5.6)
\end{aligned}$$

because V^{-1} is full rank. □

Corollary 2. *The unbiased estimates of the degrees of freedom in the extreme case of* $\lambda \searrow 0$*, are*

- *Standardized group LASSO:* $\operatorname{rank}\left(\widetilde{X}_{\mathcal{S}}\right) = \operatorname{rank}\left(X_{\mathcal{S}}\right)$,
- *Overlapped group LASSO:* $\operatorname{rank}\left(X_{\mathcal{S}}^{*}\right)$,
- *Hierarchical group LASSO:* $\operatorname{rank}\left(\widetilde{X}_{\mathcal{S}} P_{\mathcal{S}}\right)$.

In fact, when the tuning parameter λ converges to zero, the degrees of freedom in the mentioned procedures converge to the degrees of freedom of the related least squares estimate.

Chapter 6

Some Additional Aspects of Group LASSO in Fixed Effects Factorial Designs

6.1 Introduction

In this chapter, we show in Theorem 9 that in the case of balanced designs, sum-to-zero constraints are satisfied in each group of estimates of group LASSO. Hence, for small values of λ, estimates of group LASSO match with constrained LSE. We provide a selection method for λ based on AIC and BIC in Section 6.3. Group LASSO and constrained LSE are compared via a toy example in Section 6.4. We study the selection process of a standardized group LASSO and show how replication affects the selection process in the standardized group LASSO.

In the classical ANOVA, parameters of a model are estimated by minimizing quadratic loss. However, by considering all levels of a factor in the model, the LSE is not identifiable. There are two frameworks to overcome this problem. The first one drops out one level from each factor, which is known as baseline constraint. The second one considers all levels of each factor, and adds sum-to-zero constraints for identifiability, which we call a constrained LSE.

The question is which one of the baseline constraints or the sum-to-zero constraints should be used in group LASSO. To answer this question, suppose that X be the full design matrix of a one factor layout with three levels. Therefore, $X_3 = 1 - X_1 - X_2$ and

$$\begin{aligned}\mu + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 &= \mu + \beta_1 X_1 + \beta_2 X_2 + \beta_3(1 - X_1 - X_2) \\ &= \mu + \beta_3 + (\beta_1 - \beta_3)X_1 + (\beta_2 - \beta_3)X_2 \\ &= \mu' + \beta_1' X_1 + \beta_2' X_2,\end{aligned}$$

which leads to baseline constrained design in ANOVA models. Note that the group LASSO of these two designs lead to different Lagrangians:

$$\|Y - [\mu + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3]\|_2^2 + \lambda\|(\beta_1, \beta_2, \beta_3)\|_2$$

and

$$\|Y - [\mu + \beta_3 + (\beta_1 - \beta_3)X_1 + (\beta_2 - \beta_3)X_2]\|_2^2 + \lambda\|(\beta_1 - \beta_3, \beta_2 - \beta_3)\|_2,$$

especially when λ is far from zero. This shows that we should consider full design matrix with sum-to-zero constraints in group LASSO.

Another question raised here is whether the sum-to-zero constraints in group LASSO shall be considered when there is a consideration of full design matrix. Group LASSO with full design matrix and sum-to-zero constraints is a convex optimization problem; however, the number of sum-to-zero constraints increases dramatically in high-dimensional data sets as is shown in Section 2.2. Lim [17] showed that overlapped group LASSO with sum-to-zero constraints is equivalent to group LASSO. Nonetheless, we are thinking more simply and we show in Theorem 9 that sum-to-zero constraints are satisfied in group LASSO in the case of balanced designs. Note that sum-to-zero constraints are used in constrained LSE for identifiability and could be replaced with other identifiability constraints or penalties

such as the group LASSO penalty.

6.2 A connection between constrained LSE and group LASSO

Is there any connection between the estimate of a group LASSO and constrained LSE? Suppose that X be the design matrix with groups $g \in \mathcal{G}$ as factors. We consider a group LASSO with full design X, where all levels of a factor are in the model without sum-to-zero constraints. Therefore, the group LASSO problem is

$$\underset{(\mu.\beta)}{\operatorname{argmin}} \|Y - \mu 1 - X\beta\|_2^2 + \lambda \sum_{g\in\mathcal{G}} \|\beta_g\|_2 .$$

Yuan and Lin [18] showed that if $X^T X = I$ then

$$\hat{\beta}_g^{GL} = \left(1 - \frac{\lambda\sqrt{l_g}}{\|\hat{\beta}_g^{LSE}\|_2}\right)_+ \hat{\beta}_g^{LSE}, \tag{6.2.1}$$

where $\hat{\beta}^{LSE}$ is the least squares regression estimate. It means that when $\lambda \searrow 0$ then $\hat{\beta}^{GL} \to \hat{\beta}^{LSE}$. This equality is correct when X is orthonormal but this assumption is rarely satisfied in actual situations. We will investigate this fact in detail.

By definition, full design matrix refers to a design in which all levels of a factor are considered and none of them is dropped out. Now consider a two-factor layout with full design matrix. The least squares estimate, $\hat{\beta}^{LSE}$, exists when sum-to-zero constraints are applied. In fact, sum-to-zero constraints are used for identifiability. However, the group LASSO penalty preserves the sum-to-zero property under a specific condition. Therefore, instead of using sum-to-zero constraints, we can use the group LASSO penalty for identifiability. Also in the case of high-dimensional data, even by using sum-to-zero constraints, there is no constrained LSE. After all, simply by using the group LASSO penalty, we will get a

sparse solution path. In the following theorem we show that sum-to-zero constraints are satisfied for a group LASSO estimate in each group when the design is balanced and it is independent of the value of tuning parameter λ.

Theorem 9. *Suppose X to be a full balanced design matrix in such a way that each group $X_g, \quad g \in \mathcal{G}$ is an indicator matrix, i.e. each row consists of exactly a single 1. Thus, the solution $\hat{\beta}$ of*

$$\underset{\mu,\beta}{\operatorname{argmin}} \|Y - \mu.1 - X\beta\|_2^2 + \lambda \sum_g \|\beta_g\|_2 ,$$

satisfies

$$\bar{\hat{\beta}}_g = 0.$$

Proof.

Note that $X_g.c1_g = c1_n$ for any constant c since X_g contains a single 1 in each row. Suppose that there are 2 groups, hence it follows that if $\hat{\mu}$ and $\hat{\beta}$ are solutions, then

$$\begin{aligned}\left\|Y - \hat{\mu}.1 - X\hat{\beta}\right\|_2^2 &= \left\|Y - \hat{\mu}.1 - c_1 1_n - c_2 1_n - X\hat{\beta} + c_1 1_n + c_2 1_n\right\|_2^2 \\ &= \left\|Y - (\hat{\mu}.1 + c_1 1_n + c_2 1_n) - X_{g_1}(\hat{\beta}_{g_1} - c_1 1_{g_1}) - X_{g_2}(\hat{\beta}_{g_2} - c_2 1_{g_2})\right\|_2^2\end{aligned}$$

Therefore, $\hat{\mu} + (c_1 + c_2)1_n$ and $\begin{bmatrix} \hat{\beta}_{g_1} - c_1 1_{g_1} \\ \hat{\beta}_{g_2} - c_2 1_{g_2} \end{bmatrix}$ minimize the loss function, while the penalty

$$\begin{aligned}\left\|\hat{\beta}_{g_1} - c1_{g_1}\right\|_2^2 + \left\|\hat{\beta}_{g_2} - c1_{g_2}\right\|_2^2 &= \left(\hat{\beta}_{g_1}^T\hat{\beta}_{g_1} - 2c_1 1_{g_1}^T\hat{\beta}_{g_1} + c_1^2 1_{g_1}^T 1_{g_1}\right) + \left(\hat{\beta}_{g_2}^T\hat{\beta}_{g_2} - 2c_2 1_{g_2}^T\hat{\beta}_{g_2} + c_2^2 1_{g_2}^T 1_{g_2}\right) \\ &= \left(\hat{\beta}_{g_1}^T\hat{\beta}_{g_1} - 2c_1\sum_{i=1}^{l_{g_1}}\hat{\beta}_i + nc_1^2\right) + \left(\hat{\beta}_{g_2}^T\hat{\beta}_{g_2} - 2c_2\sum_{i=l_{g_1}+1}^{l_{g_2}}\hat{\beta}_i + nc_2^2\right)\end{aligned}$$

is minimized for $c_1 = \bar{\hat{\beta}}_{g_1}$ and $c_2 = \bar{\hat{\beta}}_{g_2}$. The intercept μ is not penalized and its estimate $\hat{\mu}$ minimizes only the loss function. Let us find $\hat{\mu}$:

$$\frac{\partial\|Y - \mu.1_n - X\beta\|_2^2}{\partial\mu} = -2.1_n^T(Y - \mu.1_n - X\beta) = 0,$$

then

$$\begin{aligned}\hat{\mu} &= \overline{Y} - \frac{1}{n} 1_n^T . \begin{bmatrix} X_{g_1} & X_{g_2} \end{bmatrix} \begin{bmatrix} \hat{\beta}_{g_1} \\ \hat{\beta}_{g_2} \end{bmatrix} \\ &= \overline{Y} - \frac{r \sum_{i=1}^{l_{g_1}} \hat{\beta}_i}{r l_{g_1} l_{g_2}} - \frac{r \sum_{i=l_{g_1}+1}^{l_{g_2}} \hat{\beta}_i}{r l_{g_1} l_{g_2}} \\ &= \overline{Y} - \frac{c_1}{l_{g_2}} - \frac{c_2}{l_{g_1}},\end{aligned}$$

where r is the number of repetition in each treatment of full balanced design X. On the other side,

$$\frac{\partial \left\| Y - (\mu.1_n + c_1 1_n + c_2 1_n) - X_{g_1}(\beta_{g_1} - c_1 1_{g_1}) - X_{g_2}(\beta_{g_2} - c_2 1_{g_2}) \right\|_2^2}{\partial \mu} =$$

$$-2.1_n^T \left(Y - (\mu.1_n + c_1 1_n + c_2 1_n) - X_{g_1}(\beta_{g_1} - c_1 1_{g_1}) - X_{g_2}(\beta_{g_2} - c_2 1_{g_2}) \right).$$

Therefore,

$$\begin{aligned}\hat{\mu} &= \overline{Y} - c_1 - c_2 - \frac{1}{n} 1_n^T X_{g_1}(\beta_{g_1} - c_1 1_{g_1}) - \frac{1}{n} 1_n^T X_{g_2}(\beta_{g_2} - c_2 1_{g_2}) \\ &= \overline{Y} - c_1 - c_2 - \frac{1}{n} r 1_{g_1}^T (\beta_{g_1} - c_1 1_{g_1}) - \frac{1}{n} r 1_{g_2}^T (\beta_{g_2} - c_2 1_{g_2}) \\ &= \overline{Y} - c_1 - c_2 - \frac{r}{n}(l_{g_1} c_1 - l_{g_1} c_1) - \frac{r}{n}(l_{g_2} c_2 - l_{g_2} c_2) \\ &= \overline{Y} - c_1 - c_2.\end{aligned}$$

Hence,

$$\overline{Y} - c_1 - c_2 = \overline{Y} - \frac{c_1}{l_{g_2}} - \frac{c_2}{l_{g_1}} \quad \Rightarrow \quad c_1 + c_2 = \frac{c_1}{l_{g_2}} + \frac{c_2}{l_{g_1}} \quad \Rightarrow \quad c_1 = c_2 = 0.$$

Note that $l_{g_1}, l_{g_2} \geq 2$. □

The theorem states that the summation of coefficients of each group is zero for group LASSO estimates when the design is balanced. In other words, the sum-to-zero constraint

is satisfied for each group in a group LASSO with balanced design, and this is independent of the value of tuning parameter λ. There are $l_g + l_h$ sum-to-zero constraints for each interaction of two main effects with l_g and l_h levels respectively. While based on the theorem, there is only one sum-to-zero constraint for each interaction. In fact, group LASSO with only main effects satisfies all sum-to-zero constraints which are needed in constrained LSE. This leads to the conjecture that the estimate of a group LASSO converges to the estimate of a constrained LSE when the tuning parameter λ converges to zero. However, it is not correct when interactions are in the model. The conjecture will be illustrated in Table 6.2 in Section 6.4.

The group LASSO estimate depends on the value of the tuning parameter λ. There are two extreme cases for λ in a group LASSO when it is large or small.

- $\lambda \longrightarrow \infty \quad \Rightarrow \quad \hat{\beta}^{GL} \longrightarrow 0$ and $SSE^{GL} \longrightarrow \left\| Y - \overline{Y} \right\|_2^2$,
- $\lambda \longrightarrow 0 \quad \Rightarrow \quad SSE^{GL} \longrightarrow SSE^{LSE}$,

where SSE stands for sum of squared errors. In fact, when λ is fixed with a large value, the penalty gets a huge weight and all efforts are focused on minimization of penalty which leads to $\hat{\beta}^{GL} = 0$ as the optimal estimation. When λ is fixed with a small value, the loss function gets the main weight in optimization and group LASSO penalty works instead of sum-to-zero constraints for identifiability; therefore, it leads to an SSE equal to constrained LSE. This is reflected in Tables 6.2 and 6.4.

6.3 The selection of the regularization parameter λ

The tuning parameter λ in regularization methods is selected in such a way that the corresponding model is optimal according to some criteria, such as AIC, BIC, or Mallow's

C_p. The spirit of these criteria is based on the prediction risk, Risk(f). Suppose f to be a continuous prediction rule. Thus according to [6],

$$\begin{aligned}\text{Risk}(f) = E\left[\|\mu - f(Y)\|_2^2\right] &= E\left[\|Y - f(Y)\|_2^2\right] - n\sigma^2 + 2\sum_{i=1}^{n}\text{Cov}\left(Y_i, f(Y_i)\right) \\ &= E\left[\|Y - f(Y)\|_2^2\right] - n\sigma^2 + 2\sigma^2\text{df}(f). \qquad (6.3.1)\end{aligned}$$

The Stein's Unbiased Risk Estimate is

$$\widehat{\text{Risk}}(f) = \|Y - f(Y)\|_2^2 - n\sigma^2 + 2\sigma^2\widehat{\text{df}}(f),$$

where the unbiased estimate of df(f) is

$$\widehat{\text{df}}(f) = \sum_{i=1}^{n}\frac{\partial f_i(Y)}{\partial Y_i}.$$

It means that an unbiased estimate of df(f) suffices to provide an unbiased estimate for Risk(f), where σ^2 is unknown and usually replaced with an estimate based on the largest model [26]. The largest model corresponds to the smallest λ and, when $\lambda \longrightarrow 0$, then

$$\hat{\sigma}^2 = MSE^{GL} \longrightarrow MSE^{LSE},$$

where $MSE^{GL} = \frac{SSE^{GL}}{n-\widehat{\text{df}}}$, and MSE refers to mean squared error. It means that minimization of prediction risk is involved with MSE^{LSE}.

Yuan and Lin [18] used Mallow's C_p for the selection of λ, which is

$$C_p\left(\hat{\mu}\right) = \frac{\|Y - \hat{\mu}\|_2^2}{n} + \frac{2}{n}\widehat{\text{df}}_\lambda(\hat{\mu})\sigma^2.$$

They considered $\hat{\beta}^{GL}$ from (6.2.1) and derived $\widehat{\text{df}}_\lambda(\hat{\mu})$ by

$$\text{tr}\left(\frac{\partial\left(X\hat{\beta}\right)}{\partial Y}\right) = \text{tr}\left(\frac{\partial\left(X\hat{\beta}\right)}{\partial\hat{\beta}^{LS}}\frac{\partial\hat{\beta}^{LS}}{\partial Y}\right)$$

$$= \sum_g I\left(\|\hat{\beta}_g\|_2 > 0\right) + \sum_g \frac{\|\hat{\beta}_g\|_2}{\|\hat{\beta}_g^{LS}\|_2}\left(l_g - 1\right). \tag{6.3.2}$$

However, this is correct under the assumption that $X^T X = I$, which is rarely satisfied in practical situations. They considered (6.3.2) as an approximation for the unbiased estimate of $\mathrm{df}(f)$.

Breheny and Huang [14] proposed an intuitive formulation for degrees of freedom in group LASSO by

$$\widehat{\mathrm{df}}_\lambda(\hat{\mu}) = \sum_{g\in\mathcal{G}} \sum_{k=1}^{l_g} \frac{\hat{\beta}_{gk}}{\hat{\beta}^*_{gk}},$$

where $\hat{\beta}^*_{gk}$ is the unpenalized fit for partial residuals, $\hat{\beta}^*_{gk} = X^T_{gk}\widetilde{Y}/n$ and $\widetilde{Y}$ is the current update of residuals in the proposed algorithm in [14]. This formula is justified intuitively.

We propose Stein's unbiased estimate for degrees of freedom of a group LASSO, which is derived by Vaiter et al. [4]. A simplified version of the estimate of the degrees of freedom is

$$\widehat{df}_\lambda(\hat{\mu}) = \mathrm{tr}\left(X_{\mathcal{S}}\left(X_{\mathcal{S}}^T X_{\mathcal{S}} + \lambda U\right)^{-1} X_{\mathcal{S}}^T\right), \tag{6.3.3}$$

where $U = \mathrm{diag}\left[\frac{I - \frac{\hat{\beta}_g \hat{\beta}_g^T}{\|\hat{\beta}_g\|_2^2}}{\|\hat{\beta}_g\|_2}\right]_{g\in\mathcal{S}}$ and diag is a blockwise diagonal operator. This formulation of degrees of freedom is justified in Figure 6.1 where the estimated degrees of freedom of group LASSO from (6.3.3) are compared with the actual degrees of freedom from

$$df(f) = \frac{1}{\sigma^2}\sum_{i=1}^{n} \mathrm{Cov}\left(Y_i, f(Y_i)\right). \tag{6.3.4}$$

Figure 6.1 shows the Monte Carlo estimate of $E\left[\hat{df}_\lambda\right]$ on the y-axis versus the Monte Carlo estimate of the actual df_λ, given by 6.3.4, for a sequence of λ (circular) with one standard error bar. In this simulation, $X \in \mathbb{R}^{100\times 75}$ and $\beta \in \mathbb{R}^{75}$ are fixed for 25 groups, each with 3

levels and, therefore, $B = 1000$ Monte Carlo replicates of $y^{(b)}$ are generated. The Monte Carlo estimate of $E\left[\hat{df}_\lambda\right]$ and the actual df_λ are computed from the generated values of y. For group LASSO we used *gglasso* package in R.

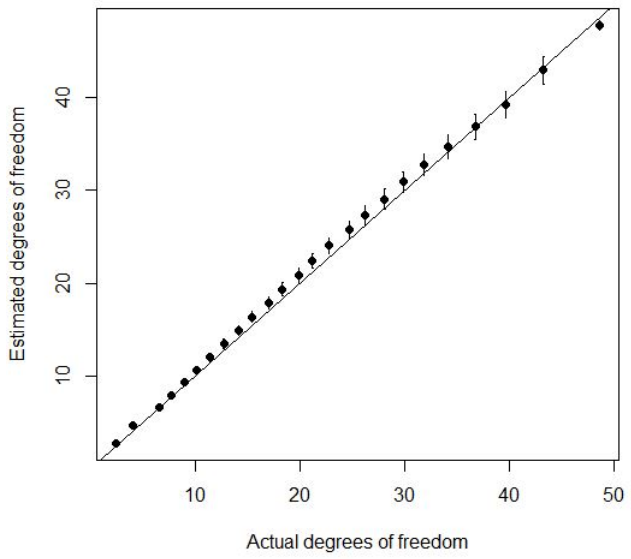


Figure 6.1: Comparing the unbiased estimate of degrees of freedom with the actual degrees of freedom for group LASSO.

We will assess the effect of proposed degrees of freedom on the selection of λ with widely used model selection criteria

$$AIC\left(\hat{\mu}\right) = \frac{\|Y - \hat{\mu}\|_2^2}{n\sigma^2} + \frac{2}{n}\widehat{\mathrm{df}}_\lambda(\hat{\mu})$$

and

$$BIC\left(\hat{\mu}\right) = \frac{\|Y - \hat{\mu}\|_2^2}{n\sigma^2} + \frac{\log(n)}{n}\widehat{\mathrm{df}}_\lambda(\hat{\mu}).$$

AIC and BIC have different properties in regression. BIC is consistent in model selection if the true model is among candidates [23]. It means that if the true model is in the candidates list, then the probability that BIC selects the true model converges to 1 as $n \to \infty$. AIC tends

to overestimate the true model and asymptotically selects the smallest average squared error when the true model is not among candidates. Zou, Hastie and Tibshirani [26] demonstrated these facts for a LASSO by a simulation study. Their simulation shows that BIC has a much higher probability than AIC to identify the true model in a LASSO. Also it shows that AIC is conservative in variable selection and it tends to overfit in LASSO, but BIC tends to find models with the right size. The following simulation experiment is analogous to that of Zou, Hastie and Tibshirani [26]: it illustrates analogous facts for AIC and BIC, this time for a group LASSO.

Simulation 1.

Consider a linear model with three factors. We will generate eight factors each with three levels, but the response depends on only three factors. Consider $Z_1, \cdots, Z_8$ *be random variables from a multivariate normal distribution with covariance between* Z_i *and* Z_j *being* $0.1^{|i-j|}$ *and a mean of zero. The three levels of each factor are* $0, 1,$ *and* 2 *if smaller than* $\Phi^{-1}(\frac{1}{3})$*, between* $\Phi^{-1}(\frac{1}{3})$ *and* $\Phi^{-1}(\frac{2}{3})$*, and greater than* $\Phi^{-1}(\frac{2}{3})$*, respectively. The response* Y *is generated by* $Y = X\beta + N(0,1)$ *where*

$$\beta^T = (7, -6.7, 8.2, -1.5, 4, -2.5, -1.7, 0, 0, 0, 0, 0, 0, -2.7, 1.8, 1.2, 0, 0, 0, 0, 0, 0, 0, 0, 0).$$

The number of observations in each run is $n \in \{100, 500, 1000, 2000\}$ *and the number of replication is 2000. The results are shown in Table 6.1.*

n	median		proportion	
	AIC	BIC	AIC	BIC
100	5	3	0.158	0.526
500	5	3	0.168	0.718
1000	5	3	0.183	0.764
2000	5	3	0.180	0.797

Table 6.1: Comparing the median number of selected nonzero factors and the probability of discovering the exact true model by AIC and BIC in group LASSO.

The exact true model is the model that has no-zero coefficients for factors of $\{1, 2, 5\}$ *and*

has zero coefficients for other factors. In this simulation both AIC and BIC select the true factors $\{1, 2, 5\}$ *in all 2000 replications, but BIC tends to select more sparse models than AIC. AIC is more conservative than BIC in shrinking of factors while BIC has higher probability than AIC in identifying the true model. This fact is shown in Figure 6.2.*

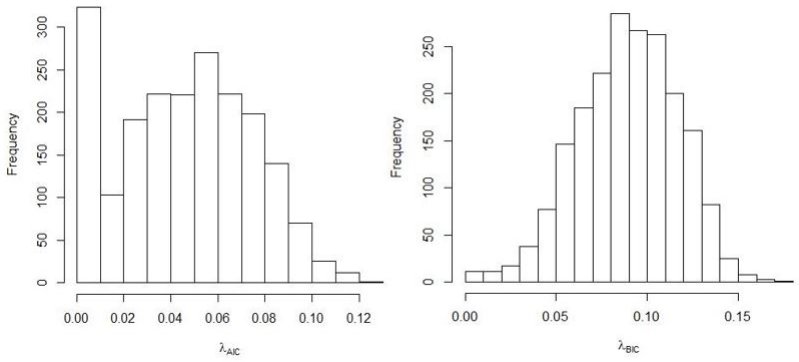


Figure 6.2: The left histogram shows the distribution of λ_{AIC} and the right one is for λ_{BIC} for the simulation in the first row of Table 6.1.

The results of Simulation 1 is comparable to the results of LASSO in Zou, Hastie and Tibshirani [26]. We will use BIC criterion with proposed degrees of freedom for the selection of λ, but AIC could be used for a conservative purpose.

6.4 The toy example revisited

In the toy example, given in Section 2.2, we compare the results of this thesis with the constrained LSE. These two should be compared on low dimensional factorial designs since classical ANOVA is not applicable on high-dimensional factorial designs. Also, this comparison is needed to find out the missed points when we apply regularization methods. First, we consider a model with two main effects in order to compare the constrained LSE with different estimates in a group LASSO. We consider the pair interaction in the model and repeat the above comparison; finally, we will consider a hierarchal model.

6.4.1 Two-factor layout without interaction

Consider the linear model for the two-factor layout without interaction:

$$y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}, \tag{6.4.1}$$

where $i = 1, 2, \quad j = 1, 2, 3, \quad k = 1, \cdots, 10, \quad \sum_i \alpha_i = 0, \quad \sum_j \tau_j = 0, \quad$ and $\quad \varepsilon_{ijk} \sim N(0, \sigma^2)$. Now, if we consider $\sum_i \alpha_i = 0$ and $\sum_j \tau_j = 0$, then $\hat{\mu} = \bar{y}_{\cdots}$. Sum-to-zero constraints are necessary for identifiability of estimates. The model (6.4.1) can be represented in the form of regression model by using dummy variables X; therefore,

$$Y = \mu + X\beta + \varepsilon,$$

where $X = [X_1; X_2; X_3; X_4; X_5]$, $\beta = [\alpha_1, \alpha_2, \tau_1, \tau_2, \tau_3]$, Y is the dependent variable, $\sum_i \alpha_i = 0$, $\sum_j \tau_j = 0$, and $\varepsilon \sim N(0, \sigma^2 I)$. The constrained LSE of (μ, β) is given by minimizing

$$(\hat{\mu}, \hat{\beta}) = \underset{\mu,\beta}{\operatorname{argmin}} \frac{1}{2} \|Y - \mu.1 - X\beta\|_2^2,$$

where sum-to-zero constraints are on parameters themselves. Here, we compare the constrained LSE with the group LASSO and the standardized group LASSO. Cross-validation tends to select small tuning parameter λ. This means that cross-validated results of group LASSO and standardized group LASSO should be close to the constrained LSE. In Table 6.2, cross- validated results of both group LASSO and standardized group LASSO coincide with the result of constrained LSE. In fact when λ converges to zero, group LASSO estimate tend to constrained LSE.

In Table 6.2, the cross-validated group LASSO, standardized group LASSO, and constrained LSE have the same SSE. Their coefficients are computationally equal. Also, sum-to-zero constraints are satisfied for all three models as a result of Theorem 9. The degrees

		β	SSE	$d.f_{model}$	$\sum_{i \epsilon g} \beta_i$	$\sum_g \|\beta_g\|_2$
Constrained LSE		(3.7, -3.7, -6.683, 8.216, -1.533)	2640.3	3	(0,0)	15.935
group LASSO	C-V	(3.7, -3.7,-6.677,8.208,-1.532)	2640.3	3.998	(0,0)	15.935
	AIC	(3.595,-3.595,-6.512,8.006,-1.494)	2642.5	3.961	(0,0)	15.511
	BIC	(3.487,-3.487,-6.337,7.791,-1.454)	2649.2	3.92	(0,0)	15.079
standardized group LASSO	C-V	(3.699, -3.699, -6.683,8.216, -1.533)	2640.3	3.997	(0,0)	15.932
	AIC	(3.262,-3.262,-6.103,7.503,-1.4)	2669.1	2.904	(0,0)	14.385
	BIC	(3.062,-3.062,-5.838,7.177,-1.339)	2701.4	2.701	(0,0)	13.678

Table 6.2: Comparison of the estimates of constrained LSE, group LASSO and standardized group LASSO for the two-factor layout without interaction.

of freedom for constrained LSE are around 1 unit smaller than the two cross-validated models. However, we expect to have the same degrees of freedom as constrained LSE by taking small values for tuning parameter λ. We will discuss this difference in more detail in Section 6.5.

The preferred model is determined by comparing AIC and BIC in both procedures. In Tabel 6.3, standardized group LASSO shows better values for AIC and BIC criteria compared to group LASSO, which was an expected result. We suggest to use the estimate of the standardized group LASSO with AIC criterion in this case.

	AIC	BIC
group LASSO	1.0661	1.2040
Standardized group LASSO	1.0404	1.1393

Table 6.3: Comparison of AIC and BIC criteria of group LASSO and standardized group LASSO for the two-factor layout without interaction.

Figure 6.3 illustrates the behaviour of AIC and BIC in standardized group LASSO.

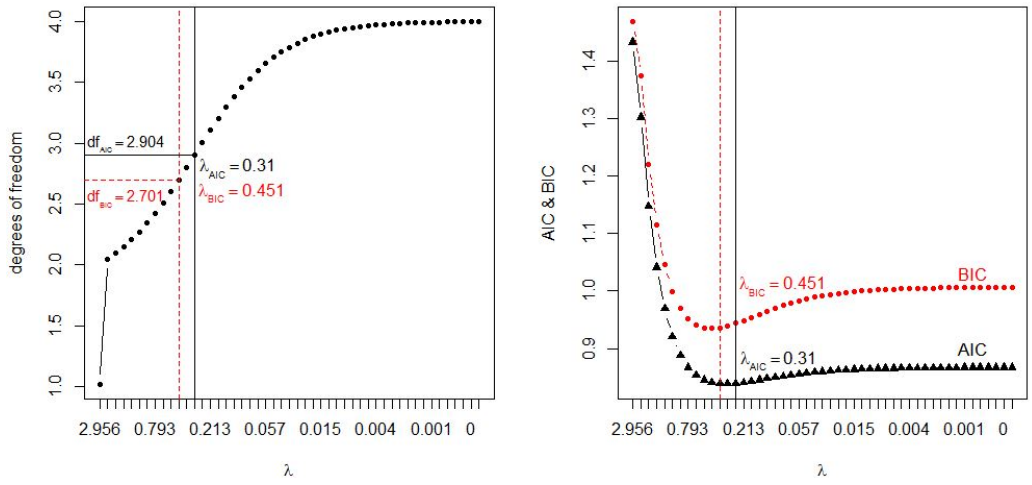


Figure 6.3: The left panel illustrates the degrees of freedom and the right one illustrates the smooth behaviour of AIC and BIC in the standardized group LASSO for the two-factor layout without interaction.

6.4.2 Two-factor layout with interaction

Consider the linear model for the two-factor layout with interaction:

$$y_{ijk} = \mu + \alpha_i + \tau_j + \omega_{ij} + \varepsilon_{ijk}, \tag{6.4.2}$$

where $i = 1, 2, \quad j = 1, 2, 3, \quad k = 1, \cdots, 10, \quad \sum_i \alpha_i = 0, \quad \sum_j \tau_j = 0, \quad \sum_j \omega_{ij} = 0, \quad \sum_i \omega_{ij} = 0,$ and $\varepsilon_{ijk} \sim N(0, \sigma^2)$. The results of constrained LSE, group LASSO, and standardized group LASSO are compared in Table 6.4.

In the two-factor layout the first two main effects require two sum-to-zero constraints, but the interaction requires five sum-to-zero interaction. Note that according to Theorem 9, there would be only one sum-to-zero constraint for interaction. Therefore, the coefficients in constrained LSE, group LASSO, and standardized group LASSO would be different. This fact is reflected in Table 6.4. We mentioned that when the tuning parameter λ converges to zero, group LASSO and constrained LSE would have the same SSE. In fact, con-

		β	SSE	$d.f_{model}$	$\sum_{i\epsilon g}\beta_i$	$\sum_g \|\beta_g\|_g$
Constrained LSE		(3.7,-3.7,-6.683,8.217,-1.533, -3.45,-1.05,4.5,3.45,1.05,-4.05)	1975.2	5	(0,0,0,0,0,0,0)	24.09
group LASSO	C-V	(2.333,-2.333,-4.465,5.489,-1.024, -4.288,3.035, 5.342,-0.135,2.403,-6.357)	1975.2	5.997	(0,0,0)	20.566
	AIC	(2.314,-2.314,-4.435,5.453,-1.018, -4.098,2.945,5.103,-0.175,2.308,-6.083)	1977.776	5.941	(0,0,0)	20.063
	BIC	(2.313,-2.313,-4.432,5.449,-1.017, -3.813,2.745,4.748,-0.166,2.148,-5.661)	1989.186	5.861	(0,0,0)	19.386
standardized group LASSO	C-V	(0.005, -0.005, -1.506, 1.851, -0.345, -4.932, 9.009, 7.006, -5.422, 3.720, -9.382)	1975.2	5.992	(0,0,0)	19.325
	AIC	(0,0,-1.329, 1.634, -0.3049, -4.647, 8.406, 6.347, -5.102, 3.581, -8.584)	1998.9	3.856	(0,0,0)	17.78
	BIC	(0,0,-1.319, 1.622, -0.303, -4.447, 8.040, 6.061, -4.882, 3.431, -8.201)	2025.5	3.629	(0,0,0)	17.08
hierarchical standardized group LASSO	C-V	(3.016, -3.016, -5.609, 6.896, -1.287, -3.839, 0.953, 4.936, 1.692, 1.686,-5.429)	1975.2	5.998	(0,0,0)	21.918
	AIC	(2.856, -2.856, -5.409, 6.650, -1.241, -3.442, 0.933, 4.423, 1.438, 1.530, -4.882)	1989.3	5.484	(0,0,0)	20.481
	BIC	(2.829, -2.829, -5.376, 6.610, -1.233, -3.356, 0.912, 4.313, 1.400, 1.493, -4.762)	1995.7	5.435	(0,0,0)	20.197

Table 6.4: Comparison of the estimates of constrained LSE, group LASSO and standardized group LASSO for the two-factor layout with interaction.

strained LSE and all cross-validated models have the same SSE, in Table 6.4. The degrees of freedom for constrained LSE are again 1 unit less than for all cross-validated models. We will explain this difference in detail in Section 6.5.

Standardized group LASSO drops the first main effect with AIC and BIC criteria. However, we know that main effects and interaction are statistically significant at the 0.05 confidence level by classical ANOVA. This leads to hierarchical standardized group LASSO. That means if an interaction is in the model, then all related main effects are in the model. Let us denote the main effects with F_1 and F_2 and the pair interaction with $F_{1:2}$. We consider overlapped group LASSO with upward grouping for hierarchy and, therefore, the grand set is $\mathcal{G} = \{F_1, F_2, \{F_1, F_2, F_{1:2}\}\}$. The result is provided in Table 6.4, which shows that the

first main effect enters the model with all criteria.

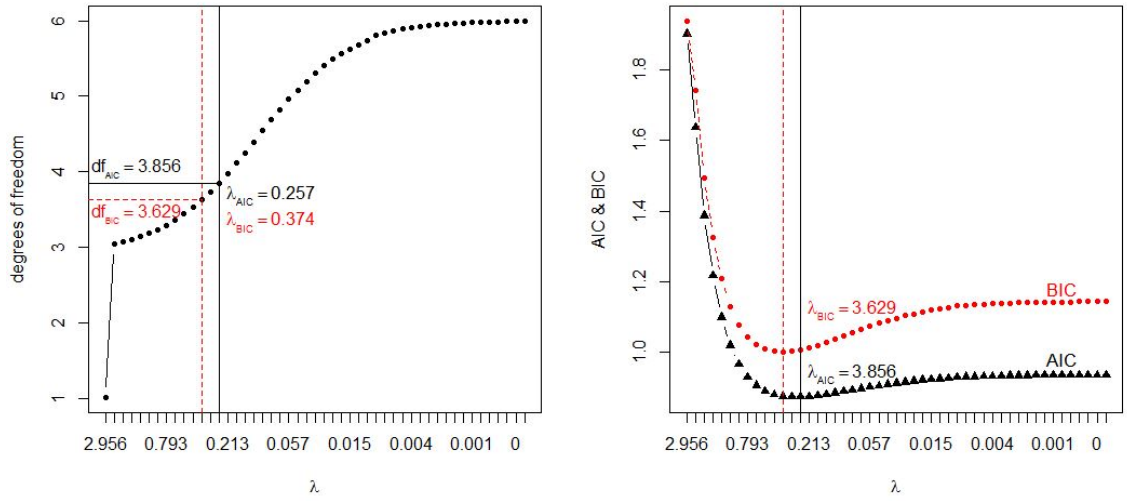


Figure 6.4: The left panel illustrates the degrees of freedom and the right one illustrates the smooth behaviour of AIC and BIC in the standardized group LASSO for the two-factor layout with interaction.

The preferred procedure is determined by comparing AIC and BIC of all models. The hierarchical standardized group LASSO is a compromise between the standardized group LASSO and the group LASSO based on AIC and BIC in Table 6.5 which also satisfies strong hierarchy. Therefore, we suggest the hierarchical standardized group LASSO with AIC.

	AIC	BIC
group LASSO	0.932	1.139
Standardized group LASSO	0.871	0.999
Hierarchical Standardized group LASSO	0.921	1.112

Table 6.5: Comparison of AIC and BIC criteria of group LASSO and standardized group LASSO for the two-factor layout with interaction.

6.5 Discussion

6.5.1 Selection process

The selection process in standardized group LASSO is given in Theorem 5 as

$$\frac{1}{\|X_g\|_2}\left\|X_g^T r_{\mathcal{S}}\right\|_2 \leq \lambda\sqrt{l_g}.$$

Also, in Section 5.5 we showed that

$$\frac{1}{\|X_g\|_2^2}\left\|X_g^T\varepsilon\right\|_2^2 \sim \sigma^2\sum_{i=1}^{l_g}\frac{\delta_{g,i}}{\max_i(\delta_{g,i})}\chi_1^2,$$

where each $\delta_{g,i}$ is an eigenvalue of $X_g^T X_g$. We know that X_g is an indicator matrix; thus, $X_g^T X_g = \text{diag}\left((n_{g,i})_{i\in g}\right)$ where $n_{g,i}$ is the replication in level $i \in g$. In fact, $n_{g,i},\ i \in g$ are eigenvalues of $X_g^T X_g$ i.e., $\delta_{g,i} = n_{g,i}$ for all $i \in g$. If the design is balanced, then all replications are equal and we have

$$\frac{1}{\|X_g\|_2}\left\|X_g^T\varepsilon\right\|_2^2 \sim \sigma^2\sum_{i=1}^{l_g}\chi_1^2 = \sigma^2\chi_{l_g}^2.$$

Now, if we use $\tilde{\beta}_g$ instead of β_g^*, then

$$\left\|Y - \widehat{Y}_{\mathcal{S}}\right\|_2^2 \sim \sigma^2\chi_{l_g}^2.$$

This is analogous to the UMP test (5.5.1). That means X_g is selected by a UMP test to be in the model if

$$\|Y - \widehat{Y}_{\mathcal{S}}\|_2^2 \geq \sigma^2\chi_{l_g}^2(1-\alpha).$$

However, standardized group LASSO [28] selects X_g to be in the model if

$$\left\|Y - \hat{Y}_{\mathcal{S}}\right\|_2^2 \geq l_g\lambda^2,$$

where $\left\|Y - \widehat{Y}_{\mathcal{S}}\right\|_2^2 \sim \sigma^2 \chi^2_{l_g}$. In this way, the selection process in standardized group LASSO is roughly based on a UMP test. Now, if the approximation $l_g \lambda^2 \approx \sigma^2 \chi^2_{l_g}(1-\alpha)$ holds, then the selection process in standardized group LASSO will be exactly according to a UMP test.

6.5.2 Degrees of freedom

The degrees of freedom of constrained LSE are always one unit smaller than that of the cross-validated group LASSO estimates. Consider the two-factor layout in Section 6.4.1. We will calculate its degrees of freedom in two different frameworks and, afterwards, it will be compared with the degrees of freedom of group LASSO. The first framework in Classical ANOVA considers baseline constraints and drops out one level from each factor. Denote the design of the first factor with $X_1 = [X_{11}; X_{12}]$. The design X_1 is an indicator matrix, i.e. there exists only a single 1 in each row and $X_{11} + X_{12} = 1$. Therefore,

$$\begin{aligned}
\|Y - X\beta\|_2^2 &= \left\| Y - \begin{bmatrix} 1 & X_{11} & X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} \mu + \beta_{12} + \beta_{23} \\ \beta_{11} - \beta_{12} \\ \beta_{21} - \beta_{23} \\ \beta_{22} - \beta_{23} \end{bmatrix} \right\|_2^2 \\
&= \left\| Y - \begin{bmatrix} 1 & X_{11} & X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} \mu^* \\ \beta_{11}^* \\ \beta_{21}^* \\ \beta_{22}^* \end{bmatrix} \right\|_2^2 .
\end{aligned}$$

Hence

$$\text{df} = \text{rank}([X_{11} \;\; X_{21} \;\; X_{22}]) = 3.$$

The second framework applies sum-to-zero constraints on parameters themselves; thus, we have

$$\beta_{11} + \beta_{12} = 0 \implies \beta_{12} = -\beta_{11},$$
$$\beta_{21} + \beta_{22} + \beta_{23} = 0 \implies \beta_{23} = -\beta_{21} - \beta_{22}.$$

Therefore, the constrained LSE is

$$\|Y - X\beta\|_2^2 = \left\| Y - \begin{bmatrix} 1 & (X_{11} - X_{12}) & (X_{21} - X_{23}) & (X_{22} - X_{23}) \end{bmatrix} \begin{bmatrix} \mu \\ \beta_{11} \\ \beta_{21} \\ \beta_{22} \end{bmatrix} \right\|_2^2.$$

Hence, the degrees of freedom are

$$\text{df} = \text{rank}(\begin{bmatrix} (X_{11} - X_{12}) & (X_{21} - X_{23}) & (X_{22} - X_{23}) \end{bmatrix}) = 3.$$

Consider the two-factor layout with standardized group LASSO penalty. Group LASSO uses a full design matrix and we have

$$\min_{\beta} \|Y - \sum_{g=1}^{2} \frac{X_g}{\sqrt{l_g}} \beta_g^*\|_2^2 + \lambda \sum_{g=1}^{2} \|X_g \beta_g^*\|.$$

We know that if $\lambda \searrow 0$, then the degrees of freedom in standardized group LASSO are

$$df = \text{rank}\,(X) = 4.$$

This is consistent with the cross-validated result of standardized group LASSO in Table 6.2. Note that $\text{rank}\,([X_{11}\ X_{12}]) = 2$ and $\text{rank}\,([X_{21}\ X_{22}\ X_{23}]) = 3$; however, $\text{rank}\,([X_1\ X_2]) = 4$. Since $X_{11} + X_{12} = 1$, then in matrix X we will have $X_{23} = X_{11} + X_{12} - X_{21} - X_{22}$.

Chapter 7

Conclusion

In many situations, additive models are insufficient for predicting an outcome in factorial experiments, and pair interactions are useful for this purpose. However, considering models with pair interactions in the case of high-dimensional data sets, adds a huge number of parameters to the model. In this thesis, we studied methods fitting interaction that are applicable to the high-dimensional data sets. They produce groupwise sparse fits, which obey strong hierarchy rules. Here, we address some questions about the significance of the studied methods and their applications.

7.1 What is the significance of the new methods?

Lim [17] and Yan et al. [1] used the overlapped group LASSO with an upward grouping penalty to produce a hierarchical fit in the case of factorial experiments. In Section 2.7, we pointed out three disadvantages for this method. The hierarchical LASSO [27] solves these issues, but it may not be applicable to linear models having factors with more than two levels. The hierarchical group LASSO fixes this particular problem; it is useful to study factorial experiments having factors with more than two levels.

The overlapped group LASSO selects the tuning parameter λ by cross-validation. However, the degrees of freedom are derived for the hierarchical group LASSO, which allow us to select the tuning parameter λ based on the criteria such as Mallows's C_p, AIC and BIC.

The overlapped group LASSO achieves hierarchy based on a geometric interpretation of the l_1-norm. At the same time, the hierarchical group LASSO enjoys a statistical principle for achieving hierarchical fits.

The hierarchical overlapped group LASSO gets more aggressive in eliminating interactions by increasing the number of factors p and, in fact, it is more likely to lead to additive models for large values of p. This fact is explained in Proposition 1 and Yan et al. [1]. However, the hierarchical group LASSO selects interactions based on their *statistical power* [27] and the selection process does not relate to the value of p.

In this thesis, we investigate further the standardized group LASSO and the overlapped group LASSO. We provide unit group weights in the standardized group LASSO by using Zhao's transformation [36]. Also, we derive the degrees of freedom for this method, which enables us to select the tuning parameter λ properly. Finally, the degrees of freedom of the overlapped group LASSO are derived.

7.2 What data can be analyzed?

The methods studied in this thesis are applicable to the factorial experiments for both designed experiments and observational studies. They are proposed for the particular case of high-dimensional data sets; however, they are also applicable on data sets with $p < n$ as well. The predictors in the data sets could be continuous, discrete, and also factors. There is no specific limitation for the number of levels of factors in the data sets. Finally, these methods could deal with continuous or binary responses.

Yang and Zou [38] provide 8 real data sets for testing their group LASSO algorithm with

different number of p and n for 4 continuous and 4 binary responses. The data set *bardet* is the most simple one. The data set is a gene expression data from the microarray experiments of mammalian eye tissue samples. It contains 120 samples with 20 factors each with 5 levels. The response is a continuous random variable giving the log transform of expression level of gene TRIM32, which causes Bardet-Biedl syndrome. It is a genetic disorder with many effects on a body system. Note that the design matrix including pair interactions contains 4850 columns.

7.3 What are the next steps toward applications?

In factorial experiments, the issue of identifiability arises in the first step of computation. There are two frameworks to solve this issue: baseline constraint, and full design matrix with sum-to-zero constraints.

In Chapter 6, we showed that the group LASSO estimate changes by dropping one level from each group or by considering the baseline constraint. In fact, the fit will be changed by changing the base level. Based on this issue, Lim [17] considered full design matrix with sum-to-zero constraints and proposed his theorem for reducing an overlapped group LASSO with sum-to-zero constraints to a group LASSO. We provide a simpler approach for this issue. At first, we showed that sum-to-zero constraints are satisfied in a group LASSO estimate when the design matrix is balanced, and this fact is similar to the case of LSE. Also we noticed that the group LASSO penalty itself can be used for identifiability. Based on these two reasons, we propose a full design matrix without sum-to-zero constraints for studied methods. In order to see the priority of this issue, recall that 2000 sum-to-zero constraints are needed for the bardet data set or in a larger sale, around 2.15×10^9 sum-to-zero constraints for the motivational example, Section 2.2.1.

The other issue in computation is the uniqueness of the solution. This problem is addressed in Chapters 3 and 5 for a hierarchical group LASSO with quadratic and logistic loss. In the case of quadratic loss, we suggested to add a tiny factor of elastic net term to the objective function. In this situation, the loss function with the elastic net term can be rewritten as a new loss function with a new augmented design matrix. For the case of logistic loss, we suggested Roth and Fischer's [8] constraint and algorithm.

For the data set bardet, the response is $Y_{120\times 1}$, the design matrix is $X_{120\times 4850}$, and the group indices are $g = 1, \cdots, 210$. There are 20 main effects each with 5 levels and 190 pair interactions each with 25 levels. To guarantee the uniqueness of the solution, we need to construct the augmented design matrix, which is explained in Section 3.2. The augmented design matrix is $\widetilde{X}_{\varepsilon_{5070\times 9700}}$ and the new response is $Y_{\varepsilon}^{T} = [Y^{T} \quad 0_{1\times 4950}]$. The matrix $\widetilde{X}_{\varepsilon}$ in the bardet example contains 2.87×10^{6} zero components. In such a case, memory consumption can be reduced by using a specialized representation storing, which is called a sparse matrix. A user friendly construction of a sparse matrix in R is *spMatrix*. The Lagrangian for this example is

$$\frac{1}{2}\|Y_{\epsilon} - \widetilde{X}_{\varepsilon}\phi\|_2^2 + \lambda \sum_{g=1}^{210} \|D_g \phi_g\|_2 + \lambda \sum_{g\neq h} \|D_{g:h}\phi_{g:h}\|_2 - \mu^T A\phi + \upsilon^T E\phi, \tag{7.3.1}$$

where ϕ and D are defined in Section 3.4; $A\phi$ and $E\phi$ represent the hierarchy constraints $\frac{1}{2}\|\Theta_g \text{diag}(\frac{1}{w})\|_1 \leq \|\beta_g\|_1$ and the symmetry constraints $\Theta = \Theta^T$, respectively.

7.4 What numerical problems may arise and how these can be addressed?

Here, we give an outline to shed a light on the computational algorithm. Coordinate descent is a derivative-free optimization algorithm. It minimizes a multivariate function along

one direction each time and iteratively minimizes it for each direction. Coordinate descent has issues with non-differentiable functions. Note that group LASSO is non-differentiable, but the non-differentiable part is blockwise separable. Tseng [31] showed that blockwise coordinate descent converges to the global minimum for a strictly convex problem when the non-differentiable part is separable. However, hierarchical group LASSO is not separable because of the symmetry constraint $\Theta = \Theta^T$. In fact, the symmetry constraint ties and couples all of the parameters together, for instance $\Theta_{g:h} = \Theta_{h:g}^T$ appears in two hierarchy constraints of $\|\Theta_{g:h}\|_1 \leq \|\beta_g^+ - \beta_g^-\|_1$ and $\|\Theta_{h:g}\|_1 \leq \|\beta_h^+ - \beta_h^-\|_1$.

In order to solve this issue, we propose to use the Alternating Direction Method of Multipliers (ADMM) [5]. The ADMM splits a convex problem into separate easier subproblems [5]. Consider a convex problem of the form $\min_\phi g(\phi) + h(\phi)$, then it can be rewritten as $\min_{\beta,\gamma} g(\phi) + h(\gamma) \quad \text{s.t.} \quad \beta = \gamma$. Note that, in the Lagrangian 7.3.1:

$$g(\phi) = \frac{1}{2}\|Y_\epsilon - \widetilde{X}_\varepsilon \phi\|_2^2 + \lambda \sum_{g=1}^{210} \|D_g \phi_g\|_2 + \lambda \sum_{g \neq h} \|D_{g:h}\phi_{g:h}\|_2 - \mu^T A\phi$$

and

$$h(\phi) = E\phi$$

Therefore, the ADMM algorithm repeats the following three steps until convergence:

(i) $\hat{\phi} = \text{argmin}_\phi \left[g(\phi) + \left(\frac{\rho}{2}\right) \left\| \phi - \hat{\gamma} + \frac{\hat{u}}{\rho} \right\|_2^2 \right]$.

(ii) $\hat{\gamma} = \text{argmin}_\gamma \left[h(\gamma) + \left(\frac{\rho}{2}\right) \left\| \gamma - \hat{\phi} + \frac{\hat{u}}{\rho} \right\|_2^2 \right]$.

(iii) $\hat{u} \leftarrow \hat{u} + \rho(\hat{\phi} - \hat{\gamma})$.

The dual variable u pulls these two subproblems together and $\rho > 0$ is the penalty parameter [5]. Note that $g(\phi)$ is a blockwise separable strictly convex problem. For minimizing $g(\beta)$, we can use one of the generalized gradient descent solvers such as FISTA [17].

Lim [17] used FISTA to solve a group LASSO. This algorithm with some modification is useful to minimize $g(\phi)$.

A full hierarchical group LASSO analysis for the above mentioned data is quite involved and exceeds the purpose of this thesis. Here, we laid the theoretical foundation and developed new methods. Specific applications of those will be topics for future research.

Bibliography

[1] X. Yan; J. Bien. Hierarchical sparse modeling: A choice of two regularizers. *arXiv:1512.01631*, 2015.

[2] H. Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24:17–36, 1996.

[3] D.R. Cox. Interaction. *International Statistical Review*, 52:1–31, 1984.

[4] S. Vaiter; C. Deledalle; G. Peyre; J. Fadili; C. Dossal. The degrees of freedom of the group lasso for a general design. *arXiv:1212.6478v1*, 2012.

[5] S. Boyd; N. Parikh; E. Chu; B. Peleato; J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–22, 2010.

[6] B. Efron. How biased is the apparent error rate of a prediction rule. *Journal of the American Statistical Association*, 81:461–470, 1986.

[7] A. M. Bruckstein; D. L. Donoho; M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51:34–81, 2009.

[8] V. Roth; B. Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. *25th International conference on machine learning, Helsinki, Finland*, 2008.

[9] T. Hastie; R. Tibshirani; J. Friedman. *The Elements of Statistical Learning.* Springer, 2009.

[10] W.J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.

[11] C.J. Wu; M. Hamada. Experiments: Planning, analysis, and parameter design optimization. 2009.

[12] K. Atkinson; W. Han. *Theoretical numerical analysis*. Springer, 2009.

[13] H. Zou; T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[14] P. Breheny; J. Huang. Penalized methods for bi-level variable selection. *Statistics and Its Interface*, 2:369380, 2009.

[15] P. Breheny; J. Huang. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 2013.

[16] M.S. Lobo; L. Vandenberghe; S. Boyd; H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:193–228, 1999.

[17] M. Lim. The group-lasso: Two novel applications. *PhD dissertation, Stanford University*, 2013.

[18] M. Yuan; Y. Lin. Piecewise linear regularized solution path. *The Annals of Statistics*, 35:1012–1030, 2007.

[19] J.A. Nelder. A reformulation of linear models. *Journal of the Royal Statistical Society, Series A*, 140:48–77, 1977.

[20] J.A. Nelder. The statistics of linear models: back to basics. *Statistics and Computing*, 4:221–234, 1994.

[21] J. Peixoto. Hierarchical variable selection in polynomial regression models. *The American Statistician*, 41:311–313, 1987.

[22] Y. Nardi; A. Rinaldo. On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistic*, 2:605–633, 2008.

[23] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[24] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9:1135–1151, 1981.

[25] R.J. Tibshirani; J. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40:1198–1232, 2012.

[26] H. Zou; T. Hastie; R. Tibshirani. On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):21732192, 2007.

[27] J. Bien; J. Taylor; R. Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41:1111–1141, 2013.

[28] N. Simon; R. Tibshirani. Standardization and the group lasso penalty. *Statistica Sinica*, 22:983–1001, 2012.

[29] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[30] R.J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:14561490, 2013.

[31] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

[32] L. Meier; S. van de Geer; P. Buhlmann. The group lasso for logistic regression. *Journal of Royal Statistics Society, B*, 70:53–71, 2008.

[33] S. Boyd; L. Vandenberghe. *Convex Optimization*. Cambridge, 2004.

[34] G. Obozinski; L. Jacob; J.P. Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv:1110.0413v1*, 2011.

[35] M. Hamada; C.J. Wu. Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24:130–137, 1992.

[36] P. Zhao; G. Rocha; B. Yu;. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37:3468–3497, 2009.

[37] X. Yu. Class notes. *https://www.math.ualberta.ca/~xinweiyu/217.1.13f/217-20131010.pdf.*

[38] Y. Yang; H. Zou. A fast unified algorithm for computing group-lasso penalized learning problems. *Statistics and Computing*, 25(6):11291141, 2015.