MASTERARBEIT | MASTER'S THESIS

Titel | Title

„Hierachical Lasso Models“

verfasst von | submitted by

Răzvan-Andrei Morariu

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 645
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Data Science
Betreut von | Supervisor:	Univ.-Prof. Dr. Tatyana Krivobokova

Acknowledgements

I would like to express my gratitude to my supervisor, Univ.-Prof. Dr. Tatyana Krivobokova, for her constant support and insightful guidance, as well as to Dr. Gianluca Finocchio for the valuable ideas shared during our meetings. Finally, I want to thank my family for their encouragement throughout this journey.

Abstract

In many disciplines the data are often collected in designed experiments. Lasso models are classical tools in most experimental sciences to identify factors and their interactions that are crucial for the variable of interest. For example, one might run a chemical reaction under various conditions in order to identify reaction components that result in the best yield. In modern applications these models have become much more complex: (i) one often is interested in many more than 2 factors, (ii) there is typically only one observation per each factor level combination, (iii) the response variable is not necessarily normal. As a result, none of the classical results are applicable. This thesis aims to address these issues. In presence of many factors and single replications per factor level combination, it is reasonable to assume that most of the factor level combinations are not significant. It is tempting to apply directly a Lasso algorithm to perform estimation and model selection in one step. However, due to the hierarchical structure of the data (there can be no interaction effect, if the main effects are not significant), the classical Lasso algorithm is not applicable. There were several Lasso algorithms developed for hierarchical data under various assumptions. However, these algorithms are limited to two-factor models. Since with the growing number of factors the complexity of the structure increases exponentially, it is a highly non-trivial task to extend such hierarchical Lasso algorithms to multifactorial models. This thesis attempts to develop a hierarchical Lasso algorithm for three factors that is suitable for responses modeled by distributions from the exponential family. The method should be applied to the data on Deoxyfluorination which has a non-normal response.

Kurzfassung

In vielen Fachrichtungen werden Daten oft durch bestimmte Experimente gesammelt. Lasso-Modelle sind klassische Werkzeuge, die in den meisten experimentellen Wissenschaften verwendet werden, um Faktoren und Interaktionen zu identifizieren, die essentiell für die gesuchte Variable sind. Zum Beispiel kann eine chemische Reaktion unter verschiedenen Bedingungen durchgeführt werden, um Reaktionskomponenten zu identifizieren, die zum besten Ertrag führen. Diese Modelle sind in modernen Anwendungen viel komplexer geworden: (i) man interessiert sich oft für viel mehr als 2 Faktoren, (ii) es gibt normalerweise nur eine Beobachtung pro Kombination der Faktorstufen, (iii) die Antwortsvariable ist nicht unbedingt normal. Im Ergebnis sind keine der klassischen Resultate anwendbar. Diese These zielt darauf ab, die oben erwähnten Aspekte anzusprechen. In der Anwesenheit von vielen Faktoren und einzelnen Replikationen pro Faktorstufenkombination ist es vernünftig anzunehmen, dass die Mehrheit der Faktorstufenkombinationen nicht signifikant ist. Es ist verlockend, einen Lasso-Algorithmus direkt anzuwenden, um Schätzungen und Modellauswahl in einem Schritt durchzuführen. Jedoch ist der klassische Lasso-Algorithmus trotz der hierarchischen Struktur der Daten (es kann keinen Interaktionseffekt geben, wenn die Haupteffekte nicht signifikant sind) nicht anwendbar. Es wurden viele Lasso-Algorithmen für hierarchische Daten unter verschiedenen Annahmen entwickelt. Allerdings sind diese Algorithmen auf Zwei-Faktormodelle limitiert. Da mit der wachsenden Anzahl von Faktoren die Komplexität der Struktur exponentiell zunimmt, ist es eine hochgradig nicht-triviale Aufgabe, solche hierarchischen Lasso-Algorithmen auf multifaktorielle Modelle zu erweitern. Diese These versucht, einen hierarchischen Lasso-Algorithmus für drei Faktoren zu entwickeln, der für Antworten geeignet ist, die durch Verteilungen aus der Exponentialfamilie modelliert werden. Die Methode sollte auf die Daten zur Deoxyfluorierung angewendet werden, die eine nicht-normale Antwort aufweisen.

Contents

List of Tables

List of Figures

1. Introduction

In experimental sciences many complex datasets are being generated from high throughput experimentation (HTE). Data scientists have to develop models that take into account all the specific data structures of these datasets. One example of a dataset that has drawn the interest of researchers is the Deoxyfluorination dataset [D23].

1.1. Motivation

1.1.1. Deoxyfluorination data

Deoxyfluorination is a chemical reaction that converts alcohols into fluorides, which are considered valuable targets in pharmaceutical research and development. An example of such a chemical reaction is shown in Figure 1.1, according to [D23]. An alcohol is combined with a sulfonyl fluoride and a base under specific reaction conditions, resulting in obtaining the final product.

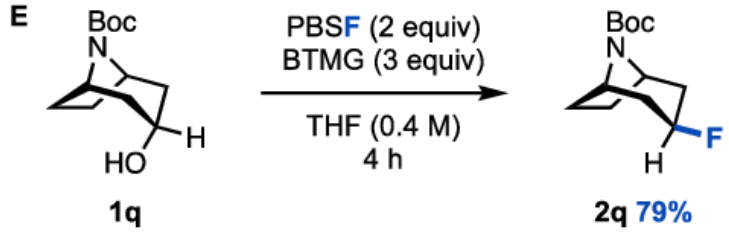

Figure 1.1: Deoxyfluorination

The dataset given in [D23] comprises 740 yield values (outputs or responses in a statistical model), taken under various reaction conditions. The yield is the ratio of moles of product to moles of reactant and can range from 0% to 100%. There are three reaction components (factors in a statistical model), referred to as alcohols, sulfonyl fluorides, and bases. These factor variables take 37, 5, and 4 possible values (also referred to as factor levels). The dataset contains exactly one observation of the yield per each factor level combination. The purpose of the analysis is to find factor level combinations that lead to high values of the yield.

From the histogram of the yield, shown in Figure 1.2, it is clear that the yield has a positive distribution with a compact support (it takes values only in $[0, 100]$, aligning with the physical interpretation of the yield). In addition to this, the distribution is right-skewed due to a higher frequency of lower yields.

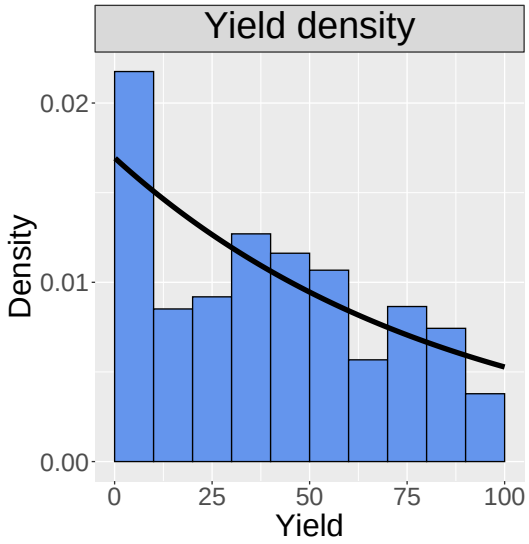

Figure 1.2: Yield density

It turns out that the distribution of the yield can be modelled by the continuous Bernoulli distribution, which, in particular, belongs to the exponential family. With this, the data generating process can be modelled via

$$\begin{aligned} g\{E(yield_{ijk})\} = \mu_0 &+ \alpha_i + \beta_j + \gamma_k \\ &+ (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} \\ &+ (\alpha\beta\gamma)_{ijk}, \\ &i = 1, \ldots, 37,\ j = 1, \ldots, 5,\ k = 1, \ldots, 4, \end{aligned} \tag{1.1}$$

where g is a canonical link function, μ_0 is the intercept, α_i, β_j and γ_k are referred to as the main effects, while the subsequent terms are the two-way and three-way interactions. More specifically, α_i is the i-th level of factor "alcohol", β_j is the j-th level of factor "sulfonyl fluoride", and γ_k is the k-th level of factor "base".

Equation 1.1 can be rewritten in a matrix form notation:

$$g\{E(yield)\} = Xv, \tag{1.2}$$

where $X \in \mathbb{R}^{740\times740}$ is a dummy matrix that describes which main effects and interactions are present in each sample, and $v \in \mathbb{R}^{740}$ is the vector of coefficients. The details on this matrix are given in the Chapter 3. Since there are as many observations as parameters, no consistent estimation of v is possible.

However, it seems reasonable to assume that v is sparse, that is, v has a relatively small number of non-zero components compared to the total number of components. Hence, it seems reasonable to apply Lasso to this problem. Lasso is a procedure that has the advantage of doing feature selection and model fitting simultaneously. First, one has to check if the data matrix is suitable for the direct usage of Lasso. Data

matrix should satisfy certain properties to ensure consistency of variable selection. In particular, restricted eigenvalue, or neighborhood stability [BvdG11] should hold. When the interpretability of the coefficients is important, the consistency of variable selection is also required. These type of conditions do not hold when the matrix X is too "ill posed" [BvdG11]. In our case, most of the columns of matrix X are obtained by element-wise multiplication of two other columns of the matrix, since two-way interactions are created by element-wise multiplication of the columns of the two corresponding main effects and three-way interactions are obtained by element-wise multiplication of the columns of the three corresponding main effects. Therefore, X is ill-posed, having a condition number of almost 200. Additionally, in most settings, the Lasso error depends on a constant similar in form to the compatibility constant from [vdG16]. The large number of zeros in the matrix X causes this constant to behave poorly, impacting the standard results on the performance of Lasso.

Moreover, one might argue that the data has a hierarchical structure. Indeed, if two components of the reaction have insignificant influence on the yield, it is reasonable to assume that their interaction cannot be significant as well. Consequently, it is sensible to assume and enforce hierarchy in the estimation. In this situation, directly using Lasso for feature selection may overlook the inherent pattern of the data, potentially leading to incorrect interpretation of the results. Consequently, one has to use a Lasso technique which is adapted for hierarchical data.

To present the problem in a thorough and formal manner, complete definitions of hierarchy will be provided.

1.1.2. Hierarchy

Let us consider model 1.1. In the following, we will present the definition of hierarchy for three-way interactions which implies two hierarchical layers: one between main effects and two-way interactions, and another between two-way interactions and three-way interactions.

Definition 1.1: Weak (Strong) Hierarchy for **three-way interactions** holds if the following two properties are satisfied:

i) Hierarchy between main effects and two-way interactions:

Weak Hierarchy: $\alpha_i = \beta_j = 0 \implies (\alpha\beta)_{ij} = 0$

Strong Hierarchy: either $\alpha_i = 0$ or $\beta_j = 0 \implies (\alpha\beta)_{ij} = 0$

ii) Hierarchy between two-way interactions and three-way interactions:

Weak Hierarchy: $(\alpha\beta)_{ij} = (\alpha\gamma)_{ik} = (\beta\gamma)_{jk} = 0 \implies (\alpha\beta\gamma)_{ijk} = 0$

Strong Hierarchy: either $(\alpha\beta)_{ij} = 0$ or $(\alpha\gamma)_{ik} = 0$ or $(\beta\gamma)_{jk} = 0 \implies (\alpha\beta\gamma)_{ijk} = 0$

One can easily observe that a model that follows the **strong hierarchy** constraints implicitly follows the **weak hierarchy** constraints. It is also worth noting that while

having the main effects to be 0 is sufficient to result in the corresponding interaction being 0, it is not necessarily required. An interaction can be 0 even if both of the corresponding main effects are not 0. Similarly, while having the two-way interactions set to 0 is sufficient to result in the corresponding three-way interaction being 0 , it is not strictly necessary. A three-way interaction can be 0 even if the three corresponding two-way interactions are not 0.

1.1.3. Existing methods

There are several works that come up with approaches to hierarchical models estimation ([JB13], [ZRY09], [NHCZ10]), but most of them are limited to hierarchy for two-way interactions and normal response. In [NHCZ10], the authors propose a hierarchical model employing equality constraints between the main effects and two-way interactions. In this thesis we aim to extend their approach to hierarchy that includes three-way interactions. This subsection will be concluded with a brief description of the method by [NHCZ10].

[NHCZ10] consider a regression model for an outcome $Y \in \mathbb{R}^n$ and a data matrix $X \in \mathbb{R}^{n \times \frac{p(p+1)}{2}}$, where n is the number of samples and p the number of main effects. The data matrix includes predictors $X_1, X_2, \ldots, X_p$ to describe the main effects, as well as the two-way interactions between these predictors $X_{1,1}, X_{1,2}, \ldots, X_{1,p}, \ldots, X_{p-1,p}$, where $X_{j,k}$ is the Hadamard product between columns X_j and X_k, for any j, $k \in \{1, \ldots, p\}$ with $j < k$. Hence, the regression model has the following form:

$$Y = \beta_0 + \sum_j \beta_j X_j + \sum_{j<k} \Theta_{jk} X_j X_k + \varepsilon, \tag{1.3}$$

where ε is the error term. The β terms are the coefficients of the main effects and the Θ terms are the coefficients of the interactions. Let us define the function $h : \mathbb{R}^{\frac{p(p+1)}{2}} \to \mathbb{R}$,

$$h(x) = \hat{\beta}_0 + \sum_j \hat{\beta}_j x_j + \sum_{j<k} \hat{\Theta}_{jk} x_j x_k,$$

which predicts the response of a sample given that the estimates $\hat{\beta}$ and $\hat{\Theta}$ are computed.

In order to induce hierarchy in the model, the authors considered the following constraints:

$$\hat{\Theta}_{jk} = \hat{\gamma}_{jk} \hat{\beta}_j \hat{\beta}_k, \ \forall\, j, \, k \in \{1, \ldots, p\} \text{ with } j < k. \tag{1.4}$$

One can observe that having a main effect equal to zero implies that all the associated two-way interactions are also zero. Therefore, strong hierarchy is satisfied. Consequently, the function h becomes:

$$h(x) = \hat{\beta}_0 + \sum_j \hat{\beta}_j x_j + \sum_{j<k} \hat{\gamma}_{jk} \hat{\beta}_j \hat{\beta}_k x_j x_k \tag{1.5}$$

For the variable selection, the authors suggest to find the hierarchical solution by minimizing the following penalized least square objective:

$$\min_{\beta,\gamma} \sum_{i=1}^{n} (y_i - h(x_i))^2 + \lambda_\beta \|\beta\|_1 + \lambda_\gamma \|\gamma\|_1, \tag{1.6}$$

where λ_β and λ_γ are two tuning parameters that control the amount of regularization for the main effects and interactions. The authors developed an algorithm that optimizes this objective iteratively for β and γ until convergence. The method can be directly extended to the generalized linear models framework by replacing the penalized least square objective with a penalized log-likelihood criterion.

1.2. Problem statement

Now that the motivation and key aspects of the dataset, which prompted the research, have been outlined, the problem can be clearly stated. Given that the dataset presented, has also three way interactions, the goal of this thesis is to extend the method from [NHCZ10] to **hierarchy** for **three-way interactions** and a **response from the exponential family**. The final goal is to apply the extended method on both synthetic datasets that contain three-way interactions and the Deoxyfluorination data.

Tackling this task proves to be challenging due to various reasons. The lack of theoretical frameworks for hierarchical models that incorporate more than two-factor interactions complicates the analysis. Additionally, extending optimization algorithms to address problems with three-way interactions increases exponentially the complexity of the problem.

Even if extending the current hierarchical methods to three-way interactions is an important step, there is future research that can still be done. There are datasets that contain four-way interactions such as the Buchwald-Hartwig amination data [AEL+18a], so extension to more hierarchical layers is of interest.

2. Literature review

The field of chemistry is currently undergoing significant changes due to the incorporation of data-driven models that facilitate the analysis of complex datasets. There are many papers that analyze datasets obtained using HTE (high throughput experimentation) ([D23], [AEL+18b], [TK23], [DPY+22]). Since the data obtained with HTE is usually very complex, it is tempting to use machine learning models to improve the efficiency of the analysis. Many of these datasets contain the yields of a reaction obtained under various reaction conditions. The purpose of the analysis of such datasets is usually to find the reaction conditions that lead to efficient reactions (i.e. high values of the yield) ([TK23], [DPY+22]). The focus of some papers ([NARD18], [AEL+18b]) is the performance of the yield prediction given the reaction conditions, and the models used are general machine learning models (e.g., random forest). These papers do not take into account all the specific data structures and the methods proposed might lead to uninterpretable or even poor results. There are also papers that focus more on the interpretability of the model ([TK23], [DPY+22]), considering the specific data structures [TK23].

To be more specific, we can use the Deoxyfluorination data [D23], which was thoroughly described in Section 1.1.1 of Chapter 1. The dataset comprises yield values (outputs or responses in a statistical model) and a matrix of descriptors (features or predictors). There are 740 yield values which result from measurements under all possible $37 \times 5 \times 4 = 740$ combinations of reaction components. These reaction components, referred to as alcohols, sulfonyl fluorides, and bases are factor variables that can take 37, 5, and 4 possible values (also known as factor levels). The dataset contains exactly one observation per factor level combination. As discussed in section 1.1.1, the designed experiment can be modeled in a matrix form by considering the contributions of main effects, two-way interactions, and three-way interactions: $g\{E(yield)\} = Xv$ (see equation (1.2)), where X can be seen as a dummy matrix, v contains the coefficients for the main effects and interactions, and the link function g is applied component-wise to the yield values. In this setting, the property of having single replicates per factor level combination implies that the number of parameters of the designed experiment is equal to the number of observations. Consequently, a model that induces sparsity has to be used.

Variable selection has been thoroughly studied in statistical literature ([Bre95], [Tib96]). Lasso [Tib96] has gained much attention in recent years due to its ability to provide sparse solutions through simultaneous variable selection and regularization. However, in many methods, such as in Lasso [Tib96], the variables are treated individually. When interaction terms are present in the data, as it is in Deoxyfluorination data, it is reasonable to assume that there may be a hierarchical pattern between the interactions and their corresponding main effects. In statistical modeling, hierarchy refers to the principle that interaction terms in a model should be included if only the corresponding main effects are

also included. This ensures that the model follows a natural structure where the presence of higher-order interactions is conditioned by the presence of the associated lower-order effects. Recalling the definitions of hierarchy from section 1.1.2 from Chapter 1 , strong hierarchy requires an interaction term to be set to zero if any of its corresponding main effects is zero, whereas weak hierarchy imposes that an interaction term is zero if both of its corresponding main effects are zero. Some statisticians even argue that violating strong hierarchy is usually not sensible [MN83]. These definitions were extended to third order interactions in section 1.1.2. Moreover, in the hierarchical framework, interactions between large main effects are expected to have a higher magnitude than the others. This behaviour would increase the interpretability of the model in practice. As a consequence, there has been significant attention given to creating hierarchical models in statistics and machine learning ([NHCZ10], [ZRY09], [JB13], [Nel94]). In some papers, the term hierarchy is alternatively referred to as heredity [NHCZ10] or marginality [Nel94].

There are various approaches to enforce hierarchy in a model, and in the subsequent part of this chapter, we will introduce three types. The first type is based on sequential modeling, the second on penalties that induce hierarchy and the third on constraints that impose the hierarchical structure.

One simple approach to build hierarchical models is to use a step-wise procedure that greedily adds or removes variables to the model with the condition that an interaction is included if only the corresponding main effects are included ([Agr02], [BRT10]). There are also models that do initially feature selection without considering hierarchy and then add all the lower order terms such that hierarchy is satisfied [NR12]. Other approaches include adapting well-known step-wise models to be hierarchical, one example being [YJL07] which modifies LARS [EHJT04] to enforce hierarchy.

A different approach to induce hierarchy in a model is to use specific penalties. CAP penalties [ZRY09] are a general class of penalties that can be used to achieve hierarchical sparsity. This type of penalty is based on the idea that it is possible to group and penalize terms such that if a specific term deviates from zero, the terms from a related group also deviate from zero. This method can be directly adapted to the hierarchical case by considering penalties that ensure that once an interaction is not zero, the associated main effects deviate from zero. Similar type of penalties were also used in [RJ10] to enforce hierarchy.

Finally, there are also models that induce hierarchy with the help of constraints. One example is [JB13], which proposes a hierarchical model obtained by using constraints on the magnitude of the interactions. The authors bound the size of the interactions by the size of the associated main effects. Intuitively, a null main effect sets all the corresponding interactions to zero due to these constraints. Another paper that proposes a hierarchical model with the help of constraints is [NHCZ10]. The authors create a model that can impose both strong and weak hierarchy by using equality constraints. Strong hierarchy is imposed by assuming that any interaction is the product of the associated main effects multiplied by a constant. As a result, any null main effect sets all the associated two-way interactions to zero. The authors also adapted the model for weak hierarchy by assuming that any interaction is the sum of the absolute values of the corresponding main effects

multiplied by a constant. This constraint ensures that an interaction corresponding to two null main effects is also null. One important aspect of this paper is that the authors managed to prove an oracle property for their model, stating that it is the first paper to prove such a property for a model that does variable selection with heredity constraints.

3. Data modelling

In this chapter, we will thoroughly explain how to model the Deoxyfluorination data by a GLM with continuous Bernoulli distribution response, as shown in the equation 1.2.

3.1. Matrix of descriptors

Recall equation 1.1, which describes the data generating process. In this section, we will show how equation 1.1 can be rewritten in the matrix form notation from equation 1.2.

For simplicity, we will describe how to obtain equation 1.2 for a model with two factors with an arbitrary number of levels. Generalization to three factors is straightforward but involves more complicated notations. We will closely follow [TK23] in this section. Assume there are two factors A and B, where factor A has I levels and factor B has J levels. Let us rewrite equation 1.1 for the case of two factors

$$g\{E(yield_{ij})\} = \mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij} \;\; i = 1, \ldots, I, \;\; j = 1, \ldots, J, \tag{3.1}$$

where μ_0 is the intercept, α_i is the contribution of the i-th level of factor A, β_j is the contribution of the j-th level of factor B, and $(\alpha\beta)_{ij}$ is the contribution of the interaction term between the two factors. For identifiability, it is assumed to have $\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$.

This model can be written as a regression model:

$$\begin{aligned} g\{E(yield_{ij})\} = \mu_0 + \alpha_1 x_1^A + \ldots + \alpha_{I-1} x_{I-1}^A + \beta_1 x_1^B + \ldots + \beta_{J-1} x_{J-1}^B \\ + (\alpha\beta)_{11} x_{1,1}^{AB} + \ldots + (\alpha\beta)_{I-1,J-1} x_{I-1,J-1}^{AB}, \end{aligned} \tag{3.2}$$

where

$$x_l^A = \begin{cases} 1, & \text{for } i = l \\ -1, & \text{for } i = I \\ 0, & \text{otherwise} \end{cases} \quad \text{for } l = 1, \ldots, I-1, \tag{3.3}$$

$$x_m^B = \begin{cases} 1, & \text{for } j = m \\ -1, & \text{for } j = J \\ 0, & \text{otherwise} \end{cases} \quad \text{for } m = 1, \ldots, J-1, \tag{3.4}$$

and $x_{l,m}^{A,B} = x_l^A \cdot x_m^B$. For example, for $I = 3$ and $J = 3$ the corresponding regression

model is:

$$g(Y) = g\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 \\ 1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\ 1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_0 \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{1,1} \\ (\alpha\beta)_{1,2} \\ (\alpha\beta)_{2,1} \\ (\alpha\beta)_{2,2} \end{pmatrix} =: Xv \tag{3.5}$$

From the identifiability conditions one finds $\alpha_3 = -\alpha_1 - \alpha_2$, $\beta_3 = -\beta_1 - \beta_2$, $(\alpha\beta)_{31} = -(\alpha\beta)_{11} - (\alpha\beta)_{21}$, $(\alpha\beta)_{32} = -(\alpha\beta)_{12} - (\alpha\beta)_{22}$, $(\alpha\beta)_{13} = -(\alpha\beta)_{11} - (\alpha\beta)_{12}$, $(\alpha\beta)_{23} = -(\alpha\beta)_{21} - (\alpha\beta)_{22}$, and $(\alpha\beta)_{33} = -(\alpha\beta)_{31} - (\alpha\beta)_{32}$.

Extension to three factors is straightforward. One should consider all the two-way interactions by multiplying all pairs of main effects that correspond to different factors, and the three-way interactions by multiplying any three main effects that correspond to different factors.

This section will be concluded by showing two important properties of our data matrix X. Firstly, we will show that it is invertible for the general case of $F \geq 2$ factors. To do this, some notations will be introduced. The number of levels of factor f will be denoted by L_f. Let us consider the matrix $\mathbf{X}$ to be our initial matrix X without the intercept column. It can be written in the following form

$$\mathbf{X} = \left[\mathbf{X}^{(1)} \,\middle|\, \cdots \,\middle|\, \mathbf{X}^{(F)} \,\middle|\, \mathbf{X}^{(1,2)} \,\middle|\, \cdots \,\middle|\, \mathbf{X}^{(1,\ldots,F)} \right],$$

where $\mathbf{X}^{(f_i)}$ represents the sub-matrix containing the columns for the levels of factor f_i and $\mathbf{X}^{(f_{i_1},\ldots,f_{i_k})}$ is the sub-matrix that contains the columns corresponding to the k-way interactions between factors $f_{i_1}, \ldots, f_{i_k}$. More specifically, these column sub-matrices have the following form

$$\mathbf{X}^{(f_{i_1},\ldots,f_{i_k})} := \left[\mathbf{X}^{(f_{i_1},\ldots,f_{i_k})}_{1,\ldots,1} \,\middle|\, \cdots \,\middle|\, \mathbf{X}^{(f_{i_1},\ldots,f_{i_k})}_{L_{f_{i_1}}-1,\ldots,L_{f_{i_k}}-1} \right],$$

where the exponents represent the factors and the indices the level corresponding to each factor. Note that this is just the generalization of the matrix X from equation (3.5) to F factors. The dummy matrix $\mathbf{X}$ has $n_F = L_1 \times \ldots \times L_F$ entries, and $p_F := n_F - 1$ columns. The latter claim follows by induction over the number of factors.

One can show that the sub-matrix $[\,\mathbf{X}^{(1)} \,|\, \cdots \,|\, \mathbf{X}^{(F)}\,] \in \mathbb{R}^{n_F \times (L_1 + \ldots + L_F - F)}$ consists of linearly independent columns. Since the remaining sub-matrix $[\,\mathbf{X}^{(1,2)} \,|\, \cdots \,|\, \mathbf{X}^{(1,\ldots,F)}\,] \in \mathbb{R}^{n_F \times (p_F - (L_1 + \ldots + L_F - F))}$ is obtained by element-wise multiplication of suitable columns of the first sub-matrix, this consists of linearly independent columns as well. This means that the enlarged matrix $[\mathbf{1}|\mathbf{X}]$ is square and invertible. Let us now state a result on the eigenvalues of the sample covariance matrix $\hat{\Sigma}_X = \mathbf{X}^T\mathbf{X}/n_F$. For simplicity, the

eigenvalues of the sample matrix $\mathbf{X}^T\mathbf{X}$ will be computed since the eigenvalues of $\hat{\Sigma}_X$ can be directly obtained by scaling with n_F We note that this result belongs to Dr. Gianluca Finocchio.

Lemma 3.1. *The eigenvalues of the sample matrix $\mathbf{X}^T\mathbf{X}$, where matrix $\mathbf{X}$ has factors with at least three levels, correspond to the products of the elements of all possible submultisets of $\{L_1, \ldots, L_F\}$ (an empty product is 1).*

Proof. Let $\mathcal{F} := \{1, 2, ..., F\}$ be the set of all factors. One can observe that the sample matrix is block diagonal and can be written as $\mathbf{X}^T\mathbf{X} = \text{diag}(B^{(1)}, ..., B^{(F)}, B^{(1,2)}, B^{(1,...,F)})$. For any $1 \leq r \leq F$ and any subset $\mathcal{F}' = \{f_{i_1}, f_{i_2}, ..., f_{i_r}\}$ of r factors, we denote by $\mathcal{L}_{\mathcal{F}'} := \mathcal{L}_{f_{i_1}} \times \mathcal{L}_{f_{i_2}} \times ... \times \mathcal{L}_{f_{i_r}}$. With $\mathcal{L}_f^0 = \mathcal{L}_f \backslash \{L_f\}$ the set obtained by removing the last element from level set $\mathcal{L}_f$, we denote $\mathcal{L}^0_{\mathcal{F}'} := \mathcal{L}^0_{f_{i_1}} \times \mathcal{L}^0_{f_{i_2}} \times \cdots \times \mathcal{L}^0_{f_{i_r}}$. With two multi-indexes $l := (l_1, l_2, ..., l_r) \in \mathcal{L}^0_{\mathcal{F}'}$ and $\tilde{l} := (\tilde{l}_1, \tilde{l}_2, ..., \tilde{l}_r) \in \mathcal{L}^0_{\mathcal{F}'}$, we denote $\|l - \tilde{l}\|_0$ the number of entries in which the two multi-indexes differ. There are $\binom{F}{r}$ r-way blocks $B^{(\mathcal{F}')} \in \mathbb{R}^{|\mathcal{L}^0_{\mathcal{F}'}| \times |\mathcal{L}^0_{\mathcal{F}'}|}$ corresponding to any subset $\mathcal{F}'$ of r factors, and their entries are

$$B^{(\mathcal{F}')}_{l,\tilde{l}} = (\mathbf{X}^{(\mathcal{F}')}_l)^T \mathbf{X}^{(\mathcal{F}')}_{\tilde{l}} = 2^{|\mathcal{F}'| - \|l-\tilde{l}\|_0} \cdot \frac{|\mathcal{L}_{\mathcal{F}}|}{|\mathcal{L}_{\mathcal{F}'}|}, \;\; (l, \tilde{l}) \in \mathcal{L}^0_{\mathcal{F}'} \times \mathcal{L}^0_{\mathcal{F}'}.$$

With the convention that $|L_\emptyset| := 1$, $|\mathcal{L}^{00}_f| := |\mathcal{L}_f| - 2$ and $|\mathcal{L}^0_f| := |\mathcal{L}_f| - 1$, their characteristic polynomials are given by

$$\begin{aligned} &\det\left(B^{(\mathcal{F}')} - \lambda \mathbf{I}_{|\mathcal{L}^0_{\mathcal{F}'}|}\right) \\ &= (-1)^{|\mathcal{L}^0_{\mathcal{F}'}|} \cdot \left(\lambda - \frac{|\mathcal{L}_{\mathcal{F}}|}{|\mathcal{L}_{\mathcal{F}'}|}\right)^{|\mathcal{L}_{\mathcal{F}^{00}}|} \cdot \prod_{k=1}^{|\mathcal{F}'|} \prod_{\substack{\mathcal{F}'' \subseteq \mathcal{F}' \\ |\mathcal{F}''| = k}} \left(\lambda - \frac{|\mathcal{L}_{\mathcal{F}}|}{|\mathcal{L}_{\mathcal{F}' \backslash \mathcal{F}''}|}\right)^{\frac{|\mathcal{L}^{00}_{\mathcal{F}}|}{|\mathcal{L}^{00}_{\mathcal{F}''}|}} \end{aligned} \tag{3.6}$$

By putting together all the different blocks in equation (3.6), one can observe that the eigenvalues of $\mathbf{X}^T\mathbf{X}$ correspond to the products of the elements of all possible subsets of $\{L_1, ..., L_F\}$ and the proof is completed. □

The range of the sample covariance matrix can be partitioned

$$\mathcal{R}(\hat{\Sigma}_x) = \bigoplus_{\mathcal{F}' \in \mathcal{P}(\mathcal{F}) \backslash \{\emptyset\}} \mathcal{R}(\hat{\Sigma}_{x_{\mathcal{F}'}}), \;\; rk(\hat{\Sigma}_{x_{\mathcal{F}'}}) = |\mathcal{L}^0_{\mathcal{F}'}|.$$

Each of the orthogonal subspaces $\mathcal{R}(\hat{\Sigma}_{x_{\mathcal{F}'}})$ is isometric to $\mathcal{R}(B^{(\mathcal{F}')})$ and can be written as trivial embedding into the enlarged space by padding suitable entries to 0. These subspaces measure the contribution of all the $|\mathcal{F}'|$-way effects involving the factors in $\mathcal{F}'$.

Let us exemplify the eigenvalues of the sample matrix $\mathbf{X}^T\mathbf{X}$, where X is the matrix defined in Equation (3.5) and $\mathbf{X}$ is the matrix X without the intercept term. Since $\mathbf{X}$ has two factors, each having three levels, the multiset $\{L_1, L_2\} = \{3, 3\}$. Hence, the submultisets are $\emptyset, \{3\}, \{3, 3\}$ and the distinct eigenvalues of the matrix are 1, 3 and 9.

3.2. Continuous Bernoulli distribution

We will closely follow the steps from the supplementary material of [TK23] to present the theory needed to use a GLM with a response modeled by a continuous Bernoulli distribution.

The density of the continuous Bernoulli distribution is given by

$$f(y;p) = \begin{cases} \frac{\log\{(1-p)/p\}}{1-2p} p^y (1-p)^{1-y}, & \text{if } p \in (0,1) \text{ and } p \neq \frac{1}{2}, \\ 2p^y(1-p)^{1-y}, & \text{if } p = \frac{1}{2}, \end{cases}$$

for $y \in [0,1]$. Rewriting the probability density function

$$f(y;p) = \exp\left[y \log\left(\frac{p}{1-p}\right) + \log(1-p) - \log(1-2p) + \log\left\{ \log\left(\frac{1-p}{p}\right)\right\}\right]$$

leads to the canonical parametrisation

$$f(y;\eta) = \exp\big\{ y\eta - \kappa(\eta)\big\} \tag{3.7}$$

with

$$\eta = \log(\frac{p}{1-p}) \quad \text{and} \quad \kappa(\eta) = \log\big\{ \exp(\eta) - 1\big\} - \log(\eta).$$

This implies for $Y \sim f(y;\eta)$

$$\kappa'(\eta) = \frac{1}{1-\exp(-\eta)} - \frac{1}{\eta} = E(Y)$$

$$\kappa''(\eta) = \frac{1}{\eta^2} - \frac{\exp(\eta)}{\{1-\exp(\eta)\}^2} = \text{var}(Y)$$

Let now $Y = (Y_1, \ldots, Y_{740})$ be the vector of yields, such that Y_i is distributed with the density $f(y_i;\eta_i)$, $i = 1, \ldots, 740$. Denote also $X \in \mathbb{R}^{740\times 740}$ the dummy matrix of predictors. To link the response vector Y and predictors X, one can employ a canonical link function, that is, $g = (\kappa')^{-1}$, so that $g\{E(Y_i)\} = g\{\kappa'(\eta_i)\} = \eta_i = X_i v$, where X_i denotes the i-th row of matrix X. This is exactly equation 1.2 written for each row. The canonical link function is not available analytically for the continuous Bernoulli distribution.

4. SHIM for three-way interactions

In this chapter, the SHIM model of [NHCZ10] will be extended to include three-way interactions. In the first section, the model for normal response and the optimization algorithm that estimates the coefficients will be described. In the second section, the model for the continuous Bernoulli response and its optimization algorithm will be presented.

Let us introduce the following notations: $\mathbf{X}_i$ denotes the i-th sample from the data matrix X, X_j represents the column corresponding to the main effect j, and $X_{i,j}$ is the value of the i-th sample for the main effect j.

4.1. SHIM for normal response

4.1.1. Model

Recall the model from equation (1.3). The model that includes three-way interactions directly extends (1.3) in the following way:

$$Y = \beta_0 + \sum_j \beta_j X_j + \sum_{j<k} \Theta_{jk} X_j X_k + \sum_{j<k<l} \Psi_{jkl} X_j X_k X_l + \varepsilon, \tag{4.1}$$

where Ψ is the coefficient for the interaction between the j-th, k-th and l-th main effect. The hierarchical conditions for two-way interactions remain as in equation (1.4), while the three-way interactions are assumed to have the form

$$\hat{\Psi}_{jkl} = \hat{\delta}_{jkl}\hat{\Theta}_{jk}\hat{\Theta}_{kl}\hat{\Theta}_{jl} = \hat{\delta}_{jkl}\hat{\gamma}_{jk}\hat{\gamma}_{kl}\hat{\gamma}_{jl}(\hat{\beta}_j\hat{\beta}_k\hat{\beta}_l)^2,$$

where $\hat{\delta}$ is a parameter that has to be estimated. Hence, the function that predicts the response for a sample, $\mathbf{X}_i$, which is defined for the two-way model in equation (1.5), has the following form in our three-way model:

$$h(\mathbf{X}_i) = \hat{\beta}_0 + \sum_j \hat{\beta}_j X_{i,j} + \sum_{j<k} \hat{\gamma}_{jk}\hat{\beta}_j\hat{\beta}_k X_{i,j}X_{i,k} + \sum_{j<k<l} \hat{\delta}_{jkl}\hat{\gamma}_{jk}\hat{\gamma}_{kl}\hat{\gamma}_{jl}(\hat{\beta}_j\hat{\beta}_k\hat{\beta}_l)^2 X_{i,j}X_{i,k}X_{i,l}.$$

Consequently, in order to find the hierarchical solution, we propose to minimize equation (4.2), which is a direct extension of (1.6):

$$\min_{\beta,\gamma,\delta} \sum_{i=1}^{n} (y_i - h(\mathbf{X}_i))^2 + \lambda_\beta \|\beta\|_1 + \lambda_\gamma \|\gamma\|_1 + \lambda_\delta \|\delta\|_1, \tag{4.2}$$

where λ_β, λ_γ, λ_δ are three tuning parameters that control the amount of regularization for the main effects, two-way interactions, and three-way interactions.

One important remark is that the model is not restricted to cases where the data matrix is discrete. Additionally, if the data matrix is discrete and involves three factors, as in our case, the model will consider only interactions between different factors. For example, equation (4.1) becomes:

$$Y = \beta_0 + \sum_j \beta_j X_j + \sum_{f(j)<f(k)} \Theta_{jk} X_j X_k + \sum_{f(j)<f(k)<f(l)} \Psi_{jkl} X_j X_k X_l + \varepsilon, \tag{4.3}$$

where $f(j)$ denotes the factor corresponding to main effect j.

4.1.2. Optimization

In this subsection, the algorithm for minimizing (4.2) will be presented. The steps of the optimization process will be enumerated initially, and then each step will be described in detail. Our algorithm is based on the algorithm proposed in [NHCZ10]. It consists of five steps, and the last four of them are repeated until a convergence criterion is met.

1. Find the intercept and initialize the other parameters with plausible values.
2. Update $\hat{\delta}$.
3. Update $\hat{\gamma}$.
4. Update $\hat{\beta}$.
5. Use a convergence criterion to decide whether the algorithm stops.

We denote by p_δ the length of $\hat{\delta}$.

1. Find the intercept and initialize the other parameters with plausible values. The intercept, $\hat{\beta}_0$, will be equal to the mean of all the responses. Hence, one can consider the model without an intercept by subtracting it from all the responses. The initialization of the parameters can be done using any consistent estimator of β, such as OLS or Lasso.

2. Update $\hat{\delta}$. The approach from [NHCZ10] will be followed to find the value of $\hat{\delta}$ that minimizes (4.2) for fixed $\hat{\beta}$ and $\hat{\gamma}$. Let us consider $\tilde{Y} \in \mathbb{R}^n$ and $\tilde{X} \in \mathbb{R}^{n \times p_\delta}$ such that

$$\tilde{Y}_i = Y_i - \sum_j \hat{\beta}_j X_{ij} - \sum_{j<k} \hat{\gamma}_{jk} \hat{\beta}_j \hat{\beta}_k X_{i,j} X_{i,k} \quad \text{and} \quad \tilde{X}_{i,jkl} = \hat{\gamma}_{jk} \hat{\gamma}_{kl} \hat{\gamma}_{jl} (\hat{\beta}_j \hat{\beta}_k \hat{\beta}_l)^2 X_{i,j} X_{i,k} X_{i,l}.$$

With this, the minimization problem becomes:

$$\hat{\delta} = \underset{\delta}{\operatorname{argmin}} \sum_i (\tilde{Y}_i - \sum_{j,k,l} \tilde{X}_{i,jkl} \delta_{jkl})^2 + \lambda_\delta \|\delta\|_1. \tag{4.4}$$

One can observe that (4.4) is a standard Lasso problem. Hence, its solution can be efficiently computed.

3. Update $\hat{\gamma}$. We will follow again the approach from [NHCZ10] and update iteratively each component of $\hat{\gamma}$, considering all other components of $\hat{\gamma}$ fixed, $\hat{\beta}$ fixed, and $\hat{\delta}$ fixed.

Therefore, equation (4.2) becomes a univariate minimization problem for each component of $\hat{\gamma}$. The algorithm for the component $\hat{\gamma}_{j'k'}$ will be presented, since the update rule is exactly the same for the other components of $\hat{\gamma}$. Let us consider $\tilde{Y} \in \mathbb{R}^n$ and $\tilde{X} \in \mathbb{R}^n$ such that

$$\tilde{Y}_i = Y_i - \sum_j \hat{\beta}_j X_{i,j} - \sum_{j,k \neq j',k'} \hat{\gamma}_{jk}(\hat{\beta}_j \hat{\beta}_k)\, X_{i,j} X_{i,k} - \sum_{\substack{jkl \\ j,k \neq j',k'}} \hat{\delta}_{jkl}(\hat{\gamma}_{jk}\hat{\gamma}_{kl}\hat{\gamma}_{jl})(\hat{\beta}_j \hat{\beta}_k \hat{\beta}_l)^2 X_{i,j} X_{i,k} X_{i,l}$$

$$\tilde{X}_i = \hat{\beta}_{j'} \hat{\beta}_{k'} X_{i,j'} X_{i,k'} + \sum_l \hat{\delta}_{j'k'l} \hat{\gamma}_{j'l} \hat{\gamma}_{k'l} (\hat{\beta}_{j'} \hat{\beta}_{k'} \hat{\beta}_l)^2 X_{i,j'k'l}.$$

With this, the minimization problem becomes:

$$\hat{\gamma}_{j'k'} = \underset{\gamma_{j'k'}}{\operatorname{argmin}} \sum_i (\tilde{Y}_i - \gamma_{j'k'} \tilde{X}_i)^2 + \lambda_\gamma \|\gamma_{j'k'}\|_1. \tag{4.5}$$

One can observe that (4.5) is a standard univariate Lasso problem. Hence, its solution can be efficiently computed.

4. Update $\hat{\beta}$. A similar approach to the update of $\hat{\gamma}$ will be used, where each component of $\hat{\beta}$ will be updated iteratively, while considering all other components of $\hat{\beta}$, $\hat{\gamma}$, and $\hat{\delta}$ fixed. Therefore, equation (4.2) becomes a univariate minimization problem for each component of $\hat{\beta}$. The algorithm for component $\hat{\beta}_{j'}$ will be presented, since the update rule is exactly the same for the other components of $\hat{\beta}$. Let us consider $\tilde{Y} \in \mathbb{R}^n$, $\tilde{X} \in \mathbb{R}^n$ and $\tilde{Z} \in \mathbb{R}^n$ such that

$$\tilde{Y}_i = Y_i - \sum_{j \neq j'} \hat{\beta}_j X_{i,j} - \sum_{\substack{j,k \\ j \neq j' \neq k}} \hat{\gamma}_{jk} \hat{\beta}_j \hat{\beta}_k X_{i,j} X_{i,k} - \sum_{\substack{jkl \\ j,k,l \neq j'}} \hat{\delta}_{jkl}(\hat{\gamma}_{jk}\hat{\gamma}_{kl}\hat{\gamma}_{jl})(\hat{\beta}_j \hat{\beta}_k \hat{\beta}_l)^2 X_{i,j} X_{i,k} X_{i,l},$$

$$\tilde{X}_i = X_{i,j'} + \sum_k \hat{\beta}_k \hat{\gamma}_{j'k} X_{i,j'} X_{i,k} \quad \text{and} \quad \tilde{Z}_i = \sum_{kl} \hat{\delta}_{j'kl} \hat{\gamma}_{j'k} \hat{\gamma}_{kl} \hat{\gamma}_{j'l} (\hat{\beta}_k \hat{\beta}_l)^2 X_{i,j'} X_{i,k} X_{i,l}.$$

With this, the minimization problem becomes:

$$\hat{\beta}_{j'} = \underset{\beta_{j'}}{\operatorname{argmin}} \sum_i (\tilde{Y}_i - \beta_{j'} \tilde{X}_i - \beta_{j'}^2 \tilde{Z}_i)^2 + \lambda_\beta \|\beta_{j'}\|_1. \tag{4.6}$$

One can observe that (4.13) is a univariate minimization problem. Hence, its solution can be efficiently computed.

5. Use a convergence criterion to decide whether the algorithm stops. The criterion from [NHCZ10] can be directly adapted to include three-way interactions. Let us compute the relative difference, $r^{(m)}$, between $Q_n(\hat{\beta}^{(m-1)}, \hat{\gamma}^{(m-1)}, \hat{\delta}^{(m-1)})$ and $Q_n(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}, \hat{\delta}^{(m)})$:

$$r^{(m)} = \frac{\left|Q_n(\hat{\beta}^{(m-1)}, \hat{\gamma}^{(m-1)}, \hat{\delta}^{(m-1)}) - Q_n(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}, \hat{\delta}^{(m)})\right|}{\left|Q_n(\hat{\beta}^{(m-1)}, \hat{\gamma}^{(m-1)}, \hat{\delta}^{(m-1)})\right|},$$

where

$$Q_n(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}, \hat{\delta}^{(m)}) = \sum_{i=1}^{n}(y_i - h^{(m)}(\mathbf{X}_i))^2 + \lambda_\beta\|\hat{\beta}^{(m)}\|_1 + \lambda_\gamma\|\hat{\gamma}^{(m)}\|_1 + \lambda_\delta\|\hat{\delta}^{(m)}\|_1$$

and $\hat{\beta}^{(m)}$, $\hat{\gamma}^{(m)}$, $\hat{\delta}^{(m)}$, $r^{(m)}$ and $h^{(m)}$ are the values obtained after iterating over steps two to five for m times. If $r^{(m)}$ is smaller than a pre-specified threshold, the algorithm stops, and the estimates of the parameters obtained at the m-th iteration are considered the final estimates of the algorithm.

It is worth noting that the convergence of the algorithm is guaranteed since the function Q decreases at each step and is positive by definition. However, as stated in [NHCZ10], the proposed criterion is not convex, and convergence to the global minimum is not theoretically guaranteed. It will be shown in the Simulations Chapter that the algorithm performs very well when a reasonable initialization of the parameters is used.

The choice of the tuning parameters, λ_β, λ_γ, and λ_δ, can be made using standard criteria such as cross-validation. Alternatively, one might try to find the λ values that shrink as many coefficients to zero as possible while still maintaining good predictive performance, to identify the main effects and interactions that significantly impact the response.

4.2. SHIM for continuous Bernoulli response

4.2.1. Model

Recall the model from equation (4.1). It can be adapted for a response modeled by a continuous Bernoulli distribution in the following way:

$$g\{E(Y)\} = \eta = \beta_0 + \sum_j \beta_j X_j + \sum_{j<k} \Theta_{jk} X_j X_k + \sum_{j<k<l} \Psi_{jkl} X_j X_k X_l, \tag{4.7}$$

where g is the canonical link function presented in Chapter 3, $\beta = (\beta_1 \dots \beta_p)^T$, and the intercept β_0 is considered a separate term. The hierarchical conditions for the interactions remain the same as in the normal response case, that is,

$$\hat{\Theta}_{j,k} = \hat{\gamma}_{j,k}\hat{\beta}_j\hat{\beta}_k \quad \text{and} \quad \hat{\Psi}_{jkl} = \hat{\delta}_{jkl}\hat{\Theta}_{jk}\hat{\Theta}_{kl}\hat{\Theta}_{jl} = \hat{\delta}_{jkl}\hat{\gamma}_{jk}\hat{\gamma}_{kl}\hat{\gamma}_{jl}(\hat{\beta}_j\hat{\beta}_k\hat{\beta}_l)^2.$$

Hence, the function that predicts the response for a sample, $\mathbf{X}_i$ has the following form: $h(\mathbf{X}_i) = \kappa'(\hat{\eta}_i)$, where

$$\hat{\eta}_i = \hat{\beta}_0 + \sum_j \hat{\beta}_j X_{i,j} + \sum_{j<k} \hat{\gamma}_{jk}\hat{\beta}_j\hat{\beta}_k X_{i,j}X_{i,k} + \sum_{j<k<l} \hat{\delta}_{jkl}\hat{\gamma}_{jk}\hat{\gamma}_{kl}\hat{\gamma}_{jl}(\hat{\beta}_j\hat{\beta}_k\hat{\beta}_l)^2 X_{i,j}X_{i,k}X_{i,l}$$

and κ is the function given in Chapter 3. Consequently, in order to find the hierarchical solution, we propose the following penalized negative log-likelihood criterion:

$$\min_{\beta,\gamma,\delta} \sum_{i=1}^{n} -\ell(\eta_i, y_i) + \lambda_\beta\|\beta\|_1 + \lambda_\gamma\|\gamma\|_1 + \lambda_\delta\|\delta\|_1, \tag{4.8}$$

where λ_β, λ_γ, λ_δ are three tuning parameters that control the amount of regularization and $\ell(\eta_i, y_i) = y_i\eta_i - \kappa(\eta_i)$ is the log-likelihood of the i-th sample.

As in the case of a normal response, we remark that the model is not restricted to cases where the data matrix is discrete. Additionally, if the data matrix is discrete and involves three factors, as in our case, the model will consider only interactions between different factors. For example, equation (4.7) becomes:

$$g\{E(Y)\} = \eta = \beta_0 + \sum_j \beta_j X_j + \sum_{f(j)<f(k)} \Theta_{jk} X_j X_k + \sum_{f(j)<f(k)<f(l)} \Psi_{jkl} X_j X_k X_l, \tag{4.9}$$

where $f(j)$ denotes the factor corresponding to main effect j.

4.2.2. Optimization

In this subsection, the algorithm for minimizing (4.8) will be presented. The optimization algorithm will be similar to that of the normal response case. The steps of the optimization process will be enumerated initially, and then each step will be described in detail. The algorithm consists of six steps, and the last five of them are repeated until a convergence criterion is met.

1. Initialize all the parameters ($\hat{\beta}_0$, $\hat{\beta}$, $\hat{\gamma}$, $\hat{\delta}$) with plausible values.
2. Update $\hat{\beta}_0$.
3. Update $\hat{\delta}$.
4. Update $\hat{\gamma}$.
5. Update $\hat{\beta}$.
6. Use a convergence criterion to decide whether the algorithm stops.

Let us denote by p_δ the length of $\hat{\delta}$. In all steps, we will use a similar approach as in the normal case and will consider all parameters fixed except for those being updated.

1. Initialize the parameters with plausible values. The initialization of the parameters can be done using any consistent estimator of β, such as MLE or Lasso for continuous Bernoulli response.

2. Update $\hat{\beta}_0$. Consider $\hat{\beta}$, $\hat{\gamma}$, $\hat{\delta}$ fixed and update $\hat{\beta}_0$ such that (4.8) is minimized. Let us consider $\tilde{\eta} \in \mathbb{R}^n$ such that

$$\tilde{\eta}_i = \sum_j \hat{\beta}_j X_{i,j} + \sum_{j<k} \hat{\gamma}_{jk}\hat{\beta}_j\hat{\beta}_k X_{i,j}X_{i,k} + \sum_{j<k<l} \hat{\delta}_{jkl}\hat{\gamma}_{jk}\hat{\gamma}_{kl}\hat{\gamma}_{jl}(\hat{\beta}_j\hat{\beta}_k\hat{\beta}_l)^2 X_{i,j}X_{i,k}X_{i,l}.$$

With this, the minimization problem becomes:

$$\hat{\beta}_0 = \underset{\beta_0}{\operatorname{argmin}} \sum_i \ell(\tilde{\eta}_i + \beta_0, Y_i) \tag{4.10}$$

One can observe that (4.10) is a univariate minimization problem. Hence, its solution can be efficiently computed.

3. Update $\hat{\delta}$. Find the value of $\hat{\delta}$ that minimizes (4.8) for $\hat{\beta}_0$ fixed, $\hat{\beta}$ fixed and $\hat{\gamma}$ fixed. Let us consider $C \in \mathbb{R}^n$ and $\tilde{X} \in \mathbb{R}^{n\times p_\delta}$ such that

$$C_i = \hat{\beta}_0 + \sum_j \hat{\beta}_j X_{ij} + \sum_{j<k} \hat{\gamma}_{jk}\hat{\beta}_j\hat{\beta}_k X_{i,j}X_{i,k} \text{ and } \tilde{X}_{i,jkl} = \hat{\gamma}_{jk}\hat{\gamma}_{kl}\hat{\gamma}_{jl}(\hat{\beta}_j\hat{\beta}_k\hat{\beta}_l)^2 X_{i,j}X_{i,k}X_{i,l}.$$

With this, the minimization problem becomes:

$$\hat{\delta} = \underset{\delta}{\operatorname{argmin}} \sum_i -\ell(C_i + \sum_{j,k,l} \tilde{X}_{i,jkl}\delta_{jkl}, Y_i) + \lambda_\delta \|\delta\|_1. \tag{4.11}$$

One can observe that (4.11) is a Lasso problem for a GLM with an offset term of C ($\tilde{X}$ contains the predictors, Y is the response, and δ is the coefficient to be estimated). The IRLS (iteratively reweighted least squares) formulation of the Fisher scoring algorithm can be efficiently used to compute the minimizer of the penalized negative log-likelihood for a standard GLM with a canonical link. In section A.1 of the Appendix, it is shown that the IRLS algorithm can be easily adapted when the model contains an offset term. Hence, the solution of the minimization problem (4.11) can be efficiently computed.

4. Update $\hat{\gamma}$. As in the normal response case, each component of $\hat{\gamma}$ will be updated iteratively. Consider all other components of $\hat{\gamma}$ fixed, $\hat{\beta}_0$ fixed, $\hat{\beta}$ fixed, and $\hat{\delta}$ fixed. Therefore, equation (4.8) becomes a univariate minimization problem for each component of $\hat{\gamma}$. The algorithm will be presented for the component $\hat{\gamma}_{j'k'}$ since the update rule is the same for the other components of $\hat{\gamma}$. Let us consider $C \in \mathbb{R}^n$ and $\tilde{X} \in \mathbb{R}^n$ such that

$$C_i = \sum_j \hat{\beta}_j X_{i,j} + \sum_{j,k \neq j',k'} \hat{\gamma}_{jk}(\hat{\beta}_j \hat{\beta}_k) X_{i,j} X_{i,k} + \sum_{\substack{jkl \\ j,k \neq j',k'}} \hat{\delta}_{jkl}(\hat{\gamma}_{jk}\hat{\gamma}_{kl}\hat{\gamma}_{jl})(\hat{\beta}_j \hat{\beta}_k \hat{\beta}_l)^2 X_{i,j} X_{i,k} X_{i,l}$$

$$\tilde{X}_i = \hat{\beta}_{j'}\hat{\beta}_{k'} X_{i,j'} X_{i,k'} + \sum_l \hat{\delta}_{j'k'l}\hat{\gamma}_{j'l}\hat{\gamma}_{k'l}(\hat{\beta}_{j'}\hat{\beta}_{k'}\hat{\beta}_l)^2 X_{i,j'k'l}.$$

With this, the minimization problem becomes:

$$\hat{\gamma}_{j'k'} = \underset{\gamma_{j'k'}}{\operatorname{argmin}} \sum_i -\ell(C_i + \gamma_{j'k'}\tilde{X}_i, Y_i) + \lambda_\gamma \|\gamma_{j'k'}\|_1. \tag{4.12}$$

One can observe that (4.12) is a univariate minimization problem. Hence, its solution can be efficiently computed.

5. Update $\hat{\beta}$. As in the case of $\hat{\gamma}$, each component of $\hat{\beta}$ will be updated iteratively, considering all other components of $\hat{\beta}$ fixed, $\hat{\beta}_0$ fixed, $\hat{\gamma}$ fixed, and $\hat{\delta}$ fixed. Therefore, equation (4.8) becomes a univariate minimization problem for each component of $\hat{\beta}$. The algorithm will be presented for $\hat{\beta}_{j'}$ since the update rule is exactly the same for the other components of $\hat{\beta}$. Let us consider $C \in \mathbb{R}^n$, $\tilde{X} \in \mathbb{R}^n$ and $\tilde{Z} \in \mathbb{R}^n$ such that

$$C = \sum_{j \neq j'} \hat{\beta}_j X_{i,j} + \sum_{\substack{j,k \\ j \neq j' \neq k}} \hat{\gamma}_{jk}\hat{\beta}_j\hat{\beta}_k X_{i,j}X_{i,k} + \sum_{\substack{jkl \\ j,k,l \neq j'}} \hat{\delta}_{jkl}(\hat{\gamma}_{jk}\hat{\gamma}_{kl}\hat{\gamma}_{jl})(\hat{\beta}_j\hat{\beta}_k\hat{\beta}_l)^2 X_{i,j}X_{i,k}X_{i,l},$$

$$\tilde{X}_i = X_{i,j'} + \sum_k \hat{\beta}_k \hat{\gamma}_{j'k} X_{i,j'} X_{i,k} \quad \text{and} \quad \tilde{Z}_i = \sum_{kl} \hat{\delta}_{j'kl}\hat{\gamma}_{j'k}\hat{\gamma}_{kl}\hat{\gamma}_{j'l}(\hat{\beta}_k\hat{\beta}_l)^2 X_{i,j'}X_{i,k}X_{i,l}.$$

With this, the minimization problem becomes:

$$\hat{\beta}_{j'} = \underset{\beta_{j'}}{\operatorname{argmin}} \sum_i \ell(C_i + \beta_{j'}\tilde{X}_i + \beta_{j'}^2 \tilde{Z}_i, Y_i) + \lambda_\beta \|\beta_{j'}\|_1. \tag{4.13}$$

One can observe that (4.13) is a univariate minimization problem. Hence, its solution can be efficiently computed.

6. Use a convergence criterion to decide whether the algorithm stops. This step will be almost identical to **Step 5** from the model for normal response. Let us compute the relative difference, $r^{(m)}$, between $Q_n(\hat{\beta}^{(m-1)}, \hat{\gamma}^{(m-1)}, \hat{\delta}^{(m-1)})$ and $Q_n(\hat{\beta}^{(m)}, \hat{\gamma}^{(m)}, \hat{\delta}^{(m)})$:

$$r^{(m)} = \frac{\left| Q_n(\hat{\beta}_0^{(m-1)}, \hat{\beta}^{(m-1)}, \hat{\gamma}^{(m-1)}, \hat{\delta}^{(m-1)}) - Q_n(\hat{\beta}_0^{(m)}, \hat{\beta}^{(m)}, \hat{\gamma}^{(m)}, \hat{\delta}^{(m)}) \right|}{\left| Q_n(\hat{\beta}_0^{(m-1)}, \hat{\beta}^{(m-1)}, \hat{\gamma}^{(m-1)}, \hat{\delta}^{(m-1)}) \right|},$$

where

$$Q_n(\hat{\beta}_0^{(m)}, \hat{\beta}^{(m)}, \hat{\gamma}^{(m)}, \hat{\delta}^{(m)}) = \sum_{i=1}^{n} -\ell(\hat{\eta}_i, y_i) + \lambda_\beta \|\hat{\beta}^{(m)}\|_1 + \lambda_\gamma \|\hat{\gamma}^{(m)}\|_1 + \lambda_\delta \|\hat{\delta}^{(m)}\|_1, \quad (4.14)$$

and $\hat{\beta}_0^{(m)}$, $\hat{\beta}^{(m)}$, $\hat{\gamma}^{(m)}$, $\hat{\delta}^{(m)}$, and $r^{(m)}$ are the values obtained after iterating over steps two to six for m times. If $r^{(m)}$ is smaller than a pre-specified threshold, the algorithm stops, and the estimates of the parameters obtained at the m-th iteration are considered the final estimates of the algorithm.

Similarly to the normal case, the convergence of the algorithm is guaranteed. However, as stated in [NHCZ10], the proposed criterion is not convex, and convergence to the global minimum is not theoretically guaranteed. It will be shown in the Simulations Chapter that the algorithm performs very well when a reasonable initialization of the parameters is used.

The choice of the tuning parameters, λ_β, λ_γ, and λ_δ, can be made using standard criteria such as cross-validation. Alternatively, one might try to find the λ values that shrink as many coefficients to zero as possible while still maintaining good predictive performance, to identify the main effects and interactions that significantly impact the response.

5. Theoretical aspects

The main goal of this thesis is to extend the method by [NHCZ10] to include hierarchy for three-way interactions. As mentioned in Chapter 1, SHIM is strongly based on Lasso because it incorporates ℓ_1-norm penalization. Consequently, this chapter will present several results on Lasso. In the first two sections [vdG16] is closely followed, while the third section is based on [BLT18]. In the fourth section, we will present an explanation of why the optimal rates cannot be achieved on the Deoxyfluorination data. Theoretical properties of SHIM will be studied in a separate work.

5.1. Preliminaries

Let us consider $X \in \mathbb{R}^{n \times p}$ an input matrix, $Y \in \mathbb{R}^n$ a vector of responses and the linear model

$$Y = X\beta^0 + \varepsilon, \tag{5.1}$$

where $\beta^0 \in \mathbb{R}^p$ is an unknown vector of coefficients and $\varepsilon \in \mathbb{R}^n$ is a mean-zero noise vector. The least squares method estimates the unknown β^0 by minimizing the Euclidean distance between Y and the space spanned by the columns in X:

$$\hat{\beta}_{\text{LS}} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2. \tag{5.2}$$

The least squares estimator $\hat{\beta}_{\text{LS}}$ is thus obtained by taking the coefficients of the projection of Y onto the column space of X. It satisfies the normal equations

$$X^T Y = X^T X \hat{\beta}_{\text{LS}}.$$

If $X^T X$ is invertible, which is equivalent to X having full rank p, one can get:

$$\hat{\beta}_{\text{LS}} = (X^T X)^{-1} X^T Y$$

However, in many cases, the matrix X has $p \geq n$ and the LS estimator from (5.2) is not feasible. It will just reproduce the data by returning an estimator such that $X\hat{\beta}_{\text{LS}} = Y$. This phenomenon is known as overfitting. To address this issue, one can use least squares loss with an ℓ_1-regularization penalty in order to prevent overfitting. This method is called the **Lasso** and has the following form:

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left(\frac{1}{n} \|Y - X\beta\|_2^2 + 2\lambda \|\beta\|_1 \right).$$

There are various definitions of Lasso in the literature that are equivalent up to a scaling factor, allowing them to be used interchangeably. In this chapter, bounds on the **estimation error** and **prediction error** of Lasso will be presented. Consequently, let us define these terms formally. The ℓ_q-norm **estimation error** for the regression model (5.1) is

$$\|\hat{\beta} - \beta^0\|_q,$$

while the **prediction error** is

$$\frac{1}{n}\|Y - X\hat{\beta}\|_2^2.$$

We note that a small prediction error does not necessarily imply a small estimation error, particularly when there is a high dependency between the columns of X.

The Lasso estimator $\hat{\beta}$ satisfies the Karush-Kuhn-Tucker conditions which state that

$$X^T(Y - X\hat{\beta})/n = \lambda\hat{z},$$

where $\hat{z}$ is a p-dimensional vector with $\|\hat{z}\|_\infty \leq 1$ and with $\hat{z}_j = \text{sign}(\hat{\beta}_j)$ if $\hat{\beta}_j \neq 0$. The latter can also be written as

$$\hat{z}^T\hat{\beta} = \|\hat{\beta}\|_1.$$

The KKT conditions imply the *almost orthogonality* of X and the residuals, in the sense that

$$\|X^T(Y - X\hat{\beta})\|_\infty/n \leq \lambda$$

Let us introduce a set of notations that will be useful later in this chapter. For a vector $v \in \mathbb{R}^n$, we define $\|v\|_n^2 = v^Tv/n = \|v\|_2^2/n$, where $\|\cdot\|_2$ is the ℓ_2-norm. The normalized Gram matrix is written as $\hat{\Sigma} = X^TX/n$. Thus, $\|X\beta\|_n^2 = \beta^T\hat{\Sigma}\beta$, where $\beta \in \mathbb{R}^p$. Let $S \subseteq \{1, \ldots, p\}$ be an index set. The vector $\beta_S \in \mathbb{R}^p$ is defined as

$$\beta_{j,S} = \beta_j 1\{j \in S\}, \quad j = 1, \ldots, p.$$

Thus, β_S is a p-vector with entries equal to zero for indices $j \notin S$. The vector β_{-S} has all entries inside the set S set to zero, i.e., $\beta_{-S} = \beta_{S^c}$, where $S^c = \{j \in \{1, \ldots, p\} \mid j \notin S\}$ is the complement of the set S. These notations allow us to write $\beta = \beta_S + \beta_{-S}$. The active set S_β of a vector $\beta \in \mathbb{R}^p$ is defined as $S_\beta = \{j \mid \beta_j \neq 0\}$. For a solution, β^0, the active set is denoted by $S_0 := S_{\beta^0}$, and the cardinality of this active set by $s_0 := |S_0|$. The j-th column of X is denoted by X_j, for $j = 1, \ldots, p$ (and if there is little risk of confusion, we also write X_i for the i-th row of the matrix X, where $i = 1, \ldots, n$). For a set $S \subseteq \{1, \ldots, p\}$, the matrix containing only the columns in the set S is denoted by X_S. To fix the ordering of the columns, they are arranged in increasing order of the indices. The "complement" matrix of X_S is denoted by $X_{-S} := \{X_j\}_{j \notin S}$.

Let us introduce two important results. The *two-point inequality* for $\hat{\beta}$ says that $\forall\beta \in \mathbb{R}^p$:

$$(\beta - \hat{\beta})^T X^T(Y - X\hat{\beta})/n \leq \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1. \tag{5.3}$$

and the *two point margin* for β^0 that for any β and β':

$$2(\beta' - \beta)^T \hat{\Sigma}(\beta' - \beta^0) = \|X(\beta' - \beta^0)\|_n^2 - \|X(\beta - \beta^0)\|_n^2 + \|X(\beta' - \beta)\|_n^2. \tag{5.4}$$

Thus, the *two-point inequality* can be written in the alternative form as

$$\|Y - X\hat{\beta}\|_n^2 - \|Y - X\beta\|_n^2 + \|X(\hat{\beta} - \beta)\|_n^2 \leq 2\lambda\|\beta\|_1 - 2\lambda\|\hat{\beta}\|_1, \quad \forall\, \beta. \tag{5.5}$$

One important property of the solution of Lasso is that it satisfies (5.3). In order to present the theoretical bounds of Lasso, is necessary to define the *compatibility constant*.
Definition 5.1: For a constant $L > 1$ and an index set S, the *compatibility constant* is defined as

$$\hat{\phi}^2(L, S) = \min\left\{|S|\|X\beta_S - X\beta_{-S}\|_n^2 \,:\, \|\beta_S\|_1 = 1 \text{ and } \|\beta_{-S}\|_1 \leq L\right\},$$

where L is called the stretching factor.
The compatibility constant reflects the degree of dependency between the columns of X. A small compatibility constant indicates greater dependency. We will see that some conclusions of the following results will depend on bounding this constant away from 0.

Let us finish this subsection by stating two inequalities that are used to prove the main theorem from the following section. The so-called *triangle property* is a direct consequence of the fact that $\beta = \beta_S + \beta_{-S}$ and states that for any β, β' and any set S:

$$\|\beta\|_1 - \|\beta'\|_1 \leq \|\beta_S - \beta'_S\| - \|\beta_{-S}\|_1 - \|\beta'_{-S}\|_1.$$

The *dual norm inequality* says that for any two vectors w and β

$$|w^T\beta| \leq \|w\|_\infty\|\beta\|_1. \tag{5.6}$$

5.2. Standard bounds for Lasso

This section presents standard bounds for Lasso prediction and estimation errors. Let us start with a result that is a particular case of a more general theorem.

Theorem 5.1. *Let λ_ε satisfy*

$$\lambda_\varepsilon \geq \frac{\|X^T\varepsilon\|_\infty}{n}, \tag{5.7}$$

where ε is the noise vector in regression model (5.1). Define for $\lambda > \lambda_\varepsilon$

$$\underline{\lambda} = \lambda - \lambda_\varepsilon, \;\; \overline{\lambda} = \lambda + \lambda_\varepsilon$$

and stretching factor

$$L := \frac{\overline{\lambda}}{\underline{\lambda}}.$$

Then

$$\|X(\hat{\beta} - \beta^0)\|_n^2 \leq \min_S \left\{ \min_{\beta \in \mathbb{R}^p, S_\beta = S} \|X(\beta - \beta^0)\|_n^2 + \frac{\overline{\lambda}^2|S|}{\hat{\phi}^2(L, S)} \right\}$$

An important remark regarding this theorem is that λ is strictly bounded below by a quantity dependent on ε. Since ε is generally unknown in practice, the bound on λ typically holds with "high probability". Further details on how λ can be selected to ensure the results of the theorem hold with "high probability" will be provided later in this section. Another remark is that the Theorem 5.1 implies that if β^0 can be closely approximated by a sparse vector β, i.e., a vector β with a small active set S_β, then the prediction error of the Lasso is small, with high probability, when the compatibility constant stays away from zero. Theorem 5.2 improves upon Theorem 5.1. The prediction error can still be bounded by a small quantity when the compatibility constant, $\hat{\phi}^2(L, S_\beta)$ behaves badly, but the ℓ_1-norm of the target is not too large.

Theorem 5.2. *Let λ_ε satisfy*

$$\lambda_\varepsilon \geq \frac{\|X^T\varepsilon\|_\infty}{n}.$$

Let $0 \leq \delta < 1$ be arbitrarily and define for $\lambda > \lambda_\varepsilon$

$$\underline{\lambda} = \lambda - \lambda_\varepsilon, \;\; \overline{\lambda} = \lambda + \lambda_\varepsilon + \delta\underline{\lambda},$$

as well as, a stretching factor

$$L := \frac{\overline{\lambda}}{(1-\delta)\underline{\lambda}}.$$

Then for all $\beta \in \mathbb{R}^p$ and all sets S

$$2\delta\underline{\lambda}\|\hat{\beta} - \beta\|_1 + \|X(\hat{\beta} - \beta^0)\|_n^2 \leq \|X(\beta - \beta^0)\|_n^2 + \frac{\overline{\lambda}^2|S|}{\hat{\phi}^2(L, S)} + 4\lambda\|\beta_{-S}\|_1.$$

Proof. The proof of the theorem will be divided into two parts depending on which of $(\hat{\beta} - \beta)^T\hat{\Sigma}(\hat{\beta} - \beta^0)$ and $-\delta\underline{\lambda}\|\hat{\beta} - \beta\|_1 + 2\lambda\|\beta_{-S}\|_1$ is larger. If

$$(\hat{\beta} - \beta)^T\hat{\Sigma}(\hat{\beta} - \beta^0) \leq -\delta\underline{\lambda}\|\hat{\beta} - \beta\|_1 + 2\lambda\|\beta_{-S}\|_1,$$

one can find from the two-point margin that

$$\begin{aligned}
&2\delta\underline{\lambda}\|\hat{\beta} - \beta\|_1 + \|X(\hat{\beta} - \beta^0)\|_n^2 \\
&= 2\delta\underline{\lambda}\|\hat{\beta} - \beta\|_1 + \|X(\beta - \beta^0)\|_n^2 - \|X(\beta - \hat{\beta})\|_n^2 + 2(\hat{\beta} - \beta)^T\hat{\Sigma}(\hat{\beta} - \beta^0) \\
&\leq \|X(\beta - \beta^0)\|_n^2 + 4\lambda\|\beta_{-S}\|_1,
\end{aligned}$$

and the desired result is obtained. From now on, one may therefore assume that

$$(\hat{\beta} - \beta)^T\hat{\Sigma}(\hat{\beta} - \beta^0) \geq -\delta\underline{\lambda}\|\hat{\beta} - \beta\|_1 + 2\lambda\|\beta_{-S}\|_1.$$

Using that $Y = X\beta^0 + \varepsilon$ in two-point inequality (5.3), one can get

$$(\hat{\beta} - \beta)^T\hat{\Sigma}(\hat{\beta} - \beta^0) \leq (\hat{\beta} - \beta)^T X^T\varepsilon/n + \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1.$$

By the dual norm inequality (5.6)

$$|(\hat{\beta}-\beta)^T X^T \varepsilon|/n \leq \lambda_\varepsilon \|\hat{\beta}-\beta\|_1.$$

Thus

$$\begin{aligned}
&(\hat{\beta}-\beta)^T \hat{\Sigma}(\hat{\beta}-\beta^0) \\
&\leq \lambda_\varepsilon \|\hat{\beta}-\beta\|_1 + \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1 \\
&\leq \lambda_\varepsilon \|\hat{\beta}_S - \beta_S\|_1 + \lambda_\varepsilon \|\hat{\beta}_{-S}\|_1 + \lambda_\varepsilon \|\beta_{-S}\|_1 + \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1.
\end{aligned}$$

By the triangle property and invoking $\underline{\lambda} = \lambda - \lambda_\varepsilon$, this implies:

$$(\hat{\beta}-\beta)^T \hat{\Sigma}(\hat{\beta}-\beta^0) + \underline{\lambda}\|\hat{\beta}_{-S}\|_1 \leq (\lambda+\lambda_\varepsilon)\|\hat{\beta}_S - \beta_S\|_1 + (\lambda+\lambda_\varepsilon)\|\beta_{-S}\|_1,$$

and by adding $\underline{\lambda}\|\beta_{-S}\|_1$ to both sides one obtains

$$(\hat{\beta}-\beta)^T \hat{\Sigma}(\hat{\beta}-\beta^0) + \underline{\lambda}\|\hat{\beta}_{-S} - \beta_{-S}\|_1 \leq (\lambda+\lambda_\varepsilon)\|\hat{\beta}_S - \beta_S\|_1 + 2\lambda\|\beta_{-S}\|_1.$$

Hence, invoking $\overline{\lambda} = \lambda + \lambda_\varepsilon + \delta\underline{\lambda}$ and adding $\delta\underline{\lambda}\|\hat{\beta}_S - \beta_S\|_1$ to both terms, one can get

$$(\hat{\beta}-\beta)^T \hat{\Sigma}(\hat{\beta}-\beta^0) + \underline{\lambda}\|\hat{\beta}_{-S} - \beta_{-S}\|_1 + \delta\underline{\lambda}\|\hat{\beta}_S - \beta_S\|_1 \leq \overline{\lambda}\|\hat{\beta}_S - \beta_S\|_1 + 2\lambda\|\beta_{-S}\|_1. \quad (5.8)$$

Since

$$(\hat{\beta}-\beta)^T \hat{\Sigma}(\hat{\beta}-\beta^0) \geq -\delta\underline{\lambda}\|\hat{\beta}-\beta\|_1 + 2\lambda\|\beta_{-S}\|_1,$$

this gives:

$$(1-\delta)\underline{\lambda}\|\hat{\beta}_{-S} - \beta_{-S}\|_1 \leq \overline{\lambda}\|\hat{\beta}_S - \beta_S\|_1,$$

or

$$\|\hat{\beta}_{-S} - \beta_{-S}\|_1 \leq L\|\hat{\beta}_S - \beta_S\|_1.$$

But then by the definition of the compatibility constant:

$$\|\hat{\beta}_S - \beta_S\|_1 \leq \sqrt{|S|}\|X(\hat{\beta}-\beta)\|_n/\hat{\phi}(L,S).$$

Continuing with inequality (5.8) and using that $ab \leq (a^2+b^2)/2$:

$$\begin{aligned}
&(\hat{\beta}-\beta)^T \hat{\Sigma}(\hat{\beta}-\beta^0) + \underline{\lambda}\|\hat{\beta}_{-S} - \beta_{-S}\|_1 + \delta\underline{\lambda}\|\hat{\beta}_S - \beta_S\|_1 \\
&\leq \overline{\lambda}\sqrt{|S|}\|X(\hat{\beta}-\beta)\|_n/\hat{\phi}(L,S) + 2\lambda\|\beta - \beta_S\|_1 \\
&\leq \frac{1}{2}\tilde{\lambda}^2|S|/\hat{\phi}^2(L,S) + \frac{1}{2}\|X(\hat{\beta}-\beta)\|_n^2 + 2\lambda\|\beta_{-S}\|_1.
\end{aligned}$$

Invoking the two-point margin:

$$2(\hat{\beta}-\beta)^T \hat{\Sigma}(\hat{\beta}-\beta^0) = \|X(\hat{\beta}-\beta^0)\|_n^2 - \|X(\beta-\beta^0)\|_n^2 + \|X(\hat{\beta}-\beta)\|_n^2,$$

one can infer that

$$\|X(\hat{\beta}-\beta^0)\|_n^2 + 2\underline{\lambda}\|\hat{\beta}_{-S} - \beta_{-S}\|_1 + 2\delta\underline{\lambda}\|\hat{\beta}_S - \beta_S\|_1$$

$$\leq \|X(\beta - \beta^0)\|_n^2 + \overline{\lambda}^2 |S| / \hat{\phi}^2(L, S) + 4\lambda \|\beta_{-S}\|_1.$$

which implies

$$2\delta\underline{\lambda}\|\hat{\beta} - \beta\|_1 + \|X(\hat{\beta} - \beta^0)\|_n^2 \leq \|X(\beta - \beta^0)\|_n^2 + \frac{\overline{\lambda}^2 |S|}{\hat{\phi}^2(L, S)} + 4\lambda \|\beta_{-S}\|_1.$$

□

One can easily observe that setting $\delta = 0$ and $S = S_\beta$ leads to Theorem 5.1. As with the previous theorem, the first important remark is that the result of the Theorem 5.2 is also probabilistic, but it can be ensured to hold with high probability. Secondly, by setting $\beta = \beta^0$ and $S = S_0$, it can be inferred from Theorem 5.2 that the prediction error $\|X(\hat{\beta} - \beta^0)\|_n^2$ is bounded by $\overline{\lambda}^2 s_0 / \hat{\phi}^2(L, S_0)$, where S_0 is the active set of β^0. Let us now state an important corollary of this theorem that provides a bound on the estimation error.

Corollary 5.1. *By taking $\beta = \beta^0$, $S = S_0$, and denoting $s_0 := |S_0|$, the estimation error, under the conditions of Theorem 5.2, is bounded as follows:*

$$\|\hat{\beta} - \beta^0\|_1 \leq \frac{\overline{\lambda}^2 s_0}{2\delta\underline{\lambda}\hat{\phi}^2(L, S_0)}.$$

As an example, one can choose $\lambda = 2\lambda_\varepsilon$ and $\delta = 1/5$ to obtain:

$$\|\hat{\beta} - \beta^0\|_1 \leq C_0 \frac{\lambda_\varepsilon s_0}{\hat{\phi}^2(4, S_0)},$$

where $C_0 = 16^2/10$.

We will now state the lemma that guarantees the condition $\lambda_\varepsilon \geq \|X^T\varepsilon\|_\infty / n$ holds with a given probability. This lemma is useful for ensuring that the results from Theorem 5.2 and its corollaries hold with high probability.

Lemma 5.1. *Let $\varepsilon \sim N_n(0, \sigma_0^2 I)$ and let X be a fixed $n \times p$ matrix with $diag(\hat{\Sigma})/n = I$. Let $0 < \alpha < 1$ be a given error level. Then, for*

$$\lambda_\varepsilon = \sigma_0 \sqrt{\frac{2\log(2p/\alpha)}{n}},$$

it is true that

$$P\left(\frac{\|X^T\varepsilon\|_\infty}{n} \leq \lambda_\varepsilon\right) \geq 1 - \alpha.$$

Corollary 5.2. *For $\lambda_\varepsilon = \sigma_0\sqrt{2\log(2p/\alpha)/n}$ the results from Theorem 5.2 and Corollary 5.1 hold with probability $1 - \alpha$.*

This subsection will be finished by specifying the order of the bound for Lasso estimation error that is obtained in Corollary 5.1 with λ_ε suitably chosen as in Lemma 5.1. To this end, let us consider the ℓ_1 estimation error bound:

$$C_0 \frac{\lambda_\varepsilon s_0}{\hat{\phi}^2(4, S_0)} = O(\lambda_\varepsilon s_0) = O\left(s_0\sqrt{\frac{\log(p)}{n}}\right) \tag{5.9}$$

5.3. Minimax optimal bounds for Lasso

In this subsection, the minimax optimal bounds for the prediction and estimation errors for Lasso are presented, following the approach outlined in [BLT18]. The bounds are obtained through a similar mechanism to the previous subsection. First, the randomness is removed by considering an event with a "high probability" of occurring. Then, the error bounds are derived on this event using a purely deterministic argument.

Let us start by introducing two notations. For a given $\delta_0 \in (0,1)$ and for any $u = (u_1, \dots, u_p) \in \mathbb{R}^p$, we set

$$H(u) := (4+\sqrt{2}) \sum_{j=1}^{p} u_j^{\#} \sigma \sqrt{\frac{\log(2p/j)}{n}}, \tag{5.10}$$

$$G(u) := (4+\sqrt{2}) \sigma \sqrt{\frac{\log(1/\delta_0)}{n}} \|Xu\|_n, \tag{5.11}$$

where $(u_1^{\#}, \dots, u_p^{\#})$ is a nonincreasing rearrangement of $(|u_1|, \dots, |u_p|)$. Now, the theorem that removes the randomness can be stated.

Theorem 5.3. *Let $\delta_0 \in (0,1)$ and let $X \in \mathbb{R}^{n \times p}$ be a matrix such that*

$$\max_{j=1,\dots,p} \|Xe_j\|_n \leq 1,$$

where $\{e_j | j = \overline{1:p}\}$ is the canonical basis of $\mathbb{R}^p$. Let $H(\cdot)$ and $G(\cdot)$ be defined as in (5.10) and (5.11). If $\varepsilon \sim N(0, \sigma^2 I_{n \times n})$, then the random event

$$\left\{ \frac{1}{n} \varepsilon^T X u \leq \max\left\{H(u), G(u)\right\}, \ \forall u \in \mathbb{R}^p \right\} \tag{5.12}$$

is of probability at least $1 - \delta_0/2$.

In order to present the theorem with the optimal rates, a restricted eigenvalue condition should be introduced.

Definition 5.2: The design matrix X satisfies the *strong restricted eigenvalue condition*, $SRE(s, c_0)$, if $\|Xe_j\|_n \leq 1$ for all $j = 1, \dots, p$, and

$$\theta(s, c_0) := \min_{\delta \in C_{SRE}(s,c_0): \delta \neq 0} \frac{\|X\delta\|_n}{\|\delta\|_2} > 0, \tag{5.13}$$

where $C_{SRE}(s, c_0) := \{\delta \in \mathbb{R}^p : \|\delta\|_1 \leq (1+c_0)\sqrt{s}\|\delta\|_2\}$ is a cone in $\mathbb{R}^p$. The last notation that is needed is

$$\delta(\lambda) := \exp\left[-\left\{ \frac{\gamma \lambda \sqrt{n}}{(4+\sqrt{2})\sigma} \right\}^2 \right],$$

which is equivalent to

$$\lambda = \frac{(4+\sqrt{2})\sigma}{\gamma} \sqrt{\frac{\log\left(1/\delta(\lambda)\right)}{n}},$$

where γ, n, σ are already fixed. The main theorem of this subsection can now be clearly stated.

Theorem 5.4. *Let $s \in \{1, \ldots, p\}$, $\gamma \in (0,1)$, and $\tau \in [0, 1-\gamma)$. Assume that the $SRE(s, c_0)$ condition holds with $c_0 = c_0(\gamma, \tau) = (1+\gamma+\tau)/(1-\gamma-\tau)$. Let λ be a tuning parameter such that $\delta(\lambda) \leq s/(2ep)$ and $\delta_0 \in (0,1)$. Then, on the event (5.12), the Lasso estimator $\hat{\beta}$ with tuning parameter λ satisfies*

$$2\tau\lambda\|\hat{\beta} - \beta\|_1 + \|X\hat{\beta} - EY\|_n^2 \leq \|X\beta - EY\|_n^2 + C_{\gamma,\tau}(s, \lambda, \delta_0)\lambda^2 s, \tag{5.14}$$

for all $\beta \in \mathbb{R}^p$ such that $\|\beta\|_0 \leq s$, where

$$C_{\gamma,\tau}(s, \lambda, \delta_0) := (1+\gamma+\tau)^2 \, max \left\{ \frac{\log(1/\delta_0)}{s \, log(1/\delta(\lambda))}, \frac{1}{\theta^2(s, c_0(\gamma, \tau))} \right\}$$

Furthermore, if $EY = X\beta^*$ for some $\beta^* \in \mathbb{R}^p$ with $|\beta^*|_0 \leq s$, then on the event (5.12), for any $1 \leq q \leq 2$,

$$\|\hat{\beta} - \beta^*\|_q \leq \left(\frac{C_{\gamma,\tau}}{2\tau} \right)^{2/q-1} \left(\frac{C_{\gamma,0}}{1+\gamma} \right)^{2-2/q} \lambda s^{1/q}. \tag{5.15}$$

Proof. Using the result from Lemma A.2 with $h(\cdot) = 2\lambda\|\cdot\|_1$, one can get that, almost surely, for all $\beta \in \mathbb{R}^p$,

$$2\tau\lambda\|\hat{\beta} - \beta\|_1 + \|X\hat{\beta} - EY\|_n^2 \leq \|X\beta - EY\|_n^2 - \|X(\hat{\beta} - \beta)\|_n^2 + \Delta^*, \tag{5.16}$$

where

$$\Delta^* := 2\tau\lambda\|\hat{\beta} - \beta\|_1 + \frac{2}{n}\varepsilon^T X(\hat{\beta} - \beta) + 2\lambda\|\beta\|_1 - 2\lambda\|\hat{\beta}\|_1.$$

Let $u = \hat{\beta} - \beta$ and assume that $\|\beta\|_0 \leq s$. Define

$$\tilde{H}(u) = \frac{(4+\sqrt{2})\sigma}{\sqrt{n}} \left\{ \|u\|_2 \left(\sum_{j=1}^{s} \log\left(\frac{2p}{j}\right) \right)^{1/2} + \sum_{j=s+1}^{p} u_j^{\#} \sqrt{\log\left(\frac{2p}{j}\right)} \right\}. \tag{5.17}$$

Using the Cauchy-Schwarz inequality, it is easy to see that

$$H(u) \leq \tilde{H}(u) \leq \gamma\lambda \left(\sqrt{s}\|u\|_2 + \sum_{j=s+1}^{p} u_j^{\#} \right) := F(u) \quad \forall u \in \mathbb{R}^p, \tag{5.18}$$

where the last inequality follows from Stirling's formula and $\delta(\lambda) \leq s/(2ep)$. On the event (5.12), using (5.18) and Lemma A.1 one can get

$$\begin{aligned} \Delta^* &\leq 2\lambda \left(\tau\|\hat{\beta} - \beta\|_1 + \|\beta\|_1 - \|\hat{\beta}\|_1 \right) + 2\max\left(H(u), G(u)\right) \\ &\leq 2\lambda \left((1+\tau)\sqrt{s}\|u\|_2 - (1-\tau)\sum_{j=s+1}^{p} u_j^{\#} \right) + 2\max\left(F(u), G(u)\right). \end{aligned}$$

By definition of $\delta(\lambda)$,

$$G(u) = \lambda\sqrt{s}\gamma\sqrt{\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda)}}\|Xu\|_n.$$

The rest of the proof will be divided into two parts depending on which of the $G(u)$ and $F(u)$ is larger.

(i) $G(u) > F(u)$

In this case,

$$\|u\|_2 \leq \sqrt{\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}}\|Xu\|_n. \tag{5.19}$$

Hence by using that $2ab \leq a^2 + b^2$,

$$\begin{aligned}\Delta^* &\leq 2\lambda(1+\tau)\sqrt{s}\|u\|_2 + 2G(u)\\ &\leq 2\lambda\sqrt{s}(1+\tau+\gamma)\sqrt{\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}}\|Xu\|_n\\ &\leq \lambda^2 s(1+\tau+\gamma)^2\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))} + \|Xu\|_n^2.\end{aligned} \tag{5.20}$$

(ii) $G(u) \leq F(u)$

In this case,

$$\Delta^* \leq 2\lambda\left((1+\gamma+\tau)\sqrt{s}\|u\|_2 - (1-\gamma-\tau)\sum_{j=s+1}^{p} u_j^{\#}\right) := \Delta.$$

If $\Delta > 0$, then u belongs to the cone $C_{SRE}(s, c_0)$, one can use the $SRE(s, c_0)$ condition, which yields $\|u\|_2 \leq \|Xu\|_n/\theta(s, c_0)$. Therefore, by using $2ab \leq a^2 + b^2$,

$$\Delta^* \leq \Delta \leq 2(1+\gamma+\tau)\lambda\sqrt{s}\frac{\|Xu\|_n}{\theta(s,c_0)} \leq \left((1+\gamma+\tau)\lambda\sqrt{s}\frac{1}{\theta(s,c_0)}\right)^2 + \|Xu\|_n^2. \tag{5.21}$$

If $\Delta \leq 0$, then (5.21) holds trivially since the upper bound is positive.

Combining (5.20) and (5.21) with (5.16) completes the proof of (5.14).

Let now $EY = X\beta^*$ for some $\beta^* \in \mathbb{R}^p$ with $\|\beta^*\|_0 \leq s$. Then (5.14) with $\beta = \beta^*$ implies

$$2\tau\|\hat{\beta} - \beta^*\|_1 \leq C_{\gamma,\tau}(s, \lambda, \delta_0)\lambda s. \tag{5.22}$$

Let use now prove that

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{C_{\gamma,0}(s,\lambda,\delta_0)}{1+\gamma} \frac{\lambda\sqrt{s}}{\theta^2\left(s, \frac{1+\gamma}{1-\gamma}\right)}. \tag{5.23}$$

To prove (5.23), one can take $\tau = 0$, and consider again two cases as above with $u = \hat{\beta} - \beta^*$.

(i) $G(u) > F(u)$

Equations (5.16) and (5.20) with $\tau = 0$ and $\beta = \beta^*$ imply that

$$\|X(\hat{\beta} - \beta^*)\|_n^2 \leq \lambda^2(1+\gamma)^2 \frac{\log(1/\delta_0)}{\log(1/\delta(\lambda))}.$$

Using (5.19), one can obtain

$$\|\hat{\beta} - \beta^*\|_2 \leq (1+\gamma)\lambda\sqrt{s}\left(\frac{\log(1/\delta_0)}{s\log(1/\delta(\lambda))}\right). \tag{5.24}$$

(ii) $G(u) \leq F(u)$

From (5.16) with $\beta = \beta^*$ one can infer that $\Delta^* \geq 0$ almost surely, and thus $\Delta \geq \Delta^* \geq 0$. Hence, $u = \hat{\beta} - \beta^* \in C_{SRE}\left(s, (1+\gamma)/(1-\gamma)\right)$. Thus, the $SRE\left(s, (1+\gamma)/(1-\gamma)\right)$ condition can be applied, which yields

$$\|\hat{\beta} - \beta^*\|_2 \leq \frac{\|Xu\|_n}{\theta\left(s, \frac{1+\gamma}{1-\gamma}\right)} \leq \frac{(1+\gamma)\lambda\sqrt{s}}{\theta^2\left(s, \frac{1+\gamma}{1-\gamma}\right)}, \tag{5.25}$$

where the second inequality is due to the combination of (5.16) and (5.21) with $\beta = \beta^*$ and $\tau = 0$.

Putting together (5.24) and (5.25) proves (5.23). Consequently, it is enough to notice that $\theta^2\left(s, (1+\gamma)/(1-\gamma)\right) \geq \theta^2(s, c_0)$ and then to bound $\|\hat{\beta} - \beta^*\|_q$ from above using (5.22), (5.23) and the norm interpolation inequality $\|\hat{\beta} - \beta^*\|_q \leq \|\hat{\beta} - \beta^*\|_1^{2/q-1} \|\hat{\beta} - \beta^*\|_2^{2-2/q}$. □

The conclusions of Theorem 5.4 hold on event (5.12), which is independent of γ and τ. Hence, on the event 5.12, for all choices of γ, τ, and λ such that $\delta(\lambda) \leq s/(2ep)$, the two inequalities from Theorem 5.4 are satisfied. Let us now present a corollary of the theorem, by setting $\gamma = 1/2$ and $\tau = 1/4$. Hence, the constants in Theorem 5.4 have the form

$$c_0 = 7, \quad (1+\gamma+\tau)^2 = \frac{49}{16}, \quad \frac{1+\gamma}{1-\gamma} = 3, \quad 1+\gamma = \frac{3}{2},$$

while inequality (5.15) can be transformed into

$$\|\hat{\beta} - \beta^*\|_q \leq \frac{49}{8} \max\left[\frac{log(1\delta_0)}{s\log(1/\delta(\lambda))}, \frac{1}{\theta^2(s,7)}\right] \lambda s^{1/q}, \tag{5.16}$$

where we have used the inequalities $\theta^2(s, c_0) \leq \theta^2(s, (1+\gamma)/(1-\gamma))$ and $(2\tau)^{1-2/q} \leq 2$. By taking a closer look at the constant $C_{\gamma,\tau}(s, \lambda, \delta_0)$, one can observe that it is always greater than or equal to $(1+\gamma+\tau)^2/\theta^2(s, c_0)$. Furthermore, the value

$$\delta_0^* = \{\delta(\lambda)\}^{\frac{s}{\theta^2(s,c_0)}} \tag{5.17}$$

is the smallest $\delta_0 \in (0,1)$ such that $C_{\gamma,\tau}(s, \lambda, \delta_0) = (1+\gamma+\tau)^2/\theta^2(s, c_0)$. If $\delta(\lambda) \leq s/(2ep)$, then

$$\delta_0^* \leq \left(\frac{s}{2ep}\right)^{\frac{s}{\theta^2(s,c_0)}}. \tag{5.18}$$

Using these remarks, we obtain the following corollary of Theorems 5.3 and 5.4 with the choice $\gamma = 1/2$, $\tau = 1/4$, and $\delta_0 = \delta_0^*$.

Corollary 5.3. *Let $s \in \{1, \ldots, p\}$. Assume that the SRE(s, 7) condition holds. Let $\hat{\beta}$ be the Lasso estimator with tuning parameter λ satisfying $\delta(\lambda) \leq s/(2ep)$ for $\gamma = 1/2$. Then, with probability at least $1 - \{s/(2ep)\}^{s/\theta^2(s,7)}/2$,*

$$\frac{\lambda}{2}\|\hat{\beta} - \beta\|_1 + \|X\hat{\beta} - EY\|_n^2 \leq \|X\beta - EY\|_n^2 + \frac{49\lambda^2 s}{16\theta^2(s,7)} \tag{5.19}$$

for all $\beta \in \mathbb{R}^p$ such that $\|\beta\|_0 \leq s$. Furthermore, if $EY = X\beta^*$ for some $\beta^* \in \mathbb{R}^p$ with $\|\beta^*\|_0 \leq s$, then for any $1 \leq q \leq 2$,

$$P\left(\|\hat{\beta} - \beta^*\|_q \leq \frac{49\lambda s^{1/q}}{8\theta^2(s,7)}\right) \geq 1 - \frac{1}{2}\left(\frac{s}{2ep}\right)^{\frac{s}{\theta^2(s,7)}}. \tag{5.26}$$

Since $\theta^2(s,7) \leq 1$, the probability in Corollary 5.3 is greater than $1 - \{s/(2ep)\}^s/2$. If the tuning parameter is chosen such that $\delta(\lambda) \leq s/(2ep)$ holds with equality, then $\lambda^2 s$ equals $\sigma^2 s \log(2ep/s)/n$, up to a multiplicative constant. Hence, $\lambda^2 s = O(s/n \log(p/s))$. This represents the minimax rate for the prediction error over the class of all s-sparse vectors $B_0(s) = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq s\}$. The rate $\lambda s^{1/q}$ is minimax optimal for the ℓ_q estimation problem. A more detailed discussion of the minimax rates can be found in [BLT18].

5.4. Lasso on Deoxyfluorination data

This section explains why the estimation error rates from the previous sections do not apply to data with the same structure as the Deoxyfluorination data. Recall the matrix form notation (3.5) from Chapter 3 for a two-way model (extension to three-way is straightforward). An important property of the matrix is that all columns are sparse, which leads to poor behavior of the compatibility constant from Definition 5.1, or the θ term defined in the SRE condition 5.2. To this end, we will provide a complete explanation of why Corollary 5.3 does not yield the optimal rate in this case. The rate in equation (5.26) from Corollary 5.3 is

$$\frac{49\lambda s^{1/q}}{8\theta^2(s,7)}.$$

In order to get the optimal rate $\lambda s^{1/q}$,which leads to $O\left(\sqrt{s/n\log(p/s)}\right)$ for $q = 2$, $\theta^2(s,7)$ should not scale with n. However, for our specific data matrix this is not true. Indeed, by considering $\delta = (1,0,...,0)^T \in \mathbb{R}^{n-1}$ in Definition 5.2, one can observe that $\theta^2(s,7) = O(f_2/n)$, where f_i is the number of levels of the i-th factor, since only $2f_2$ entries from the first column of X are non-zeros. By choosing $f_1 = f_2$, one can obtain $\theta^2(s,7) = O(1/\sqrt{n})$. Hence, the optimal estimation rate does not hold.

6. Simulations

In this chapter, our method, SHIM for three-way interactions, is compared with the standard Lasso across different simulations and, in the end, used on the Deoxyfluorination data. Simulations will be conducted for both a normal response and a response modeled by a continuous Bernoulli distribution, under the assumption that the truth is hierarchical. The following setting will be used for the simulations. Firstly, the tuning parameter λ for Lasso is selected using standard cross-validation. Secondly, tuning parameters of our model, λ_β and λ_γ, will be selected using a modified version of cross-validation that ensures that at least one sample from each possible two-way interaction is present in both the training and validation sets. SHIM will use the parameters estimated by Lasso as initialization and the performance of the models in recovering the unknown parameters will be compared. Further details on model comparison will be provided in the following section. The R^2 scores and the number of times the hierarchy was broken in the first and second layers will also be reported in all the simulations.

6.1. Metrics used to compare different models

Five different metrics will be used to compare models in this chapter. The R^2 score will measure the predictive performance of the models, while four additional metrics will assess the quality of parameter recovery. True positive rates and false positive rates for identifying zeros will measure the model's ability to distinguish between significant and non-significant coefficients. The ℓ_1-norm difference between the true and predicted parameters and the mean squared error will be used to assess how close the recovered parameters are to the true values. The two differences will be scaled by the length of the coefficient vector to obtain results that are independent of its size. Let us introduce the the metrics in a mathematical way:

- R^2 **Score:** The R^2 score measures the predictive performance of the models. It is defined as:
$$R^2 = \rho^2_{y,\hat{y}},$$
where $\rho_{y,\hat{y}}$ is the Pearson correlation, y is the vector of true values, $\hat{y}$ is the vector of predicted values, and $\bar{y}$ is the mean of the true values. A higher R^2 score indicates better prediction performance. An R^2 score of 1 indicates a perfect prediction, while an R^2 score of 0 indicates a prediction performance equal to that of predicting the mean value for all samples.

- **True Positive Rate for Identifying Zeros (TPR-0s):** Let β denote the true vector of coefficients, $\hat{\beta}$ denote the predicted coefficients, and p the length of β. The

TPR for identifying zeros is given by:

$$\text{TPR}(\beta, \hat{\beta}) = \frac{|i : \hat{\beta}_i = \beta_i = 0|}{|i : \beta_i = 0|}.$$

- **False Positive Rate for Identifying Zeros (FPR-0s):** The FPR for identifying zeros is given by:
$$\text{FPR}(\beta, \hat{\beta}) = \frac{|i : \hat{\beta}_i = 0 \ \& \ \beta_i \neq 0|}{|i : \beta_i \neq 0|}.$$
- **ℓ_1-Norm Difference (ℓ_1-dif):** The scaled ℓ_1-norm difference between the true parameter vector, β, and the predicted parameter vector, $\hat{\beta}$, is defined as:
$$\frac{1}{p}\|\beta - \hat{\beta}\|_1 = \frac{1}{p}\sum_{j=1}^{p} |\beta_j - \hat{\beta}_j|.$$
- **Mean squared error (MSE):** The MSE between the true parameter vector, β, and the predicted parameter vector, $\hat{\beta}$, is defined as:
$$\frac{1}{p}\|\beta - \hat{\beta}\|_2^2 = \frac{1}{p}\sum_{j=1}^{p} (\beta_j - \hat{\beta}_j)^2.$$

6.2. SHIM vs Lasso with normal response

6.2.1. Data generation process

A synthetic dataset was created, comprising three factors with 9, 8, and 7 levels, respectively. Out of 24 main effects, 13 were selected as significant. Among the 191 two-way interactions ($9 \times 8 + 9 \times 7 + 8 \times 7 = 191$), 30 were chosen to be significant, and 20 out of the 504 three-way interactions ($9 \times 8 \times 7 = 504$) were considered significant. In the selection of significant coefficients, it was ensured that the true model follows strong hierarchy. The process of assigning values to the significant coefficients was conducted as follows. Recall the matrix form notation described in Chapter 3. Among the coefficients that appear explicitly in the notation, 10 main effects, 14 two-way interactions, and 4 three-way interactions were randomly chosen as significant. Their signs were assigned randomly with equal probability, and their absolute values were drawn from the Uniform distribution, Uniform $[1, 10]$. The rest of the significant coefficients were determined with the identifiability conditions presented in Chapter 3. The response vector, Y, was generated as described in equation (4.1), with noise $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{5I})$.

6.2.2. Results when using cross-validation for selecting the tuning parameters

In our specific case, the data has exactly one replication per each factor level combination. As a result, all the three-way combinations that are present in the training set do not

appear in the validation set, and vice versa. Consequently, cross-validation cannot be used for selecting λ_δ. To ensure comparability with Lasso, λ_δ is chosen such that our model achieves a similar TPR of identifying null three-way interactions as Lasso. As previously mentioned, λ for Lasso and λ_β and λ_γ for SHIM are determined using cross-validation. The obtained values are $\lambda = 0.06$, $\lambda_\beta = 0.12$, and $\lambda_\gamma = 0.6$. In the end, λ_δ is chosen to be 1. These tuning parameter values will be used when training the models on the entire dataset for comparing coefficient estimation.

The comparison of coefficient estimation will be conducted separately for main effects, two-way interactions, and three-way interactions. As shown in Table 6.1, our method outperforms Lasso. The most significant difference in performance between SHIM and Lasso occurs with two-way interactions, where SHIM consistently outperforms Lasso across all criteria. The primary advantage of our method is that it preserves the hierarchy in the estimation, whereas Lasso does not. This results in better overall parameter estimation, as a significant proportion of the null interactions are constrained to remain null. The R^2 scores on the validations sets are similar in both cases, being between 0.8 and 0.82. It should be noted that our method is slower than Lasso, and there is significant potential for improving its computational efficiency.

The main purpose of our analysis is to check whether, our method, SHIM, performs better in parameter estimation than Lasso. However, for the completeness of our analysis, a scatter plot showing SHIM's predictions versus the true values of the responses on one validation set will be shown in Figure 6.1 to confirm the quality of the model's prediction.

Table 6.1.: Results when using cross-validation (normal response)

Model	ℓ_1-dif	MSE	TPR-0s	FPR-0s	Broken hierarchy
Lasso main	0.41	**0.28**	45.45%	0%	-
SHIM main	**0.38**	0.32	**63.64%**	0%	-
Lasso two-way	0.32	0.54	75.15%	6.67%	13
SHIM two-way	**0.26**	**0.43**	**90.1%**	**0%**	0
Lasso three-way	**0.02**	**0.01**	96.28%	30%	17
SHIM three-way	0.03	0.02	**97.93%**	30%	0

6.3. SHIM vs Lasso with continuous Bernoulli response

6.3.1. Data generation process

A synthetic dataset was created, comprising three factors with 9, 8, and 7 levels, respectively. Out of 24 main effects, 13 were selected as significant. Among the 191 two-way interactions ($9 \times 8 + 9 \times 7 + 8 \times 7 = 191$), 30 were chosen to be significant, and 20 out of the 504 three-way interactions ($9 \times 8 \times 7 = 504$) were considered significant. In the selection of significant coefficients, it was ensured that the true model follows strong

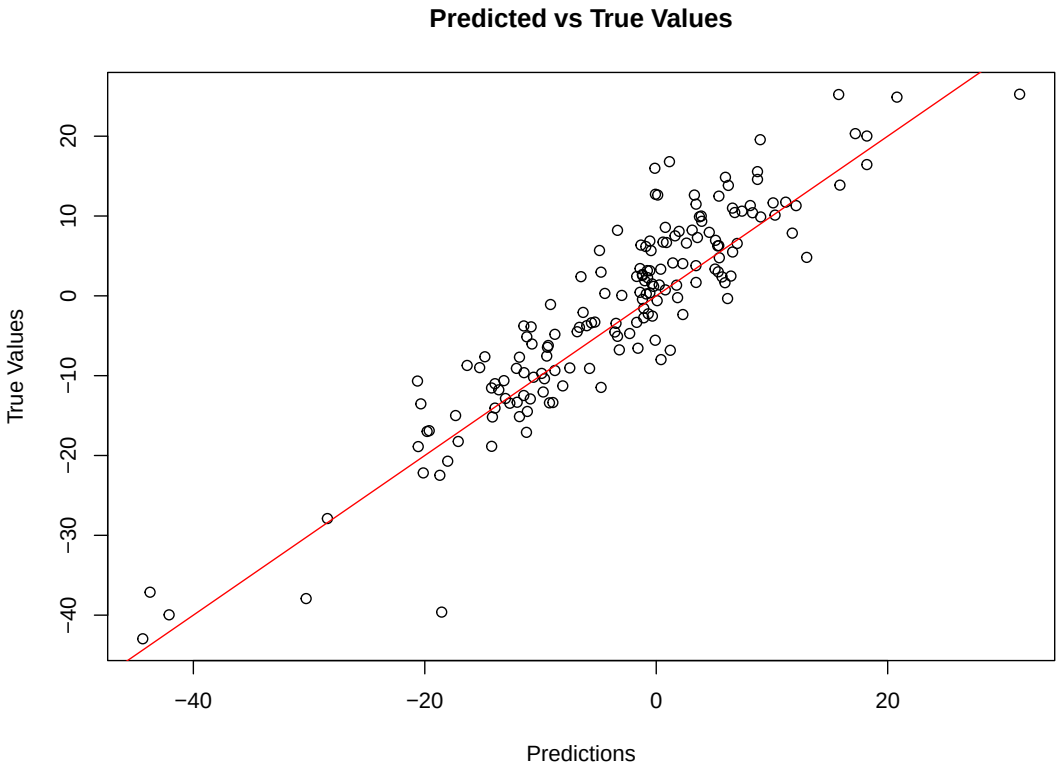

Figure 6.1: Predicted vs true values with normal SHIM on a validation set of simulated data

hierarchy. The process of assigning values to the significant coefficients was conducted as follows. Recall the matrix form notation described in Chapter 3. Among the coefficients that appear explicitly in the notation, 10 main effects, 14 two-way interactions, and 4 three-way interactions were randomly chosen as significant. Their signs were assigned randomly with equal probability. The absolute values of the main effects were drawn from the Uniform distribution, Uniform $[8, 11]$, while the absolute values of the interactions were drawn from Uniform $[5, 8]$. The rest of the significant coefficients were determined with the identifiability conditions presented in Chapter 3. The response Y was generated by sampling from a continuous Bernoulli distribution with η as described in equation (4.9).

6.3.2. Results when using cross-validation for selecting the tuning parameters

As mentioned in the previous section, cross-validation cannot be used to select λ_δ. To ensure comparability with Lasso, λ_δ is chosen such that our model achieves a similar TPR

of identifying null three-way interactions as Lasso. As previously mentioned, λ for Lasso and λ_β and λ_γ for SHIM are determined using cross-validation. The obtained values are $\lambda = 10^{-3}$, $\lambda_\beta = 10^{-3}$, and $\lambda_\gamma = 5 \times 10^{-4}$. In the end, λ_δ is chosen to be 10^{-3}.

Table 6.2.: Results when using cross-validation (continuous Bernoulli response)

Model	ℓ_1-dif	MSE	TPR-0s	FPR-0s	Broken hierarchy
Lasso main	1.01	1.83	45.45%	0%	-
SHIM main	**0.96**	**1.65**	**54.55%**	0%	-
Lasso two-way	0.95	3.85	59%	6.67%	30
SHIM two-way	**0.77**	**3.3**	**75.77%**	**3.33%**	0
Lasso three-way	**0.05**	**0.06**	88.64%	45%	33
SHIM three-way	0.08	0.12	**96.9%**	**20%**	0

The comparison of coefficient estimation will be conducted separately for main effects, two-way interactions, and three-way interactions. As shown in Table 6.2, our method outperforms Lasso. The most significant difference in performance between SHIM and Lasso occurs with main effects and two-way interactions, where our method performs consistently better. In the case of main effects, our method identifies a higher percentage of null coefficients than Lasso, and achieves better parameter estimation. For two-way interactions, our method outperforms Lasso in all criteria. This can be explained by the hierarchical constraints of our method which automatically set to zero a high fraction of the two-way interactions. The performance for the three-way interactions is similar for the both methods. SHIM is better at identifying whether an interaction is null or not, while Lasso achieves better parameter estimation. The average R^2 scores on the validation sets are 0.7 for Lasso and 0.66 for SHIM. It should be noted that, as in the case of a normal response, our method is slower than Lasso.

The main purpose of our analysis is to check whether our method, SHIM, performs better in parameter estimation than Lasso. However, for the completeness of our analysis, a scatter plot showing SHIM's predictions versus the true values of the responses on one validation set is shown in Figure 6.2.

6.4. SHIM and Lasso on Deoxyfluorination data

In this section, normal Lasso, Lasso for the continuous Bernoulli distribution, and SHIM for the continuous Bernoulli distribution will be used on the Deoxyfluorination data. Firstly, our purpose is to show the necessity of using a suitable distribution for modeling the response. Secondly, Lasso and SHIM for the continuous Bernoulli distribution will be compared. In this case, we are not able to compare the quality of the coefficients estimation since the true parameters are unknown. However, if the R^2 scores are similar, one might argue that using a hierarchical model is reasonable since the hierarchical constraints do not affect the prediction.

Figure 6.2: Predicted vs true values with SHIM for continuous Bernoulli distribution on a validation set of simulated data

6.4.1. Normal Lasso

In this subsection, the focus will be on showing that a model suited for a normal response cannot be safely applied to Deoxyfluorination data. To show this, the normal Lasso model is used, with the regularization parameter selected by cross-validation. The model's predictions, using this regularization parameter, are then presented on the entire dataset after fitting it on the whole data.

There are two important issues that can be observed in the model's predictions in Figure 6.3. Firstly, it can be seen that, as expected, some samples are predicted to take values outside the possible range $[0, 1]$. Secondly, the errors in the predictions are not uniformly distributed. The predictions for lower yields (except for those outside the possible range) tend to be higher than the true yields, while for higher yields, the predictions are consistently lower. Similar issues arise when using our hierarchical model for a normal response on the Deoxyfluorination data.

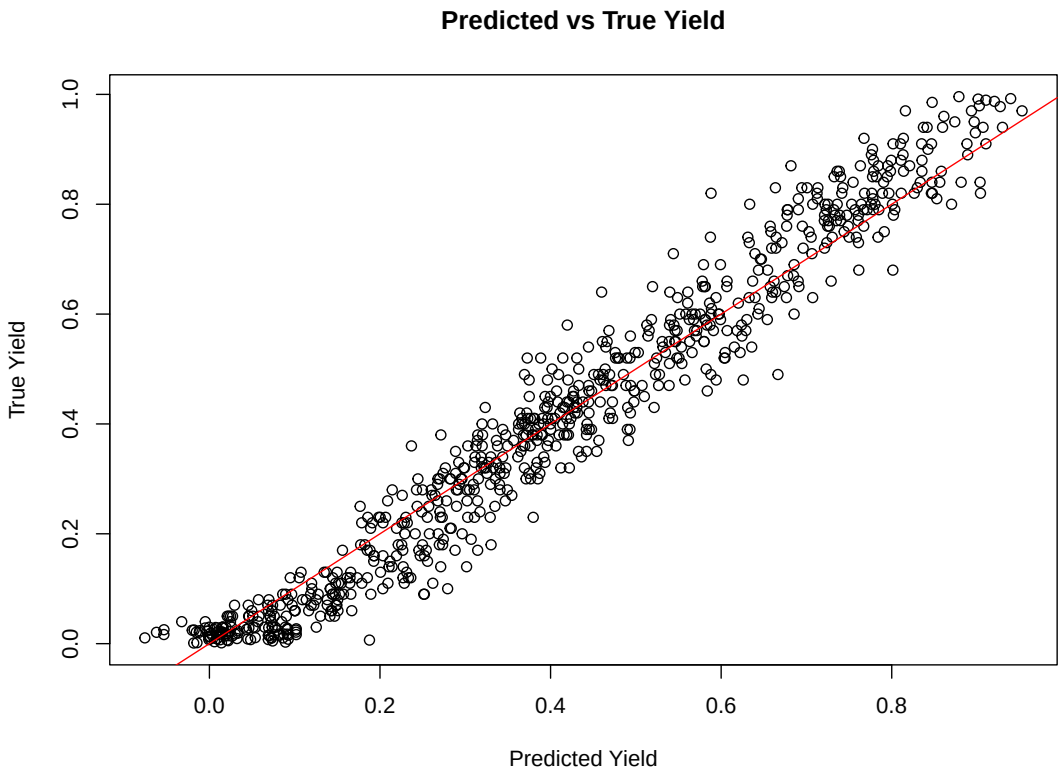

Figure 6.3: Predicted vs true values with normal Lasso on Deoxyfluorination data with entire dataset as training set

6.4.2. Lasso and SHIM for continuous Bernoulli distribution

In this section, the results obtained using Lasso and SHIM for the continuous Bernoulli distribution will be presented. The tuning parameters are obtained through cross-validation, as outlined in the previous simulations. The comparison will focus on the prediction quality of both models. In Figure 6.4 one can observe how the quality of the prediction on the entire data is improved when using continuous Bernoulli distribution to model the response. All the predictions fall within the possible range of values, $[0, 1]$, and the errors are closer to being uniformly distributed compared to those from normal Lasso. Similar observations can be made when analyzing the results obtained using SHIM for the continuous Bernoulli distribution, shown in Figure 6.5.

To prove the prediction quality on data not included in the training set, when using Lasso or SHIM with the response modeled by the continuous Bernoulli distribution, two plots comparing predictions to observed values on a validation set are presented in Figure 6.6. Both models obtain similar R^2 scores on the validation sets, having the average around 0.82. Both SHIM and Lasso preserve all main effects. However, Lasso

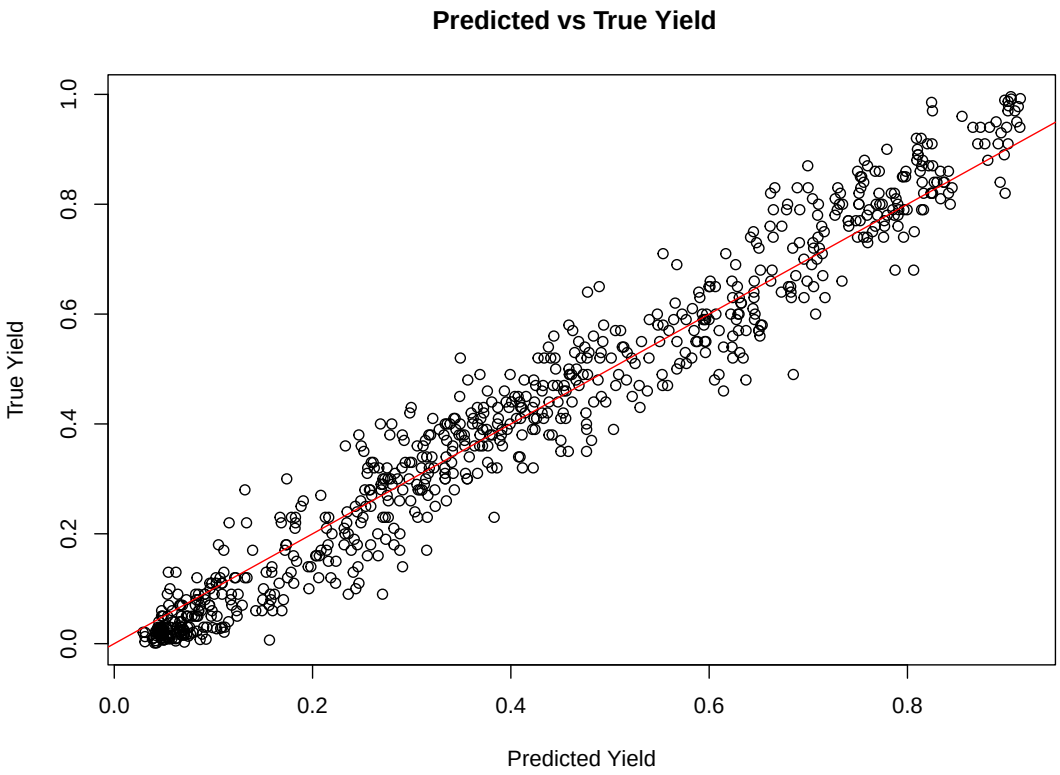

Figure 6.4: Predicted vs true values with Lasso for continuous Bernoulli distribution on Deoxyfluorination data with entire dataset as training set

removes 33% of the two-way interactions, while SHIM eliminates only 17%. On the other hand, SHIM kills few more three-way interactions (95%) due to its hierarchical constraints, compared to Lasso (92%). The exact number of main effects, two-way interactions and three way interactions that are considered significant is shown in Table 6.3 (including the ones obtained using the identifiability conditions), and the number of non-zero parameters of the model (number of non zero coefficients excluding the ones obtained by the identifiability conditions) is shown in Table 6.4.

Table 6.3.: Number of main effects, two-way interactions, and three-way interactions predicted to be non-zero

Model	No. of main effects	No. of two-way interactions	No. of three-way interactions	Total number
Lasso	46/46	238/353	58/740	342/1139
SHIM	46/46	295/353	39/740	380/1139

It is worth noting that SHIM and Lasso do similar parameter estimation. The 10 most important coefficients (highest absolute values) for SHIM are the same as with Lasso, but

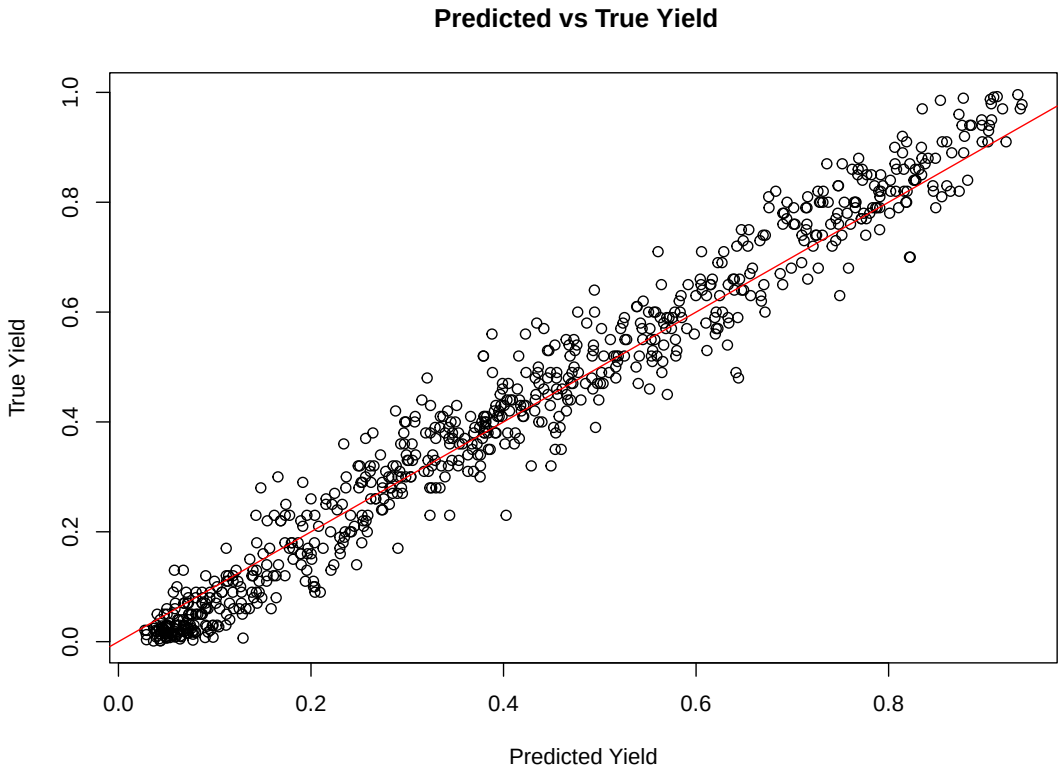

Figure 6.5: Predicted vs true values with SHIM for continuous Bernoulli distribution on Deoxyfluorination data with entire dataset as training set

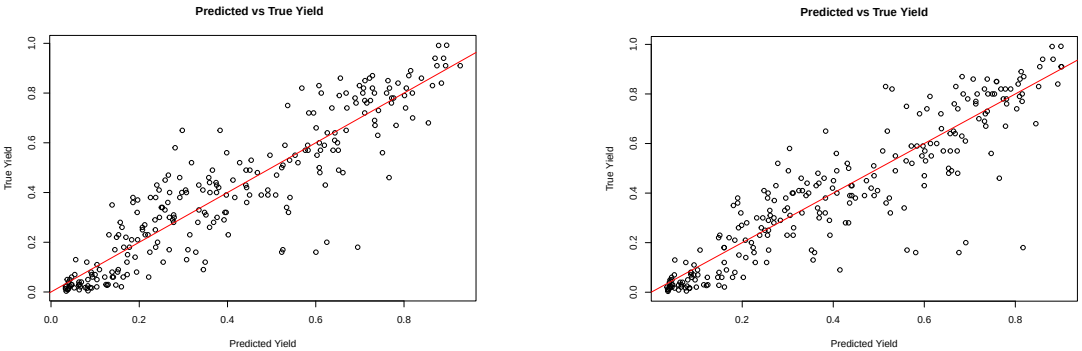

Figure 6.6: Predicted vs true values with SHIM (left) and Lasso (right) for continuous Bernoulli distribution on a validation set of Deoxyfluorination data

in a different order. Additionally, for both models, the signs of the coefficients are the same, when the coefficients are not null. A summary table of the coefficients can be found

Table 6.4.: Number of non-zero coefficients selected by the model

Model	No. of main effects	No. of two-way interactions	No. of three-way interactions	Total number (with intercept)
Lasso	43/43	158/264	12/432	214/740
SHIM	43/43	206/264	8/432	258/740

at the following link. The code that produced all the results presented can be found at the following link.

7. Conclusion

The thesis focused on hierarchical Lasso models. Since most existing models are designed only for two-way interactions and assume a normal response, we decided to address these limitations. Consequently, one of the current hierarchical models was chosen for further extension. We managed to extend the method of [NHCZ10] to handle data with three-way interactions. Additionally, the method was adapted to work for responses modeled by the continuous Bernoulli distribution as well. In the end, it was shown in simulations that our model achieves better parameter estimation than Lasso on hierarchical data with three-way interactions for responses from both the normal distribution and the continuous Bernoulli distribution.

Future work may focus on two directions. The first involves proving the theoretical properties of our model, while the second aims to extend the current methods to accommodate four-way interactions.

Bibliography

[AEL+18a] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle. Predicting reaction performance in c-n cross-coupling using machine learning. *Science*, 360:186, 2018.

[AEL+18b] Daniel T. Ahneman, John G. Estrada, Simon Lin, Steven D. Dreher, and Abigail G. Doyle. Predicting reaction performance in c-n cross-coupling using machine learning. *Science*, 360(6385):186–190, 2018. Erratum in: Science. 2018 Apr 13;360(6385).

[Agr02] Alan Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], New York, second edition, 2002.

[BLT18] Pierre C. Bellec, Guillaume Lecué, and Alexandre B. Tsybakov. Slope meets Lasso: Improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603 – 3642, 2018.

[Bre95] Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[BRT10] Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Hierarchical selection of variables in sparse high-dimensional regression. *IMS Collections*, 6:56–69, 2010.

[BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[DPY+22] Jing Dong, Lichao Peng, Xiaohui Yang, Zelin Zhang, and Puyu Zhang. Xgboost-based intelligence yield prediction and reaction factors analysis of amination reaction. *Journal of Computational Chemistry*, 43(4):289–302, 2022.

[EHJT04] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2), April 2004.

[JB13] ROBERT TIBSHIRANI JACOB BIEN, JONATHAN TAYLOR. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013.

[MN83] P. McCullagh and J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1983.

[NARD18] Matthew K. Nielsen, Derek T. Ahneman, Orestes Riera, and Abigail G. Doyle. Deoxyfluorination with sulfonyl fluorides: Navigating reaction space with machine learning. *Journal of the American Chemical Society*, 140(15):5004–5008, 2018. PMID: 29584953.

[Nel94] J. A. Nelder. The statistics of linear models: back to basics. *Statistics and Computing*, 4(4):221–234, 1994.

[NHCZ10] William Li Nam Hee Choi and Ji Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010.

[NR12] Yuval Nardi and Alessandro Rinaldo. The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli*, 18:945–974, 2012.

[RJ10] Peter Radchenko and Gareth M. James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553, 2010.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[TK23] Boris Maryasin Tatyana Krivobokova, Gianluca Finocchio. Factorizing yields in buchwald-hartwig amination. Manuscript under preparation, 2023.

[vdG16] Sara van de Geer. *Estimation and Testing Under Sparsity*. Lecture Notes in Mathematics. Springer International Publishing, Zurich, Switzerland, 2016.

[YJL07] Ming Yuan, V. Roshan Joseph, and Yi Lin. An efficient variable selection approach for analyzing designed experiments. *Technometrics*, 49:430–439, 2007.

[ZRY09] Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

[D23] Andrzej M. Żurański, Shivaani S. Gandhi, and Abigail G. Doyle. A machine learning approach to model interaction effects: Development and application to alcohol deoxyfluorination. *Journal of the American Chemical Society*, 145(14):7898–7909, 2023. PMID: 36988153.

A. Appendix

A.1. GLM with offset

In this section, it will be shown how the MLE (maximum likelihood estimator) is modified for a GLM with an offset term. Specifically, we will explain how the iteratively reweighted least squares form of the Fisher scoring is adjusted.

A GLM with an offset term is a specific type of GLM where a fixed term is added to the linear predictor.

Let us consider the data $Y_1, \dots, Y_n$ such that $Y_1 \mid X_1, \dots, Y_n \mid X_n$ are independent and $Y_i \mid X_i$ has the probability density function

$$f_{\eta,\psi}(y_i \mid x_i) = \exp\left\{\frac{\eta_i y_i - \kappa(\eta_i)}{\psi_i}\right\} h(y_i, \eta_i), \quad i = 1, \dots, n.$$

For simplicity, and due to the fact that in our case the parameterization of the GLM has the form given in (3.7), the over-dispersion parameter, ψ, and the function h will be ignored. The probability density function becomes

$$f_\eta(y_i \mid x_i) = \exp\left\{\eta_i y_i - \kappa(\eta_i)\right\}, \quad i = 1, \dots, n.$$

It is true that

$$\mu(\eta_i) := \mathbb{E}(Y_i \mid X_i) = \kappa'(\eta_i) \quad \text{and} \quad \operatorname{var}(Y_i \mid X_i) = \kappa''(\eta_i), \quad i = 1, \dots, n.$$

The systematic component changes due to the addition of the offset term and becomes $X_i^T\beta + C_i$, where $C \in \mathbb{R}^n$ is the offset term and C_i is the i-th component.

The link function between the random and systematic component is described by

$$g\left\{\mu(\eta_i)\right\} = X_i^T\beta + C_i, \quad i = 1, \dots, n.$$

Let us consider a canonical link function, $g = \mu^{-1}$, since we use the canonical link for the continuous Bernoulli distribution in our problem. Consequently, the systematic component is $\eta_i = X_i^T\beta + C_i$.

With the setting of a GLM with an offset described, one can write the formula for the log-likelihood in the following way:

$$\ell(\beta) = \sum_{i=1}^{n} \left\{(X_i^T\beta + C_i)Y_i - \kappa(X_i^T\beta + C_i)\right\},$$

which implies

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left\{ Y_i - \kappa'(X_i^T \beta + C_i) \right\} X_i =: S_n(\beta) \tag{A.1}$$

and

$$-\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = \sum_{i=1}^{n} \kappa''(X_i^T \beta + C_i) X_i X_i^T =: F_n(\beta). \tag{A.2}$$

The update rule, in Fisher scoring algorithm is:

$$\hat{\beta}^{(j+1)} = \hat{\beta}^{(j)} + F_n\left(\hat{\beta}^{(j)}\right)^{-1} S_n\left(\hat{\beta}^{(j)}\right).$$

Using (A.1) and (A.2), the update rule becomes

$$\hat{\beta}^{(j+1)} = \hat{\beta}^{(j)} + \left\{ \sum_{i=1}^{n} \kappa''(X_i^T \hat{\beta}^{(j)} + C_i) X_i X_i^T \right\}^{-1} \left[\sum_{i=1}^{n} \left\{ Y_i - \kappa'(X_i^T \hat{\beta}^{(j)} + C_i) \right\} X_i \right].$$

Let us define $W = W(\beta) = \operatorname{diag}\left\{\kappa''(X_i^T \beta + C_i)\right\}$ and the so-called working vector $Z(\beta) = X\beta + W^{-1}\left\{Y - \kappa'(X^T\beta + C)\right\}$. Then, one can represent the Fisher scoring algorithm as a series of weighted least squares estimators. That is, $\hat{\beta}^{(j+1)}$ is a weighted least-squares estimator in the model

$$Z(\hat{\beta}^{(j)}) = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}_n(0, W^{-1}) \tag{A.3}$$

Indeed,

$$\begin{aligned} \hat{\beta}^{(j+1)} &= (X^T W X)^{-1} X^T W Z(\hat{\beta}^{(j)}) \\ &= (X^T W X)^{-1} X^T W \left[X\hat{\beta}^{(j)} + W^{-1}\{Y - \kappa'(X\hat{\beta}^{(j)} + C)\} \right] \\ &= \hat{\beta}^{(j)} + (X^T W X)^{-1} X^T \{Y - \kappa'(X\hat{\beta}^{(j)} + C)\}, \quad j = 0, 1, 2, \ldots. \end{aligned} \tag{A.4}$$

Therefore, Fisher scoring is also an iteratively re-weighted least squares (IRLS) algorithm, and the only modification from the well-known algorithm for a standard GLM with a canonical link function is that the systematic component, η_i, is $X_i^T\beta + C_i$ instead of $X_i^T\beta$.

The modification of using an ℓ_1-norm penalized log-likelihood criterion is the same as in the standard GLM case and involves using iteratively standard LASSO instead of least squares.

A.2. Lemmas for LASSO optimal rate

In this section, the lemmas used for proving Theorem 5.4 are presented.

Firstly, let us introduce the definition of $|\beta|_*$. Let $\lambda = (\lambda_1, \dots, \lambda_p) \in \mathbb{R}^p$ be a vector of tuning parameters, not all equal to 0, such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

For any $\beta = (\beta_1, \dots, \beta_p) \in \mathbb{R}^p$, let $(\beta_1^\#, \dots, \beta_p^\#)$ be a nonincreasing rearrangement of $(|\beta_1|, \dots, |\beta_p|)$ and set

$$|\beta|_* = \sum_{j=1}^{p} \lambda_j \beta_j^\#, \tag{A.5}$$

which defines a norm on $\mathbb{R}^p$.

Lemma A.1. *Let $s \in \{1, \dots, p\}$ and $\tau \in [0, 1]$. For any two $\beta, \hat{\beta} \in \mathbb{R}^p$ such that $\|\beta\|_0 \leq s$, it is true that*

$$\tau|u|_* + |\beta|_* - |\hat{\beta}|_* \leq (1+\tau)\left(\sum_{j=1}^{s} \lambda_j^2\right)^{1/2} \|u\|_2 - (1-\tau)\sum_{j=s+1}^{p} \lambda_j u_j^\#, \tag{A.6}$$

where $u = \hat{\beta} - \beta = (u_1, \dots, u_p)$ and $(u_1^\#, \dots, u_p^\#)$ is a non-increasing rearrangement of $(|u_1|, \dots, |u_p|)$. If $\lambda_1 = \cdots = \lambda_p = \lambda$ for some $\lambda > 0$, then $|\cdot|_* = \lambda|\cdot|_1$ and A.6 yields

$$\tau\lambda\|u\|_1 + \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1 \leq (1+\tau)\lambda\sqrt{s}\|u\|_2 - (1-\tau)\lambda\sum_{j=s+1}^{p} u_j^\#. \tag{A.7}$$

Lemma A.2. *Let $h : \mathbb{R}^p \to \mathbb{R}$ be a convex function, let $f, \varepsilon \in \mathbb{R}^n$, $y = f + \varepsilon$, where ε is normal $\mathcal{N}(0, \sigma^2 I_{n\times n})$, and let X be any $n \times p$ matrix. If $\hat{\beta}$ is a solution of the minimization problem*

$$\min_{\beta\in\mathbb{R}^p} \|X\beta - y\|_n^2 + h(\beta),$$

then $\hat{\beta}$ satisfies, for all $\beta \in \mathbb{R}^p$,

$$\|X\hat{\beta} - f\|_n^2 - \|X\beta - f\|_n^2 \leq \frac{2}{n}\varepsilon^T X(\hat{\beta} - \beta) + h(\beta) - h(\hat{\beta}) - \frac{1}{n}\|X(\hat{\beta} - \beta)\|_n^2. \tag{A.8}$$

The proofs of the lemmas can be found in [BLT18].

A.3. Use of AI Tools for writing the thesis

In this thesis, I used both AI and non-AI tools a few times solely for translation and rephrasing. No additional information was added based on AI tools.