Select the Best Factor Level Combinations in Finite Population Factorial Experiments

by

Lei Shi

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Art

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Jingshen Wang, Chair
Associate Professor Peng Ding
Professor Lexin Li

Spring 2022

Select the Best Factor Level Combinations in Finite Population Factorial Experiments

Abstract

Select the Best Factor Level Combinations in Finite Population Factorial Experiments

by

Lei Shi

Master of Art in Biostatistics

University of California, Berkeley

Assistant Professor Jingshen Wang, Chair

Factorial designs have been widely utilized and studied in many fields. Beyond the purpose of quantifying the factorial effect sizes, many factorial experiments are designed and conducted to seek the most (or the least) effective combinations of factor levels on a relevant outcome of interest subject to certain resource constraint. These demands bring new challenges to this realm, such as the well-recognized "winner's curse" phenomenon, the overly large number of treatment groups or the methodological concerns of how to proceed under resource constraints. To handle these challenges, we propose a general workflow that incorporates several crucial components: (i) forward model selection; (ii) factor level combination selection; (iii) statistical inference over the ties. Theoretically, using a forward selection framework can achieve family-wise error rate control as well as model selection consistency in an asymptotic perspective. We prove finite sample probability bounds to justify the ability of factor level combination selection. Upon the selected results, we justify the inference procedure over the selected ties and report valid confidence intervals for estimation of the effects. These statistical properties are further illustrated by several numerical experiments.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and our contribution

Factorial designs have been widely utilized and studied in many fields, witnessing success in agricultural, industrial, and biomedical applications([1–3]). The power of factorial designs lies in their ability to simultaneously accommodate multiple factors and provide informative assessment for the magnitude of the main causal effects and interactions.

In recent years, beyond the purpose of quantifying the factorial effect sizes, many factorial experiments are designed and conducted to seek the most (or the least) effective combinations of factor levels on a relevant outcome of interest. For example, in cases where the factors represent a set of strategies or policies, decision makers are usually interested in identifying the most promising combination of factor levels that can maximize the utility and produce the highest reward. As another example, when the factors encode a set of characteristics or demographics (race, gender, ethnicity, etc.) for some population, researchers might have peculiar interest in determining which combination of levels is most impacted (positively or adversely) in terms of certain measurement. In general, how to accurately select these extreme groups as well as quantify the effect sizes constitutes the core of these practices.

However, several challenges are hindering the application of heuristic statistical methods. First, due to the well-recognized "winner's curse" phenomenon, a naive peek at the combinations with the largest effect sizes might lead to overly optimistic estimates and a deficiency in the coverage rates [4, 5]. Second, the number of treatment groups in factorial experiments are in general quite large and can increase exponentially in scale as the number of factors grows, which leads to less accurate estimators or confidence intervals and prohibits the implementation of many common analytical strategies such as covariate adjustments. Moreover, beyond these methodological concerns, there are a variety of practical constraints that practitioners might wish to impose. For example, in the strategy combination example presented in the previous paragraph, it is of interest to position ourselves in a scenario where only restricted resources or budgets are available and we are confronted with a maximum limit in the total number of strategies we can apply. How to incorporate such realistic

constraints in the analysis stands as another important problem. Last but not least, when entangled with a finite population discussion, the aforementioned problems are all missing pieces in the puzzle of factorial experiment theories.

In this work we propose a general workflow that targets the issues posited above. The procedure consists of several crucial components: (i) forward model selection; (ii) factor level combination selection; (iii) statistical inference over the ties. The model selection part plays an important role especially when the number of factors considered is large. There is a natural hierarchical structure in factorial experiments. Theoretically speaking, such structure can lend additional information gain if one exploits it cleverly. From a practical perspective, a more parsimonious model leads to more controllability in design and more interpretability in analysis. We show that using a forward selection framework can achieve family-wise error rate control as well as model selection consistency in an asymptotic perspective. The factor level combination selection is indispensable for our initial purpose of identifying most effective groups. We show that such a purpose can be achieved even if the number of factors are increasing. Lastly, we perform statistical inference over the selected ties and report valid confidence intervals for estimation of the effects. Interestingly, as demonstrated by our theoretical insights and simulation results, the forward model selection would lead to a great statistical efficiency gain on the inferential reports for the effect size of the best combinations.

1.2 Literature review

In the realm of factorial experiments, the factor-based regression typically serves as a dominant strategy for delivering point estimators and valid confidence regions, due to its simplicity and flexibility in real-life applications. For example, [6] extended the classical notion of factorial effects to causal counterparts by introducing potential outcome framework and contrast designs. [2] studied the use of both saturated and unsaturated linear models for estimating the factorial causal effects. They discussed the parameter specifications of the regression models and justified the commonly used ordinary least squares (OLS) practice from a theoretical perspective. [7] highlighted the desirable property of regression schemes combined with fractional factorial designs when full designs are possible due to constraints on resources such as units or cost. [8] explores the possibility of incorporating covariate information and applying restricted least squares (RLS) for multiple treatment experimental designs, including factorial studies as a special instance.

A closely related thread of research in factorial designs focuses on variable screening and model selection. Powerful variable selection procedures can significantly reduce the complexity of the working model and lead to aditional benefits in statistical estimation and inference. In practice pre-screening serves as an appealing scheme for optimizing allocation and utilization of resources. To this end, [9] introduced forward regression for main effects screening and proves its screening consistency property. [10] further included second-order interactions into the linear model and proposes a two-step procedure for ultra-high dimensional variable screening. Meanwhile, to save resources and build an interpretable model with high

prediction power, variable selection or screening must be employed. [11] considered convex modelling of the factorial effects estimation and introduces strong heredity condition to achieve adaptive selection. [12] utilized a regularization scheme to tackle the curse of high dimensionality and perform valid variable screening with quadratic regression. Other works including [13, 14], proposed procedures for learning interactions based on ℓ_1 regularized least squares based on a purely algorithmic perspective without statistical guarantee.

1.3 Notations

We summarize the commonly used asymptotic notation. $a_N = O(b_N)$ means that there exists $C > 0$ such that $a_N \leq Cb_N$. $a_N = o(b_N)$ means that $a_N/b_N \to 0$. $a_N = \widetilde{O}(b_N)$ means that there exists $c, C > 0$ such that $cb_N \leq a_N \leq Cb_N$.

To define factorial designs and factorial effects in a general style, we need to apply different level of sets. For an integer K, let $[K] = \{1, \cdots, K\}$. For subsets of $[K]$, we will use a calligraphic font for notation, say $\mathcal{K}$. For collection of such subsets, or subsets of the power set of $[K]$, we would use a blackboard bold font for presentation, say $\mathbb{M} \subset \{\mathcal{K} \mid \mathcal{K} \subset [K]\}$. In particular, we put the power set of $[K]$ as $\mathbb{K}$.

Chapter 2

Factorial experiment setup

We consider a 2^K factorial experimental design for some $K \geq 2$, which encompasses K factors with binary levels indexed by $z_k \in \{0,1\}, k = 1, \cdots, K$. Let $\boldsymbol{z}_{\mathcal{K}} = (z_k)_{k\in\mathcal{K}}$ index the combination of factors in $\mathcal{K} \subset [K]$, with $\boldsymbol{z}_{[K]}$ abbreviated to $\boldsymbol{z}$ specially. These factors define a collection of $Q = 2^K$ treatments, which we denote as $\mathcal{T} = \{\boldsymbol{z} = (z_1 \cdots z_K) \mid z_k \in \{0,1\}, k = 1, \cdots, K\}$. If for some $\boldsymbol{z}$ we have $z_k = 1$, then we call factor z_k is *active*; otherwise z_k is *inactive*. We specially introduce the subsets $\mathcal{T}_{K_0}$ of $\mathcal{T}$, which contains the combinations with at most K_0 factors set as active. N units are enrolled in the experiment, with $N(\boldsymbol{z})$ units in the group with the treatment $\boldsymbol{z}$. For simplicity, we can also index the treatments $\boldsymbol{z}$ in $\mathcal{T}$ using lexicographical order so that we have $\mathcal{T} = \{\boldsymbol{z}_j\}_{j=1}^{Q}$. Let $\overline{Y} = \{\overline{Y}(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}}$ be a vector defined as follows:

$$\overline{Y}(\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^{N} Y_i(\boldsymbol{z}), \boldsymbol{z} \in \mathcal{T}.$$

Under the counterfactual outcome framework or the Neyman-Rubin causal model [15, 16], the i-th unit has potential outcome $Y_i(\boldsymbol{z})$ if assigned to treatment $\boldsymbol{z}$. We aggregate the counterfactual outcomes into vectors $\boldsymbol{Y}_i = \{Y_i(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}}$ using lexicographic order. Let Z_i encode the treatment that the i-th unit received under a random permutation. More concretely, for given $N(\boldsymbol{z})$, we have

$$\mathbb{P}\{Z_i = \boldsymbol{z}_j, i \in [N], j \in [Q]\} = \frac{1}{\binom{N}{N(\boldsymbol{z}_1)}\binom{N-N(\boldsymbol{z}_1)}{N(\boldsymbol{z}_2)}\cdots\binom{N-\sum_{j=1}^{Q-2}N(\boldsymbol{z}_j)}{N(\boldsymbol{z}_{Q-1})}}.$$

The observation for the i-th unit contains only a single realization among these potential outcomes, which we denote as (Y_i, Z_i). We also abbreviate $N(Z_i)$ as N_i to denote the number of units for the treatment group to which the i-th individual is assigned.

We can define factorial effects for any subset $\mathcal{K}$ of the K factors following the discussion of [1, 2, 6]. We introduce a set of vectors $\{g_{\mathcal{K}} \mid g_{\mathcal{K}} \in \mathbb{R}^Q\}$ defined in the following way: for

$|\mathcal{K}| = |\{k\}| = 1$,

$$g_{\mathcal{K}} = \{g_{\mathcal{K}}(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}},\ g_{\mathcal{K}}(\boldsymbol{z}) = \begin{cases} 1, & z_k = 1; \\ -1, & z_k = 0. \end{cases}$$

For $|\mathcal{K}| >= 2$, we have

$$g_{\mathcal{K}} = \{g_{\mathcal{K}}(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}},\ g_{\mathcal{K}}(\boldsymbol{z}) = \prod_{k\in\mathcal{K}} g_{\{k\}}(\boldsymbol{z}).$$

It is convenient to introduce the vector of ones; in other words, we also define:

$$g_{\varnothing} = \mathbf{1}_Q.$$

$\tau_{\varnothing} = \frac{1}{2^K} g_{\varnothing}^{\top}\overline{Y}$ captures the total average of potential outcomes. The main effects and k-way interaction ($k \geq 2$) among factors in $\mathcal{K}$ are denoted by $\tau_{\mathcal{K}}$, which are defined by the inner product of $g_{\mathcal{K}}$ and $\overline{Y}$: $\tau_{\mathcal{K}} = \frac{1}{2^K} g_{\mathcal{K}}^{\top}\overline{Y}$. By introducing an orthonormal matrix $G \in \mathbb{R}^{Q\times Q}$ with columns designated to be these contrast vectors, we can stack these effects into one vector:

$$\tau = (\tau_{\mathcal{K}})_{\mathcal{K}\subset[K]} = \frac{1}{2^K} G^{\top}\overline{Y},\ G = (g_{\mathcal{K}})_{\mathcal{K}\subset[K]}. \tag{2.1}$$

See Example 1 for an elaboration on these definitions in a 2^3 design. For ease of presentation, we call the effect $\tau_{\mathcal{K}}$ a *parent* of $\tau_{\mathcal{K}'}$ if $\mathcal{K} \subset \mathcal{K}'$ and $|\mathcal{K}| = |\mathcal{K}'| - 1$.

Following the presentation of [2], we consider factor-based regression. For the saturated regression, the covariates X_i is a vector indexed by combination of factors (equivalently, subsets of $[K]$). More precisely speaking, X_i is constructed from $Z_i = (z_{i,k})_{k=1}^K$ such that

$$X_{i,\mathcal{K}} = \begin{cases} 1, & \mathcal{K} = \varnothing; \\ \prod_{k\in\mathcal{K}} (2z_{i,k} - 1), & \mathcal{K} \subset [K]. \end{cases}$$

Then *the saturated regression* simply means regressing Y_i on the full X_i. More generally, we denote a collection of combination of factors by $\mathbb{M}$, $\mathbb{M} \subset \mathbb{K} = \{\mathcal{K} \mid \mathcal{K} \subset [K]\}$. The true model is defined as $\mathbb{M}^{\star}$, which contains the indices of the nonzero factorial effects. In particular, we let $\mathbb{K}_d = \{\mathcal{K} \mid |\mathcal{K}| = d\}$ be the collection of indices corresponding to all the k-way interactions. Clearly we have

$$\mathbb{K} = \bigcup_{d=0}^{K} \mathbb{K}_d.$$

For *the unsaturated regression*, we only use a sub-vector of X_i indexed by $\mathcal{K} \in \mathbb{M}$. Denote this sub-vector by $X_{i,\mathbb{M}}$, then the unsaturated regression over $\mathbb{M}$ translates to regressing Y_i on $X_{i,\mathbb{M}}$.

Regarding the results, for regression over $\mathbb{M}$, we use $\widehat{\tau}(\mathbb{M})$ denote the coefficients. Moreover, the population version of the saturated and unsaturated regression are also of particular

interest in our study. We use a non-hat counterpart $\tau(\mathbb{M})$ and τ to denote the population effects over model $\mathbb{M}$. In view of (2.1), we can see the population averages $\overline{Y}$ can be represented by the factorial effects:

$$\overline{Y} = G\tau = G_{\mathbb{M}}\tau(\mathbb{M}) + G_{\mathbb{M}^c}\tau(\mathbb{M}^c). \tag{2.2}$$

Here we use $G_{\mathbb{M}}$ to denote the columns in G indexed by $\mathbb{M}$.

Remark 1. *The above construction of covariates X_i can be generalized by introducing a location-shift scheme as in [2]. For a given centering vector $\delta = (\delta_k)_{k=1}^K$, the X_i can be redefined as*

$$X_{i,\mathcal{K}} = \prod_{k\in\mathcal{K}} (z_{i,k} - \delta_k), \mathcal{K} \subset [K].$$

Example 1 (An explanation in the balanced 2^3 factorial design)**.** *Suppose we have three binary factors, F_1, F_2, F_3, each with level 0 and 1. Combinations of different levels amount to 8 treatment groups, indexed by a triple $(z_1z_2z_3)$ with $z_1, z_2, z_3 \in \{0, 1\}$:*

$$\mathcal{T} = \{(000), (001), (010), (011), (100), (101), (110), (111)\}.$$

There are $N = \sum_{z_1,z_2,z_3} N(z_1z_2z_3) = 2^3N_0$ units from this design, where $N(z_1z_2z_3) = N_0$ denotes the group size under treatment $(z_1z_2z_3)$. Each unit i corresponds to a potential outcome vector $\boldsymbol{Y}_i = \{Y_i(z_1z_2z_3)\}_{z_1,z_2,z_3=0,1}^{\top}$. The parameters of interest are the factorial effects (plus a total average for convenience) $\tau = \left(\tau_\varnothing, \tau_{\{1\}}, \tau_{\{2\}}, \tau_{\{3\}}, \tau_{\{23\}}, \tau_{\{13\}}, \tau_{\{12\}}, \tau_{\{123\}}\right)^{\top}$, which are defined as $\tau = \frac{1}{2^3}G^{\top}\overline{Y}$ through a contrast matrix G.

$$G = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 \end{pmatrix}.$$

Under the Neyman-Rubin model, we observe (Y_i, Z_i) for the unit i, where $Z_i = (z_{i,1}, z_{i,2}, z_{i,3})$. For the purpose of regression, we construct an X_i from Z_i:

$$\begin{aligned} X_i = \Big[&1,\ 2z_{i,1} - 1,\ 2z_{i,2} - 1,\ 2z_{i,3} - 1, \\ &(2z_{i,2} - 1)(2z_{i,3} - 1),\ (2z_{i,1} - 1)(2z_{i,3} - 1), \\ &(2z_{i,1} - 1)(2z_{i,2} - 1),\ (2z_{i,1} - 1)(2z_{i,2} - 1)(2z_{i,3} - 1)\Big]. \end{aligned}$$

Note by our general notation, we see that X_i is defined as a vector indexed by a combination of factors or subset of $[K]$, with $K = 3$ in this case. A saturated regression can then be expressed as

$$Y_i \sim X_i.$$

Sometimes we are also interested in the unsaturated regression. For example, if we only include indices $\varnothing$ (the intercept), $\{1\}, \{12\}, \{13\}, \{123\}$ in our regression, we can form a set of indices $\mathbb{M} = \{\varnothing, \{1\}, \{12\}, \{13\}, \{123\}\}$ and perform

$$Y_i \sim X_{i,\mathbb{M}}, \text{ where } X_{i,\mathbb{M}} = \Big[1,\ 2z_{i,1} - 1,\ \ (2z_{i,1} - 1)\,(2z_{i,2} - 1)\,,\ \ (2z_{i,1} - 1)\,(2z_{i,3} - 1)\,,$$
$$(2z_{i,1} - 1)\,(2z_{i,2} - 1)\,(2z_{i,3} - 1)\Big].$$

Chapter 3

Target parameters and methodology

3.1 Formulation of target parameters

Our study is motivated by the following question: can we identify the most effective factor level combinations in a factorial experiment as well as provide statistical quantification of the effect size under certain constraint? Mathematically speaking, the following set of mean of potential outcomes are of particular interest in our study:

$$\Gamma_{K_0} = \left\{ \overline{Y}(z_1 \cdots z_K) \mid \sum_{k=1}^{K} z_k \leq K_0 \right\}. \tag{3.1}$$

The elements in this set depict the effect size of different treatment levels in the factorial experiments. We mainly care about *the leading population averages* within Γ_{K_0}. The constraint $\sum_{k=1}^{K} z_k \leq K_0$ is practically tempting because it incorporates a type of realistic constraint: at most K_0 factors can be set as active.

Remark 2. *We can consider other types of constraints as substitutes of that in* (3.1)*. For example, we might want to hold certain factors to be always active while some always inactive:*

$$z_k = 1, \text{ for all } k \in \mathcal{K}_{\text{active}}; \; z_k = 0, \text{ for all } k \in \mathcal{K}_{\text{inactive}}.$$

One can tailor the constraints on set of mean potential outcomes to meat different real-life requirements. Here for ease of presentation we focus on a restriction on the maximum number of active factors.

Before formally introducing the target parameters, we first add a careful discussion over the notion of "leading population averages". The subtlety originates from the possible ties inside Γ_{K_0}. For example, it is likely that more than one combination of factor levels reached the highest value among Γ_{K_0}. More generally, we introduce the following definition of tiers, from which we derive the target parameters of interest:

Definition 1 (Tiers and leading population averages)**.** *Assume that* Γ_{K_0} *consists of* H *tiers which satisfy:*

1. Γ_{K_0} *is a disjoint union of the tiers:*

$$\Gamma_{K_0} = \bigcup_{h=1}^{H} \Gamma_{K_0;h}, \text{ where } \Gamma_{K_0;h} \bigcap \Gamma_{K_0;h'} = \varnothing \text{ for } h \neq h' \in [H].$$

2. *The mean of potential outcomes within each tier has the same size:*

$$\overline{Y}(\boldsymbol{z}) \equiv \overline{Y}_{(h)}, \text{ for all } \overline{Y}(\boldsymbol{z}) \in \Gamma_{K_0;h}.$$

3. *Without of loss generality, assume* $\{\overline{Y}_{(h)}\}_{h=1}^{H}$ *are strictly ordered:*

$$\overline{Y}_{(1)} > \overline{Y}_{(2)} > \cdots > \overline{Y}_{(H)}.$$

The largest H_0 *elements in* $\{\overline{Y}_{(h)}\}_{h=1}^{H}$ *are also termed as the* H_0 *leading population averages. Moreover, we denote the set of treatment levels corresponding to the* h*-th tier by* $\mathcal{T}_{K_0;h} = \{\boldsymbol{z} \in \mathcal{T} \mid \overline{Y}(\boldsymbol{z}) \in \Gamma_{K_0;h}\}$*. Then we have*

$$\mathcal{T}_{K_0} = \bigcup_{h=1}^{H} \mathcal{T}_{K_0;h}, \text{ where } \mathcal{T}_{K_0;h} \bigcap \mathcal{T}_{K_0;h'} = \varnothing \text{ for } h \neq h' \in [H].$$

Remark 3. *The definitions of the tiers can be more general. For example, instead of collecting population averages that share the* same *value, we could aggregate combinations that share* similar *values into one tier. For example, we can define the first tier as the set of population averages whose sizes are greater than some pre-specified thresholds* L_1*:*

$$\Gamma_{K_0;1} = \left\{\overline{Y}(\boldsymbol{z}) \in \Gamma_{K_0} \mid \overline{Y}(\boldsymbol{z}) \geq L_1\right\}.$$

And similar definitions can be generalized to $\Gamma_{K_0;h}$*.*

Based on Definition 1, the key question of this paper can be summarized as follows:

> Can we identify the combinations $\cup_{h=1}^{H_0} \mathcal{T}_{K_0;h}$ for the top H_0 tiers and perform statistical inference on the corresponding leading population averages $\{\overline{Y}_{(h)}\}_{h=1}^{H_0}$?

3.2 Challenges and solutions

It is intuitive and straightforward to construct a naive plug-in estimator for the elements in Γ_{K_0} by simply taking group-wise sample average:

$$\widehat{\Gamma}_{K_0;\text{AVG}} = \left\{\widehat{Y}_{\text{AVG}}(z_1 \cdots z_K) \mid \sum_{k=1}^{K} z_k \leq K_0\right\}, \ \widehat{Y}_{\text{AVG}}(z_1 \cdots z_K) = \frac{1}{N(\boldsymbol{z})} \sum_{i:Z_i=\boldsymbol{z}} Y_i.$$

Then it seems that we are approaching our goal by carefully manipulating $\widehat{\Gamma}_{K_0;\mathrm{AVG}}$. However, such a naive practice is confronted with many subtleties both in theory and in practice. The first challenge is that, the number of treatment groups in factorial experiments grows exponentially in K. This could lead to the less favorable situation where some of the target treatment groups suffer from an insufficiency in the number of assigned units. That being said, these $\widehat{Y}_{\mathrm{AVG}}(\boldsymbol{z})$ would be lacking in statistical accuracy. Secondly, such a practice tends to ignore the connection between different treatment groups and discard the unique structure of factorial experiments. One obvious critic is that: the units from many treatment groups would not be able to play any role in constructing $\widehat{\Gamma}_{K_0;\mathrm{AVG}}$. By calculation we can see the number of treatment groups being ignored is

$$2^K - \sum_{l=1}^{K_0} \binom{K}{l}.$$

A large volume of data would be redundant due to the neglect of high level connections between factorial treatment groups.

The key idea to circumvent these issues is by taking advantage of the structural benefits of factorial effects. [1] described several important principles when designing and analyzing factorial experiments:

- *Effect Hierarchy Principle.* (i) Lower-order effects are more likely to be important than higher-order effects. (ii) Effects of the same order are equally likely to be important.
- *Effect Sparsity Principle.* The number of relatively important effects in a factorial experiment is small.
- *Effect Heredity Principle.* In order for an interaction to be significant, at least one of its parent main effects should be significant.

The hierarchy and sparsity principle are natural and consistent among the literature. For the heredity principle, there are some generalization that has received more attention. Based on different levels of belief on the factorial structure, the following versions of heredity are commonly used:

- *Weak heredity.* If a higher order interaction effect is nonzero, at least one of its parent lower order interaction effect is nonzero.
- *Strong heredity.* If a higher order interaction effect is nonzero, all of its parent lower order interaction effects are nonzero.

In light of these principles, although it is not likely to assume any special structure for the Γ_{K_0}, a linear transformation of these mean of potential outcomes into factorial effects might have a nice hierarchical and sparse structure. This naturally inspires us to reparametrize the elements in Γ_{K_0} using factorial effects:

$$\overline{Y} = G\tau. \tag{3.2}$$

Coordinate-wisely we have

$$\overline{Y}(z_1 \cdots z_K) = \sum_{\mathcal{K}:\mathcal{K}\subset[K]} \tau_{\mathcal{K}} \prod_{k\in\mathcal{K}} (2z_k - 1)\,. \tag{3.3}$$

Although the dimension of τ is the same as $\overline{Y}$, in practice we might have evidence that the effect size of high-order interactions are negligible. That is, a plausible approximation can be obtained using only the significant lower-order effects (say $\tau_{\mathcal{K}}$ with $|\mathcal{K}| \le D$) in (3.1):

$$\overline{Y}(z_1 \cdots z_K) \approx \sum_{\mathcal{K}:|\mathcal{K}|\le D} \tau_{\mathcal{K}} \prod_{k\in\mathcal{K}} (2z_k - 1)\,.$$

What remains now is to find accurate point estimates $\widehat{\tau}_{\mathcal{K}}$ for $|\mathcal{K}| \le D$ and report a set of reparametrization-based estimates:

$$\widehat{\Gamma}_{K_0;\mathrm{RP}} = \left\{ \widehat{Y}_{\mathrm{RP}}(z_1 \cdots z_K) \mid \sum_{k=1}^{K} z_k \le K_0 \right\},\ \widehat{Y}_{\mathrm{RP}}(z_1 \cdots z_K) = \sum_{\mathcal{K}:|\mathcal{K}|\le D} \widehat{\tau}_{\mathcal{K}} \prod_{k\in\mathcal{K}} (2z_k - 1)\,.$$

Here $\widehat{\tau}_{\mathcal{K}}$ can be obtained from a weighted unsaturated regression [2]:

$$Y_i \sim X_{i,\mathbb{M}},\ \text{with weights } w_i = \frac{N}{N_i},\ \text{and } \mathbb{M} = \{\mathcal{K} \subset [K] \mid |\mathcal{K}| \le D\}. \tag{3.4}$$

An important fact pointed out by [2] is that the obtained $\widehat{\tau}_{\mathcal{K}}$ by running (3.4) are unbiased estimates for the true effects.

How well can $\widehat{\Gamma}_{K_0;\mathrm{RP}}$ approximate the truth? Can we witness additional statistical gain when using $\widehat{\Gamma}_{K_0;\mathrm{RP}}$ compared with $\widehat{\Gamma}_{K_0;\mathrm{AVG}}$? Intuitively speaking, if the high-order interaction signals are weak, $\widehat{Y}_{\mathrm{RP}}(\boldsymbol{z})$ will be able to recover the truth accurately. Moreover, if there are some heredity structure in the lower-order interactions, we would expect a further improvement in the estimation since we are able to obtain better estimates for $\widehat{\tau}_{\mathcal{K}}$. In Section 4 we will translate these intuitions into rigorous mathematical formulations and add more detailed theoretical discussion.

3.3 Identify the best factor level combinations

We propose to incorporate such heredity condition into a selection procedure, then perform statistical inference on the selected model. The first algorithm presents a forward model selection procedure:

Remark 4. *We add some comments regarding the forward model selection algorithm (Algorithm 1):*

Algorithm 1: Forward model selection under heredity

Input: Factorial data (Y_i, Z_i); predetermined integer D; initial model for factorial effects $\widehat{\mathbb{M}} = \{\varnothing\}$; significance level $\alpha = \sum_{d=1}^{D} \alpha_d$.

Output: Selected model $\widehat{\mathbb{M}}$.

Define an intermediate model $\widehat{\mathbb{M}}' = \widehat{\mathbb{M}}$ for convenience.

for $d = 1, \cdots, D$ **do**

  Update intermediate model to include all the d-order terms: $\widehat{\mathbb{M}}' = \widehat{\mathbb{M}} \cup \mathbb{K}_d$.

  Prune interactions according to the heredity principle and still denote the pruned working model as $\widehat{\mathbb{M}}'$. When $d \geq 2$,

  if *under weak heredity* **then**

    remove all the d-way interaction term indexed by $\mathcal{K}$ from $\widehat{\mathbb{M}}'$ if

$$\mathcal{K}' \notin \widehat{\mathbb{M}}' \text{ for all } \mathcal{K}' \subset \mathcal{K},\ |\mathcal{K}'| = |\mathcal{K}| - 1.$$

  else if *under strong heredity* **then**

    remove all the d-way interaction term indexed by $\mathcal{K}$ from $\widehat{\mathbb{M}}'$ if

$$\mathcal{K}' \notin \widehat{\mathbb{M}}' \text{ for some } \mathcal{K}' \subset \mathcal{K},\ |\mathcal{K}'| = |\mathcal{K}| - 1.$$

  Run weighted least squares on the model $\widehat{\mathbb{M}}'$:

$$Y_i \sim X_{i,\widehat{\mathbb{M}}'},\ \text{with weights } w_i = N/N_i.$$

  Obtain coefficients $\widehat{\tau}(\widehat{\mathbb{M}}')$ and robust covariance estimation $\widehat{\mathbf{\Sigma}}(\widehat{\mathbb{M}}')$:

$$\widehat{\mathbf{\Sigma}}(\widehat{\mathbb{M}}') = \frac{1}{Q^2} G_{\widehat{\mathbb{M}}'}^{\top} \mathbf{Diag}\left\{ N(\boldsymbol{z})^{-1} \widehat{S}(\boldsymbol{z}, \boldsymbol{z}) \right\} G_{\widehat{\mathbb{M}}'}.$$

  Extract $\widehat{\tau}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ and $\widehat{\sigma}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ for all $\mathcal{K} \in \widehat{\mathbb{M}}'$ with $|\mathcal{K}| = d$.

  Run marginal t-test using the above $\widehat{\tau}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ and $\widehat{\sigma}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ under significance level $\alpha_d/(|\widehat{\mathbb{M}}'| - |\widehat{\mathbb{M}}|)$ and remove the non-significant terms from $\widehat{\mathbb{M}}' \backslash \widehat{\mathbb{M}}$.

  Set $\widehat{\mathbb{M}} = \widehat{\mathbb{M}}'$.

return $\widehat{\mathbb{M}}$

1. *In Step 3 we make use of the heredity principles to prune the model before running a formal selection procedure. Under weak heredity, we remove $\mathcal{K}'$ from $\widehat{\mathbb{M}}'$ if* all *of its parent effects are zero (not selected by $\mathbb{M}$ from last loop). Under strong heredity, we remove $\mathcal{K}'$ from $\widehat{\mathbb{M}}'$ if* some *of its parent effects are zero (not selected by $\mathbb{M}$ from last loop);*

2. *In Step 9 - 11 we apply a Bonferroni corrected marginal t-test for model selection. Generally speaking, these steps can be replaced by many popular model selection procedures. In Section 4 we establish several conditions on the layer-wise selection algorithms that can guarantee an overall model selection consistency property. Empirically many popular algorithms (such as ℓ_1 regularized least squares) work satisfactorily too, as verified by our simulation study in Section 5.*

The second algorithm targets the best factor level combinations and report confidence intervals for the selected leading population averages.

Remark 5. *We add some brief comments on the details of the algorithm:*

1. *Algorithm 2 involves two thresholding parameters b_L and b_R. In Section 4 we show that it suffices to require b_L, b_R to be smaller the gap between the tiers to achieve consistent selection. Empirically we will propose a tuning procedure to choose the most appropriate b_L, b_R from a data-driven perspective.*

2. *Alternatively, in Step 3.3 - 3.4, we can use smoothed bootstrap to report confidence intervals: we generate B bootstrap samples: $\xi_b \sim N\left(G_{\widehat{\mathbb{M}}}\widehat{\tau}, G_{\widehat{\mathbb{M}}}\widehat{\mathbf{\Sigma}}G_{\widehat{\mathbb{M}}}^\top\right) \mid (\widehat{\tau}, \widehat{\mathbf{\Sigma}})$, $b = 1, \cdots, B$. Compute the bootstrap average:*

$$\widehat{\xi}_{b;h} = \frac{1}{|\widehat{\mathcal{T}}_{K_0;h}|} \sum_{\boldsymbol{z} \in \widehat{\mathcal{T}}_{K_0;h}} \xi_b(\boldsymbol{z}).$$

   Report quantile confidence intervals from the empirical distribution of $\widehat{\xi}_{b;h}$:

$$[\widehat{q}_{1-\alpha'/2}, \widehat{q}_{\alpha'/2}].$$

3. *If one input $\widehat{\mathbb{M}} = \mathbb{K}$, then using $\widehat{Y}_{\mathrm{RP}}$ in Algorithm 2 is equivalent to $\widehat{Y}_{\mathrm{AVG}}$.*

Algorithm 2: Factor level combination selection and statistical inference

Input: Selected model $\widehat{\mathbb{M}}$; pre-specified positive integers K_0, H_0; significance level α'; thresholds b_L, b_R; initial set $\widehat{\mathcal{T}}_{K_0;0} = \varnothing$.

Output: For $h \in [H_0]$, report selected factor level combinations $\widehat{\mathcal{T}}_{K_0;h}$, estimated leading population averages $\widehat{Y}_{(h)}$ and confidence intervals.

1 Run weighted least squares to obtain coefficients $\widehat{\tau} = \widehat{\tau}(\widehat{\mathbb{M}})$ and robust covariance estimation $\widehat{\mathbf{\Sigma}} = \widehat{\mathbf{\Sigma}}(\widehat{\mathbb{M}})$.

2 Obtain reparametrization-based estimates and restricted subset $\widehat{\Gamma}_{K_0}$:

$$\widehat{Y}_{\mathrm{RP}} = G_{\widehat{\mathbb{M}}}\widehat{\tau},\ \widehat{\Gamma}_{K_0} = \left\{ \widehat{Y}_{\mathrm{RP}}(z_1 \cdots z_K) \mid \sum_{k=1}^{K} z_k \leq K_0 \right\}.$$

3 Select the best factor level combinations. Initialize $\widehat{\mathcal{T}}_{\mathrm{Selected}} = \varnothing$ to record the selected combinations. **for** $d = 1, \cdots, D$ **do**

4 Locate the factor level combinations that belong to the h-th tier. Let $\widehat{\mathcal{T}}_{\mathrm{Selected}} = \widehat{\mathcal{T}}_{\mathrm{Selected}} \cup \widehat{\mathcal{T}}_{K_0;h-1}$. Estimate the combinations $\widehat{\mathcal{T}}_{K_0;h}$ corresponding to the h-th tier:

$$\widehat{\mathcal{T}}_{K_0;h} = \left\{ \boldsymbol{z} \in \mathcal{T}_{K_0} \mid -b_L \leq \widehat{Y}_{\mathrm{RP}}(\boldsymbol{z}) - \max_{\boldsymbol{z}' \in \mathcal{T}_{K_0} \setminus \widehat{\mathcal{T}}_{\mathrm{Selected}}} \widehat{Y}_{\mathrm{RP}}(\boldsymbol{z}') \leq b_R \right\}.$$

5 Generate point estimates for the effect size over the h-th tier by taking average:

$$\widehat{Y}_{(h)} = \frac{1}{|\widehat{\mathcal{T}}_{K_0;h}|} \sum_{\boldsymbol{z} \in \widehat{\mathcal{T}}_{K_0;h}} \widehat{Y}_{\mathrm{RP}}(\boldsymbol{z}) = \frac{1}{|\widehat{\mathcal{T}}_{K_0;h}|} \mathbf{1}_{\widehat{\mathcal{T}}_{K_0;h}}^{\top} \widehat{Y}_{\mathrm{RP}}.$$

Here $\mathbf{1}_{\widehat{\mathcal{T}}_{K_0;h}}^{\top}$ is a vector with 1 on the positions in $\widehat{\mathcal{T}}_{K_0;h}$ and 0 otherwise.

6 Obtain the variance of the estimate.

$$\widehat{s}_{(h)}^2 = \frac{1}{|\widehat{\mathcal{T}}_{K_0;h}|^2} \mathbf{1}_{\widehat{\mathcal{T}}_{K_0;h}}^{\top} G_{\widehat{\mathbb{M}}} \widehat{\mathbf{\Sigma}} G_{\widehat{\mathbb{M}}}^{\top} \mathbf{1}_{\widehat{\mathcal{T}}_{K_0;h}}.$$

7 Generate confidence intervals for the effect size over the h-th tier. Report:

$$\left[\widehat{Y}_{(h)} - z_{\alpha'/2}\widehat{s}_{(h)}, \widehat{Y}_{(h)} + z_{\alpha'/2}\widehat{s}_{(h)} \right].$$

Chapter 4

Theoretical insights

Before presenting the main theorems, we first state the conditions on the factorial designs and mathematically formalize the effect sparsity principle and effect heredity principle into two sets of assumptions.

Assumption 1 (Conditions on the factorial design)**.** *There exists universal constants* $\underline{c}, \overline{c}, \gamma > 0$*, such that*

- $N(\boldsymbol{z}) = c(\boldsymbol{z})N_0$*, where* $0 < \underline{c} \leq c(\boldsymbol{z}) \leq \overline{c}$ *for all* $\boldsymbol{z} \in \mathcal{T}$*.*
- $S(\boldsymbol{z}, \boldsymbol{z}) \leq \overline{s}$ *for all* $\boldsymbol{z} \in \mathcal{T}$*. That is, the potential outcomes have bounded second order moments.*

Assumption 2 (Sparse high-order interactions)**.** *There is an integer* $D \in [K]$*, such that all the* k*-way interactions with* $k \geq D + 1$ *are all zero.*

For effect heredity principle, we generalize the discussion in [1] and study two types of heredity [10, 13]:

Assumption 3 (Effect heredity)**.** *The following assumptions are used to model the heredity structure in the factorial effects:*

1. *Weak heredity: if* $\tau_{\mathcal{K}} \neq 0$*, then there exists some* $\mathcal{K}' \subset \mathcal{K}$ *with* $|\mathcal{K}'| = |\mathcal{K}| - 1$*, such that* $\tau_{\mathcal{K}'} \neq 0$*.*
2. *Strong heredity: if* $\tau_{\mathcal{K}} \neq 0$*, then for all* $\mathcal{K}' \subset \mathcal{K}$ *with* $|\mathcal{K}'| = |\mathcal{K}| - 1$*,* $\tau_{\mathcal{K}'} \neq 0$*.*

4.1 Probabilistic analysis of finite population estimates

Before presenting the formal theoretical analysis, we gather some useful results regarding the finite population estimates. We introduce some useful notations and facts. Let $V =$

$\mathrm{Var}(\widehat{Y}_{\mathrm{AVG}})$. Then

$$V = \mathbf{Diag}\left\{N(\boldsymbol{z})^{-1}S(\boldsymbol{z},\boldsymbol{z})\right\}_{\boldsymbol{z}\in\mathcal{T}} - N^{-1}S.$$

A robust variance estimator $\widehat{V}_R$ is given by

$$\widehat{V}_R = \mathbf{Diag}\left\{N(\boldsymbol{z})^{-1}\widehat{S}(\boldsymbol{z},\boldsymbol{z})\right\}_{\boldsymbol{z}\in\mathcal{T}}, \text{ with } V_R = \mathbb{E}\{\widehat{V}_R\} = \mathbf{Diag}\left\{N(\boldsymbol{z})^{-1}S(\boldsymbol{z},\boldsymbol{z})\right\}_{\boldsymbol{z}\in\mathcal{T}}.$$

The first result comes from [2, 17], which summarizes the asymptotic property of the factorial effect estimates and the robust variance estimations.

Lemma 1 (Asymptotic results)**.** *Let Q be fixed. Assume that, as N_0 goes to infinity, for all $\boldsymbol{z}\in\mathcal{T}$,*

1. *$N(\boldsymbol{z}) = c(\boldsymbol{z})N_0 \geq 2$ and $c(\boldsymbol{z})$ has a limit $c_{\lim}(\boldsymbol{z})$ between $\underline{c}$ and $\overline{c}$;*
2. *$\overline{Y}$ and S have finite limits $\overline{Y}_{\lim}$ and $S_{\lim}$;*
3. *$\max_{1\leq i\leq N}\{Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\}^2/N_0 \to 0$.*

Then the following results hold:

1. *By Condition 1 and 2 we know that there is a $V_{\lim}$ such that*

$$N_0 V \to V_{\lim}.$$

The treatment-wise averages $\widehat{Y}_{\mathrm{AVG}}$ are asymptotically normal:

$$\sqrt{N_0}\{\widehat{Y}_{\mathrm{AVG}} - \overline{Y}\} \xrightarrow{d} N(0, V_{\lim}).$$

2. *The variance estimation is robust and converges to its expectation in probability. That is,*

$$N_0\widehat{V}_R - N_0 V_R \xrightarrow{p} 0, \text{ where } V_R = \mathbb{E}\{\widehat{V}_R\} \succeq V.$$

The proof is omitted here.

Lemma 2 (Berry-Esseen bounds in finite population)**.** *Let $\breve{Y}_i(\boldsymbol{z}) = Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})$ and $\breve{\gamma}_i = \frac{1}{N}\sum_{\boldsymbol{z}\in\mathcal{T}} w(\boldsymbol{z})\breve{Y}_i(\boldsymbol{z})$. Consider an estimator of the form*

$$\widehat{\gamma}_N = \sum_{\boldsymbol{z}\in\mathcal{T}} w(\boldsymbol{z})\widehat{Y}_{\mathrm{AVG}}(\boldsymbol{z}),$$

with robust variance estimation

$$\widehat{v}_R^2 = \sum_{\boldsymbol{z}\in\mathcal{T}} w(\boldsymbol{z})^2 N(\boldsymbol{z})^{-1}\widehat{S}(\boldsymbol{z},\boldsymbol{z}).$$

Denote $\gamma_N = \mathbb{E}\{\widehat{\gamma}_N\}$, $v_N^2 = \mathrm{Var}(\widehat{\gamma}_N)$ and $v_R^2 = \mathbb{E}\{\widehat{v}_R^2\}$. Assume that $v_N/v_R \geq \kappa$ for some $\kappa\in(0,1]$. Then we have a Berry-Esseen bound with the true variance:

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}_N - \gamma_N}{v_N} \leq t\right\} - \Phi(t)\right| \leq \frac{C(\underline{c},\overline{c},\kappa)\|w\|_\infty}{\|w\|_2\sqrt{N_0}} \cdot \frac{\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\min_{\boldsymbol{z}\in\mathcal{T}}\sqrt{S(\boldsymbol{z},\boldsymbol{z})}}.$$

4.2 Model selection consistency

We are now ready to show the model selection guarantee for our procedure. The proof proceeds through mathematical induction:

1. (Base case) Show that Algorithm 1 selects the correct main effects with probability tending to one.
2. (Induction step) Show that if Algorithm 1 correctly specifies the model up to k-way interactions(main effects if $k = 1$), then it will correctly detect the non-nulls among all $(k + 1)$-way interactions.

Recall that the validity of a test requires that given the null is true, the probability of rejection should be less than the pre-specified significance level. The consistency of a test, on the other hand, requires that when the alternative is true, we have probability tending to 1 to reject the null. Following the tradition in the realm of model selection [18], we introduce the following definitions:

Definition 2 (Terminologies for model selection)**.** *The following terminologies are wrapped up for the discussion of model selection:*

- *Type I error: commit a false positive. Since the selected positives are contained in* $\widehat{\mathbb{M}}$*, this translates into* $\widehat{\mathbb{M}} \cap \mathbb{M}^{\star c} \neq \varnothing$*.*
- *Type II error: commit a false negative. Since the selected negatives are contained in* $\widehat{\mathbb{M}}$*, this translates into* $\widehat{\mathbb{M}}^c \cap \mathbb{M}^{\star} \neq \varnothing$*.*
- *Size: we can define the size as* $q(\widehat{\mathbb{M}}) = \mathbb{P}(\widehat{\mathbb{M}} \cap \mathbb{M}^{\star c} \neq \varnothing)$ *and the asymptotic size* $\limsup_{n\to\infty} q(\widehat{\mathbb{M}})$*.*
- *Power: the power is defined as* $\pi(\widehat{\mathbb{M}}) = \mathbb{P}(\mathbb{M} \subset \widehat{\mathbb{M}})$*.*

Clearly size measures the ability of under-selection; since zero $q(\widehat{\mathbb{M}})$ means $\widehat{\mathbb{M}} \subset \mathbb{M}^{\star}$. Power measures the ability of over-selection, since a power of 1 means $\mathbb{M}^{\star} \subset \widehat{\mathbb{M}}$. We have the following graph to visualize the idea:

For convenience, we introduce some additional notation. Let $\mathbb{M}^{\star}_d$ be the true model on the d-way interactions. Then

$$\mathbb{M}^{\star} = \mathbb{M}^{\star}_1 \cup \cdots \cup \mathbb{M}^{\star}_D.$$

Analogously we can define the sample version of these concepts. That is, let

$$\widehat{\mathbb{M}} = \widehat{\mathbb{M}}_1 \cup \cdots \cup \widehat{\mathbb{M}}_D.$$

One important part in understanding Algorithm 1 is that it introduces two operators to advance the selected models to the next layer. Concretely speaking, Step 3 - Step 7

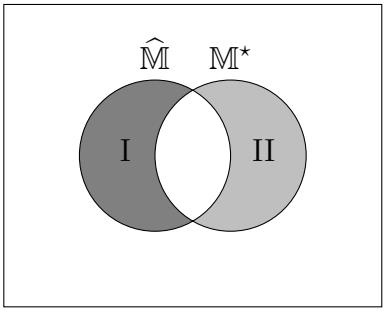

Figure 4.1: Visualization of the two types of errors

introduces a deterministic operator $\mathtt{P}(\cdot)$ on a given model $\mathbb{M}$, while Step 9 - Step 11 introduces a stochastic (or data dependent) operator $\mathtt{S}_N(\cdot) = \mathtt{S}_N(\cdot; \{Y_i, X_i\}_{i=1}^N)$ on any given model $\mathbb{M}$. We can track the models in the following diagram:

$$\widehat{\mathbb{M}}_1 \xrightarrow{\mathtt{P}} \cdots \xrightarrow{\mathtt{S}_N} \widehat{\mathbb{M}}_{d-1} \xrightarrow{\mathtt{P}} \widehat{\mathbb{M}}_{d,+} \xrightarrow{\mathtt{S}_N} \widehat{\mathbb{M}}_d \to \cdots \xrightarrow{\mathtt{S}_N} \widehat{\mathbb{M}}_D. \tag{4.1}$$

Intuitively speaking, in order to achieve satisfactory model selection results, some regularization conditions need to be imposed to characterize a "good" layer-wise selection method. A key question is that: if we are able to select the $(d-1)$-th layer correctly, can the selection procedure forwards the good results to the d-th layer? In light of this, we use $\mathbb{M}_{d,+}^\star$ to denote the pruned set of effects on the d-th layer based on the true model $\mathbb{M}_{d-1}^\star$ on the previous layer; that is,

$$\mathbb{M}_{d,+}^\star = \mathtt{P}(\mathbb{M}_{d-1}^\star).$$

These discussions motivate the following assumption on the layer-wise selection procedure $\mathtt{S}_N(\cdot)$:

Assumption 4 (Validity and consistency of the selection operator)**.** *We denote*

$$\widetilde{\mathbb{M}}_d = \mathtt{S}_N(\mathbb{M}_{d,+}^\star; \{Y_i, X_i\}_{i=1}^N),$$

where $\mathbb{M}_{d,+}^\star = \mathtt{P}(\mathbb{M}_{d-1}^\star)$ *is defined as above. Let* $\{\alpha_d\}_{d=1}^D$ *be a sequence of significance levels in* $(0,1)$*. We assume that the following* validity *and* consistency *property hold for* $\mathtt{S}_N(\cdot)$*: for* $d = 1, \cdots, D$*, we have*

$$\textit{Validity: } \limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} \neq \varnothing\right\} \leq \alpha_d, \tag{4.2}$$

$$\textit{Consistency: } \limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\} = 0. \tag{4.3}$$

This assumption can be proved for many model selection procedures. In Corollary 1 we will show it holds for the layer-wise Bonferroni corrected marginal testing procedure in Algorithm 1. Moreover, in the high dimensional super population study, data splitting and adaptation of ℓ_1 regularization can also fulfill such a requirement [18].

Jibberish 1 (Model selection consistency)**.** *Assume $\mathbb{M}^\star \neq \varnothing$. Assume Condition 3 holds. Let D be the largest positive integer such that there is at least one nonzero D-order interaction effects. Then Algorithm 1 with default weights and maximal interaction levels D has the following properties:*

1. Type I error control. *Algorithm 1 controls the Type I error rate, in the sense that*

$$\limsup_{N_k\to\infty,\ \forall k\in[K]} \mathbb{P}\left(\widehat{\mathbb{M}}\cap\mathbb{M}^{\star c}\neq\varnothing\right)\leq\alpha. \tag{4.4}$$

2. Model selection consistency. *Further assume $\alpha=\alpha_N\to 0$. Algorithm 1 consistently selects all the non-null terms with probability tending to 1:*

$$\limsup_{N_k\to\infty,\ \forall k\in[K]} \mathbb{P}\left(\widehat{\mathbb{M}}=\mathbb{M}^\star\right)=1. \tag{4.5}$$

Corollary 1 (Bonferroni corrected marginal test)**.** *Let $\widetilde{\mathbb{M}}_d=\mathcal{S}_N(\mathbb{M}^\star_{d,+})$ where $\mathbb{M}^\star_{d,+}=\mathcal{P}(\mathbb{M}^\star_{d-1})$. Assume the following scaling of parameters:*

- *$|\tau_{\mathcal{K}}|=\widetilde{O}(N_0^\delta)$ for some $\delta>-1/2$ and all $\mathcal{K}\in\mathbb{M}^\star_d$.*
- *$\alpha_d=\widetilde{O}(N_0^{-\delta_0})$ for some $\delta_0\geq 0$.*

Also assume the conditions in Lemma 1 hold. Then we have the following results for the model selection procedure based on Bonferroni corrected marginal t-test:

1. *(Validity)* $\limsup_{N\to\infty}\mathbb{P}\left\{\widetilde{\mathbb{M}}_d\cap\mathbb{M}^{\star c}_d\neq\varnothing\right\}\leq\alpha_d$ *for all* $d=1,\cdots,D$.
2. *(Consistency)* $\limsup_{N\to\infty}\mathbb{P}\left\{\widetilde{\mathbb{M}}^c_d\cap\mathbb{M}^\star_d\neq\varnothing\right\}=0$ *for all* $d=1,\cdots,D$.
3. *(Type I error control) Overall the procedure achieves family-wise error rate control:*

$$\limsup_{N\to\infty}\mathbb{P}\left(\widehat{\mathbb{M}}\cap\mathbb{M}^{\star c}\neq\varnothing\right)\leq\frac{\alpha_1}{K}\cdot|\mathbb{M}^\star_1|+\sum_{d=2}^{D}\frac{\alpha_d}{|\mathbb{M}^\star_{d,+}|}\cdot|\mathbb{M}^\star_d|\leq\alpha.$$

4. *(Perfect selection) When $\delta_0<0$, $\alpha_d\to 0$,*

$$\lim_{N\to\infty}\mathbb{P}\left(\widehat{\mathbb{M}}=\mathbb{M}^\star\right)=1.$$

Benefits of heredity. It is well-known that Bonferroni correction controls Type I error rate at the cost of inflated type II error rates. The heredity structure can alleviate this inflation by reducing the number of tests we need to work on within each level. Mathematically, from the proof, by our calculation, in the k-th step the Type II error of the Bonferroni correction is upper bounded by

$$\begin{aligned}&\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\} \\ &\leq \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^\star} \Phi\left\{r_{\mathcal{K}}^{-1}\left(Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\} - \Phi\left\{r_{\mathcal{K}}^{-1}\left(-Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\}, \end{aligned} \tag{4.6}$$

where

$$Z_d^\star = \sqrt{2\ln\frac{2|\mathbb{M}_{d,+}^\star|}{\alpha_d}} + o(1), \text{ as } \alpha_d \to 0.$$

On the other hand, one can easily formulate a lower bound:

$$\begin{aligned}&\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\} \\ &\geq \limsup_{N\to\infty} \max_{\mathcal{K}\in\mathbb{M}_d^\star} \Phi\left\{r_{\mathcal{K}}^{-1}\left(Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\} - \Phi\left\{r_{\mathcal{K}}^{-1}\left(-Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\}. \end{aligned} \tag{4.7}$$

Since we are considering fixed dimension, the bounds (4.6) and (4.7) are almost tight up to a factor $|\mathbb{M}_d^\star|$. Hence it's important to characterize the rate of the following quantity for any $\mathcal{K} \in \mathbb{M}_d^\star$:

$$\Phi\left\{r_{\mathcal{K}}^{-1}\left(Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\} - \Phi\left\{r_{\mathcal{K}}^{-1}\left(-Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\}.$$

When there is no heredity structure, a full Bonferroni correction would give

$$\widetilde{Z}_d^\star = \sqrt{2\ln\frac{2\binom{K}{d}}{\alpha_d}} + o(1), \text{ as } \alpha_d \to 0.$$

WLOG we assume $\tau_{\mathcal{K}} > 0$. Consider each summand in (4.6); the leading term would be the first component: $\Phi\left\{r_{\mathcal{K}}^{-1}\left(Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\}$. Asymptotically, when $\alpha_d \to 0$, we have

$$\begin{aligned}\lim_{N\to\infty} \frac{\Phi\left\{r_{\mathcal{K}}^{-1}\left(Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\}}{\Phi\left\{r_{\mathcal{K}}^{-1}\left(\widetilde{Z}_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\}} &= \lim_{N\to\infty} \frac{\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(\tau_{\mathcal{K}}/\sigma_{\mathcal{K}} - Z_d^\star)^2}{2r_{\mathcal{K}}^2}\right\}}{\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(\tau_{\mathcal{K}}/\sigma_{\mathcal{K}} - \widetilde{Z}_d^\star)^2}{2r_{\mathcal{K}}^2}\right\}} \text{ (L'Hopital's rule)} \\ &= \lim_{N\to\infty} \exp\left\{\frac{\widetilde{Z}_d^{\star 2} - Z_d^{\star 2}}{2r_{\mathcal{K}}^2}\right\} \cdot \exp\left\{\frac{\tau_{\mathcal{K}}(Z_d^\star - \widetilde{Z}_k^\star)}{\sigma_{\mathcal{K}} r_{\mathcal{K}}^2}\right\} \\ &= \lim_{N\to\infty} \left\{\frac{|\mathbb{M}_{d,+}^\star|}{\binom{K}{d}}\right\}^{-\frac{1}{2r_{\mathcal{K}}^2}} \cdot \exp\left\{-\frac{\tau_{\mathcal{K}}(\widetilde{Z}_d^\star - Z_d^\star)}{\sigma_{\mathcal{K}} r_{\mathcal{K}}^2}\right\}. \end{aligned}$$

We know that

$$\widetilde{Z}_d^\star - Z_d^\star = \frac{\widetilde{Z}_d^{\star 2} - Z_d^{\star 2}}{Z_d^\star + \widetilde{Z}_d^\star} = \frac{2\ln\{\binom{K}{d}/|\mathbb{M}_{d,+}^\star|\}}{2\sqrt{\ln\frac{1}{\alpha_d}}}\{1+o(1)\}.$$

Hence

$$\lim_{N\to\infty}\frac{\Phi\left\{r_{\mathcal{K}}^{-1}\left(Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\}}{\Phi\left\{r_{\mathcal{K}}^{-1}\left(\widetilde{Z}_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\}} = \lim_{N\to\infty}\left\{\frac{|\mathbb{M}_{d,+}^\star|}{\binom{K}{d}}\right\}^{\frac{\tau_{\mathcal{K}}\{1+o(1)\}}{\sigma_{\mathcal{K}} r_{\mathcal{K}}^2\sqrt{\ln(1/\alpha_k)}} - \frac{1}{2r_{\mathcal{K}}^2}}.$$

Now by our assumption, $\alpha_d = \widetilde{O}(N^{-\delta_0})$ for some $\delta_0 \geq 0$, $\tau_{\mathcal{K}} = \widetilde{O}(N^\delta)$ for some $\delta > -1/2$,

$$\frac{\tau_{\mathcal{K}}\{1+o(1)\}}{\sigma_{\mathcal{K}} r_{\mathcal{K}}^2\sqrt{\ln(1/\alpha_d)}} - \frac{1}{2r_{\mathcal{K}}^2} = \widetilde{O}\left(\frac{N^{\delta+1/2}}{\max\{\delta_0\ln N, 1\}}\right).$$

To sum up, within the d-th level, the ratio of the Type II error rate with/without heredity assumptions has the order of

$$O\left\{\left(|\mathbb{M}_{d,+}^\star|/\binom{K}{d}\right)^{\frac{CN^{\delta+1/2}}{\max\{\delta_0\ln N,1\}}}\right\}. \tag{4.8}$$

This shows that the heredity assumption can significantly reduce the type II error and improve the power of the selection procedure if the true effects have sparse structure (that is, $|\mathbb{M}_d^\star|$ is much smaller than the total number of d-way effects $\binom{K}{d}$).

4.3 Factor level combination selection from a non-asymptotic view

We generalize the definition of the tiers (1) to embrace a non-asymptotic study of the factor level combination selection results. Suppose there exists H scalars $\overline{Y}_{(1)} > \cdots > \overline{Y}_{(H)}$, with which the set of Γ_{K_0} can be partitioned into H tiers:

$$\Gamma_{K_0;h} = \left\{\overline{Y}(\boldsymbol{z}) \in \Gamma_{K_0} \mid |\overline{Y}(\boldsymbol{z}) - \overline{Y}_{(h)}| = o(N_0^{-\delta_3})\right\}, h = 1, \cdots, H. \tag{4.9}$$

Now we define

$$d_h = \max_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}}|\overline{Y}(\boldsymbol{z}) - \overline{Y}_{(h)}|,\ d_h^\star = \min_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h}}|\overline{Y}(\boldsymbol{z}) - \overline{Y}_{(h)}|.$$

which we refer to as within-group distance and between-group distance respectively.

Jibberish 2 (Convergence of the selected tie sets)**.** *Assume the following regularization condition on the variance of* $\widehat{Y}_{\mathrm{AVG}}$*,* V*:*

$$\lambda_{\min}(N_0 V) \geq \underline{\lambda}.$$

Also assume Assumption 1, 2 and 3 hold. Assume the following scaling of parameters: $d_h^\star = O(N_0^{\delta_1}), \eta_N = O(N_0^{\delta_2}), d_h = O(N_0^{\delta_3})$ *with* $-1/2 < \delta_1 < \delta_2 < \delta_3$*. Let* $\underline{\epsilon} < \bar{\epsilon}$ *be two positive real numbers. If we select the factor level combinations using Algorithm 2, taking* $b_L = b_R = 2\underline{\epsilon}\eta_N$*, when* N_0 *is large enough, i.e.,* $N_0 > n_0(\delta_1, \delta_2, \delta_3, \bar{\epsilon}, \underline{\epsilon}, \beta)$*, it holds that*

$$\begin{aligned}
&\mathbb{P}\left\{\forall\ h \in [H_0] : \widehat{\mathcal{T}}_{K_0;h} = \mathcal{T}_{K_0;h}\right\} \\
\geq& 1 - \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^\star\} - 32\left(\sum_{h=1}^{H_0} |\mathcal{T}_{K_0;h}|\right)|\Gamma_{K_0}| \frac{C \max_{i\in[N],\boldsymbol{z}\in\mathcal{T}} |\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}} \cdot \sqrt{\frac{|\mathbb{M}^\star|}{N_0 Q}}.
\end{aligned}$$

4.4 Inference of the ordered values

In the previous sections, we have shown that, under certain conditions, we can select the true model $\mathbb{M}^\star$ and $\mathcal{T}_{K_0;h}$ with high probability. Hence we could expect that Step 4 of Algorithm 1 would generate a sequence of point estimates

$$\frac{1}{|\widehat{\mathcal{T}}_{K_0;h}|} \sum_{\boldsymbol{z}\in\widehat{\mathcal{T}}_{K_0;h}} \widehat{Y}_{\mathrm{RP}}(\boldsymbol{z})$$

that has similar asymptotic behavior as

$$\frac{1}{|\mathcal{T}_{K_0;h}|} \sum_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}} \widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z})$$

This can be rigorously summarized in the following theorem:

Jibberish 3 (Asymptotic results on the estimated effects)**.** *Let* Q *be fixed. Assume the conditions in Lemma 1, Theorem 1 and Theorem 2. Then the point estimates are asymptotically jointly normal:*

$$\sqrt{N_0}\left\{\begin{pmatrix} \widehat{Y}_{(1)} \\ \vdots \\ \widehat{Y}_{(H_0)} \end{pmatrix} - \begin{pmatrix} \overline{Y}_{(1)} \\ \vdots \\ \overline{Y}_{(H_0)} \end{pmatrix}\right\} \to N(0, W^\top V_{\lim} W),$$

where $W \in \mathbb{R}^{H_0\times Q}$ *has columns*

$$W_{\cdot h} = \sum_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}} \frac{w_{\boldsymbol{z}}}{|\mathcal{T}_{K_0;h}|} = (Q|\mathcal{T}_{K_0;h}|)^{-1} \sum_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}} G_{\mathbb{M}^\star} G_{\boldsymbol{z},\mathbb{M}^\star}^\top.$$

Moreover, $W^\top \widehat{V}_R W$ *is a robust variance estimator which satisfies*

$$N_0\{W^\top \widehat{V}_R W - W^\top V_R W\} \xrightarrow{p} 0,\ N_0 V_R \to V_{R,\lim},\ W^\top V_{R,\lim} W \succeq W^\top V_{\lim} W.$$

Benefits of reparametrization under sparsity. Consider one simple scenario: only looking at the first tier ($H_0 = 1$) which only contains a single element $\boldsymbol{z}_1$ ($|\mathcal{T}_{K_0;h}| = 1$). Asymptotically speaking, we have

$$\sqrt{N_0}\left\{\widehat{Y}_{\mathrm{AVG}}(\boldsymbol{z}_1) - \overline{Y}(\boldsymbol{z}_1)\right\} \xrightarrow{d} N(0, V_{\lim}(\boldsymbol{z}_1, \boldsymbol{z}_1));$$
$$\sqrt{N_0}\left\{\widehat{Y}_{\mathrm{RP}}(\boldsymbol{z}_1) - \overline{Y}(\boldsymbol{z}_1)\right\} \xrightarrow{d} N(0, w_{\boldsymbol{z}_1}^\top V_{\lim} w_{\boldsymbol{z}_1}).$$

When constructing confidence intervals, we would use robust variance estimation $\widehat{V}_R(\boldsymbol{z}_1, \boldsymbol{z}_1)$ and $w_{\boldsymbol{z}_1}^\top \widehat{V}_R w_{\boldsymbol{z}_1}$, which (after multiplied by N_0) have limits:

$$\begin{aligned}
V_{R,\lim}(\boldsymbol{z}_1, \boldsymbol{z}_1) &= \frac{S_{\lim}(\boldsymbol{z}_1, \boldsymbol{z}_1)}{c_{\lim}(\boldsymbol{z}_1)}, \\
w_{\boldsymbol{z}_1}^\top V_{R,\lim} w_{\boldsymbol{z}_1} &= Q^{-2} G_{\boldsymbol{z}_1,\mathbb{M}^\star} G_{\mathbb{M}^\star}^\top V_{R,\lim} G_{\mathbb{M}^\star} G_{\boldsymbol{z}_1,\mathbb{M}^\star}^\top \\
&\le \|w_{\boldsymbol{z}_1}\|_\infty^2 \sum_{\boldsymbol{z}\in\mathcal{T}} V_{R,\lim}(\boldsymbol{z}, \boldsymbol{z}) \\
&= \frac{|\mathbb{M}^\star|}{Q}\left\{Q^{-1}\sum_{\boldsymbol{z}\in\mathcal{T}} \frac{S_{\lim}(\boldsymbol{z}, \boldsymbol{z})}{c_{\lim}(\boldsymbol{z})}\right\}.
\end{aligned}$$

If Q is not that large, then one can simply run saturated regression and estimate all the factorial effects. However, when Q (or equivalently K) is large, clearly the confidence intervals based on reparametrization has more advantage when the model is actually sparse.

Chapter 5

Simulation studies

We consider a factorial experiment with 5 binary factors $F_1, \cdots, F_5$. The potential outcomes are generated independently from some gaussian super populations with varying means and constant standard deviation of one. That is to say,

$$Y_i(\boldsymbol{z}) \sim N(\mu(\boldsymbol{z}), 1), \ \boldsymbol{z} = (z_1, \cdots, z_5).$$

In our experiments, we create $\mu(\boldsymbol{z})$ so that all the interaction effects of order three or higher are all zero.

5.1 Factorial structure under weak heredity

In this subsection, we consider a specification of effects given by Figure 5.1. The effect sizes are determined in the following way: to impose sparsity in the model, we put the main effects τ_4 and τ_5 to be zero. Under weak heredity, the interaction terms involving F_4, F_5 all have to be zero. For the rest of the main effects, we pick them independently from a uniform distribution $U([-5,-1] \cup [1,5])$ to ensure a nonzero magnitude. For the rest of the interactions, we also choose them independently from $U([-5,-1] \cup [1,5])$; but to allow for zero size we introduce a censoring probability of 0.3. That is, there is a variable $B \sim \text{Ber}(0.3)$ independent of a uniform $U \sim U([-5,-1] \cup [1,5])$ such that $\tau_{kl} \sim U \cdot B$. We generate the finite population matrix from gaussian distributions with standard deviation 1, then permute the units among the 2^5 treatment combinations, each arm with 100 replications. The Monte Carlo simulation is repeated for 500 runs. Table 5.1 reports the model selection results based on several methods and heredity principles, and Table 5.2 reports the tier selection results, the confidence interval (CI) coverage rates along with the length of CIs.

For the model selection results, we see that both Bonferroni correction and LASSO can achieve a high probability of selection consistency. Taking the weak heredity structure and forward procedure into account, the performance would be slightly better than running a full selection over the whole model. But since the number of factors K is not too large here, the results does not demonstrate a large advantage.

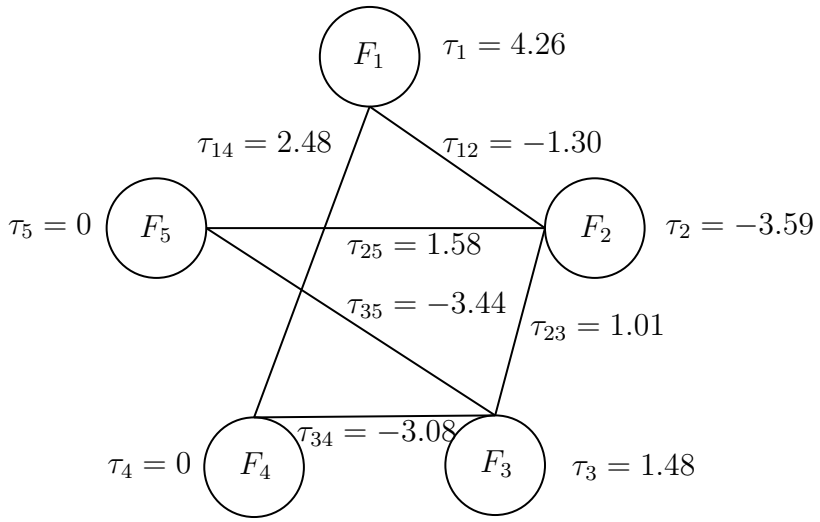

Figure 5.1: Synthetic factorial effects specification under weak heredity. No edge between two nodes means the interaction is zero.

For the Tier selection results, we see that under forward selection, the probability of perfect tier selection rises significantly. Besides, we can achieve better coverage rates for the confidence intervals. More importantly, the length of the CIs are much shorter compared to the case where we utilize naive group-wise averages, which verifies our theoretical discussion that we can borrow information across groups to improve the efficiency of the methods.

Table 5.1: Model selection for factorial structure generated under weak heredity

Selection	Heredity	Perfect	Over	Under	None of above
Bonferroni	No heredity	0.974	0.026	0	0
	Weak heredity	0.978	0.022	0	0
	Strong heredity	0	0	0.978	0.022
	Full WLS	0.972	0.026	0	0
LASSO	No heredity	0.974	0.020	0	0
	Weak heredity	0.980	0	0	0
	Strong heredity	0	0.028	1.000	0
	Full WLS	0.962	0.038	0	0

5.2 Factorial structure under strong heredity

In this subsection, we consider a specification of effects similar to the above setup but generated under strong heredity. The results are summarized in Table 5.3 and 5.4. In

Table 5.2: Tier selection and inference for factorial structure generated under weak heredity

		Tier Selection	Coverage	CI length
Bonferroni	No heredity	0.998	0.950	0.1828
	Weak	0.998	0.950	0.1828
	Strong	0.000	0.000	1.0470
LASSO	No heredity	1.000	0.952	0.1830
	Weak	1.000	0.000	0.1829
	Strong	0.000	0.950	1.0500
No selection	No heredity	0.960	0.936	0.2815

this case we have a sparser model, and since the strong heredity structure holds for the parameters, proceeding with this prior knowledge can lead to a better model selection result. Moreover, as we highlight earlier, we would also have better coverage and shorter confidence intervals after taking the structure into account.

Table 5.3: Model selection for factorial structure generated under strong heredity

Selection	Heredity	Perfect	Over	Under	None of above
Bonferroni	No heredity	0.974	0.026	0	0
	Weak heredity	0.974	0.026	0	0
	Strong heredity	0.986	0.014	0	0
	Full WLS	0.970	0.032	0	0
LASSO	No heredity	0.968	0.034	0	0
	Weak heredity	0.966	0.006	0	0
	Strong heredity	0.994	0.030	0	0
	Full WLS	0.966	0.034	0	0

Table 5.4: Tier selection and inference for factorial structure generated under strong heredity

		Tier Selection	Coverage	CI length
Bonferroni	No heredity	1.000	0.950	0.2062
	Weak	1.000	0.950	0.2062
	Strong	1.000	0.950	0.2062
LASSO	No heredity	1.000	0.950	0.2062
	Weak	1.000	0.950	0.2062
	Strong	1.000	0.950	0.2062
No selection	No heredity	0.974	0.938	0.2763

Chapter 6

Technical proofs

6.1 Proof of Theorem 1

Proof. **Induction proof of a basic fact.** According to the orthogonality of designs, the signs for all terms in the studied unsaturated population regressions are consistent with those of saturated regressions, which saves the effort of differentiating true models for partial and full regression. By induction we hope to prove the following fact under the given assumptions:

For all $D_0 \leq D$, we have

$$\left|\mathbb{P}\left(\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d = 1, \cdots, D_0\right) - \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d = 1, \cdots, D_0\right)\right| \to 0. \tag{6.1}$$

Clearly we know the following inclusion is always true: for any $D_0 \in [D]$,

$$\left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d = 1, \cdots, D_0\right\} \subset \left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d = 1, \cdots, D_0\right\}.$$

Hence the statement (6.1) is equivalent to: for all $D_0 \leq D$,

$$\mathbb{P}\left(\forall d \in [D_0], \widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star; \exists d \in [D_0], \widehat{\mathbb{M}}_d \subsetneq \mathbb{M}_d^\star\right) \to 0. \tag{6.2}$$

1. **Main effects.** First, since we assume the tests are consistent, meaning asymptotically no false negatives:

   $$\mathbb{P}\left(\widehat{\mathbb{M}}_1^c \cap \mathbb{M}_1^\star \neq \varnothing\right) \to 0.$$

   Or equivalently,

   $$\mathbb{P}\left(\mathbb{M}_1^\star \subset \widehat{\mathbb{M}}_1\right) \to 1.$$

   That being said,

   $$\mathbb{P}\left(\widehat{\mathbb{M}}_1 \subsetneq \mathbb{M}_1^\star\right) \to 0. \tag{6.3}$$

2. **Induction validity.** Generally speaking, the induction proceeds based on the following idea:

   The case for $D_0 = 1$ has been shown in the previous part. Now assume (6.1) or (6.2) is true for some $D_0 \geq 1$. For $D_0 + 1$, the following holds:

$$\begin{aligned}
0 &\leq \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0 + 1\right\}\right) - \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right) \\
&= \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0 + 1\right\} - \left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right) \\
&\leq \mathbb{P}\left(\forall d \in [D_0 + 1], \widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star; \exists d \in [D_0], \widehat{\mathbb{M}}_d \subsetneq \mathbb{M}_d^\star\right) \\
&\leq \mathbb{P}\left(\forall d \in [D_0], \widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star; \exists d \in [D_0], \widehat{\mathbb{M}}_d \subsetneq \mathbb{M}_d^\star\right) \to 0. \text{ (by (6.2))}
\end{aligned}$$

   Hence

$$\left|\mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0 + 1\right\}\right) - \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right)\right| \to 0. \tag{6.4}$$

   Now $\widehat{\mathbb{M}}_{D_0+1}$ is generated based on $\widehat{\mathbb{M}}_{D_0}$ and the set of estimates over the prescreened effect set $\widehat{\mathbb{M}}_{D_0+1,+}$. Due to the Step 3 in Algorithm 1, under Assumption 3, on the event $\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star$ we have

$$\widehat{\mathbb{M}}_{d+1} = \widetilde{\mathbb{M}}_{d+1}.$$

   Hence we can compute

$$\begin{aligned}
0 \leq& \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right) - \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0 + 1\right) \\
=& \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subsetneq \mathbb{M}_{D_0+1}^\star\right) \\
=& \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widetilde{\mathbb{M}}_{D_0+1} \subsetneq \mathbb{M}_{D_0+1}^\star\right) \\
\leq& \mathbb{P}\left(\widetilde{\mathbb{M}}_{D_0+1}^c \cap \mathbb{M}_{D_0+1}^\star \neq \varnothing\right) \to 0.
\end{aligned}$$

   The last convergence holds because of the consistency of the test.

   The induction can be proceeded.

   Proof of the first result. Now it follows

$$\begin{aligned}
&\limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}} \cap \mathbb{M}^{\star c} \neq \varnothing\right) \\
&= \limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_1 \cap \mathbb{M}_1^{\star c} \neq \varnothing\right) + \sum_{D_0=2}^{D} \mathbb{P}\left(\widehat{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} = \varnothing, d \in [D_0 - 1]; \widehat{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star c} \neq \varnothing\right)
\end{aligned}$$

$$= \limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_1 \cap \mathbb{M}_1^{\star c} \neq \varnothing\right) + \sum_{D_0=2}^{D} \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \in [D_0-1]; \widehat{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star c} \neq \varnothing\right)$$

(using (6.1) and the fact that D is a fixed integer)

$$\leq \limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_1 \cap \mathbb{M}_1^{\star c} \neq \varnothing\right) + \sum_{D_0=2}^{D} \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \in [D_0-1]; \widetilde{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star c} \neq \varnothing\right)$$

(since $\widehat{\mathbb{M}}_{D_0,+} = \mathcal{P}(\widehat{\mathbb{M}}_{D_0-1}) = \mathcal{P}(\mathbb{M}^\star_{D_0-1}) = \mathbb{M}^\star_{D_0,+}$ and $\widehat{\mathbb{M}}_{D_0} = \mathcal{S}_N(\widehat{\mathbb{M}}_{D_0,+}) = \widetilde{\mathbb{M}}_{D_0}$)

$$\leq \limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_1 \cap \mathbb{M}_1^{\star c} \neq \varnothing\right) + \sum_{D_0=2}^{D} \mathbb{P}\left(\widetilde{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star\, c} \neq \varnothing\right) \leq \sum_{D_0=1}^{D} \alpha_{D_0} = \alpha. \tag{6.5}$$

Therefore the FWER gets controlled under α.

Proof of the second result. Under $\alpha = \alpha_N \to 0$, (6.5) implies $\widehat{\mathbb{M}} \subset \mathbb{M}^\star$ with probability tending to one, or

$$\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, \text{ for } d = 1, \cdots, D.$$

Now apply (6.1), we obtain

$$\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, \text{ for } d = 1, \cdots, D,$$

with probability tending to one, which concludes the proof.

□

6.2 Proof of Corollary 1

Proof. 1. First we show validity.

$$\begin{aligned}
\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} \neq \varnothing\right\} &= \limsup_{N\to\infty} \mathbb{P}\left\{\exists \mathcal{K} \in \mathbb{M}^\star_{d,+} \backslash \mathbb{M}_d^\star, \left|\frac{\widehat{\tau}_{\mathcal{K}}}{\widehat{\sigma}_{\mathcal{K}}}\right| \geq \Phi^{-1}\left(1 - \frac{\alpha_d}{2|\mathbb{M}^\star_{d,+}|}\right)\right\} \\
&\leq \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}^\star_{d,+}\backslash\mathbb{M}_d^\star} \mathbb{P}\left\{\left|\frac{\widehat{\tau}_{\mathcal{K}}}{\widehat{\sigma}_{\mathcal{K}}}\right| \geq \Phi^{-1}\left(1 - \frac{\alpha_d}{2|\mathbb{M}^\star_{d,+}|}\right)\right\} \\
&\leq \sum_{\mathcal{K}\in\mathbb{M}^\star_{d,+}\backslash\mathbb{M}_d^\star} \frac{\alpha_d}{|\mathbb{M}^\star_{d,+}|} \leq \alpha_d.
\end{aligned}$$

2. Secondly, we show consistency.

$$\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\} = \limsup_{N\to\infty} \mathbb{P}\left\{\exists \mathcal{K} \in \mathbb{M}_d^\star, \left|\frac{\widehat{\tau}_\mathcal{K}}{\widehat{\sigma}_\mathcal{K}}\right| \leq \Phi^{-1}\left(1 - \frac{\alpha_d}{2|\mathbb{M}_{d,+}^\star|}\right)\right\}$$
$$\leq \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^\star} \mathbb{P}\left\{\left|\frac{\widehat{\tau}_\mathcal{K}}{\widehat{\sigma}_\mathcal{K}}\right| \leq \Phi^{-1}\left(1 - \frac{\alpha_d}{2|\mathbb{M}_{d,+}^\star|}\right)\right\}.$$

Asymptotically we have

$$\frac{\widehat{\tau}_\mathcal{K}}{\widehat{\sigma}_\mathcal{K}} - \frac{\tau_\mathcal{K}}{\sigma_\mathcal{K}} \xrightarrow{d} N\left(0, \frac{\mathrm{Var}(\widehat{\tau}_\mathcal{K})}{\mathbb{E}(\widehat{\sigma}_\mathcal{K}^2)}\right) := N(0, r_\mathcal{K}^2).$$

For simplicity, let

$$Z_d^\star = \Phi^{-1}\left(1 - \frac{\alpha_d}{2|\mathbb{M}_{d,+}^\star|}\right).$$

Hence

$$\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_k^\star \neq \varnothing\right\}$$
$$\leq \lim_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^\star} \mathbb{P}\left\{-Z_d^\star - \frac{\tau_\mathcal{K}}{\sigma_\mathcal{K}} \leq \frac{\widehat{\tau}_\mathcal{K}}{\widehat{\sigma}_\mathcal{K}} - \frac{\tau_\mathcal{K}}{\sigma_\mathcal{K}} \leq Z_d^\star - \frac{\tau_\mathcal{K}}{\sigma_\mathcal{K}}\right\}$$
$$= \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^\star} \Phi\left\{r_\mathcal{K}^{-1}\left(Z_d^\star - \frac{\tau_\mathcal{K}}{\sigma_\mathcal{K}}\right)\right\} - \Phi\left\{r_\mathcal{K}^{-1}\left(-Z_d^\star - \frac{\tau_\mathcal{K}}{\sigma_\mathcal{K}}\right)\right\}. \tag{6.6}$$

Note that when $\alpha_d \to 0, N \to \infty$, we have

$$Z_d^\star = \widetilde{O}\left(\sqrt{2\ln\frac{2|\mathbb{M}_{d,+}^\star|}{\alpha_d}}\right) = \widetilde{O}(\max\{\delta_0 \ln N, \ln(2|\mathbb{M}_{d,+}^\star|)\}), \quad \left|\frac{\tau_\mathcal{K}}{\sigma_\mathcal{K}}\right| = \widetilde{O}(N^{1/2+\delta}).$$

Since $\delta > -1/2$ and $\delta_0 \geq 0$, we have $|\frac{\tau_\mathcal{K}}{\sigma_\mathcal{K}}| \to \infty$ and $Z_d^\star/(|\frac{\tau_\mathcal{K}}{\sigma_\mathcal{K}}|) \to 0$. Hence the above limit (6.6) converges to zero. This concludes the proof.

3. Based on the above two parts and Theorem 1, it suffices to conclude the Type I error rate control. A more delicate analysis in this particular setup can actually lead to sharper bound. Based on (6.5), we directly compute

$$\limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}} \cap \mathbb{M}^{\star c} \neq \varnothing\right)$$
$$\leq \limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_1 \cap \mathbb{M}_1^{\star c} \neq \varnothing\right) + \sum_{D_0=2}^{D} \mathbb{P}\left(\widetilde{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star\ c} \neq \varnothing\right)$$
$$\leq \frac{\alpha_1}{K} \cdot |\mathbb{M}_1^\star| + \sum_{D_0=2}^{D} \frac{\alpha_{D_0}}{|\mathbb{M}_{D_0,+}^\star|} \cdot |\mathbb{M}_{D_0}^\star| \leq \alpha.$$

4. The perfect selection result follows without difficult from Part 1,2 and Theorem 1.

□

6.3 Proof of Theorem 2

Proof. The high level idea of the proof is that: we first prove the non-asymptotic bounds over the random event $\widehat{\mathbb{M}} = \mathbb{M}^\star$, then proceed to make up for the cost of $\widehat{\mathbb{M}} \neq \mathbb{M}^\star$. Over $\widehat{\mathbb{M}} = \mathbb{M}^\star$, from Step 1 of Algorithm 2 we have

$$\widehat{Y}_{\mathrm{RP}} = \widehat{Y}^\star_{\mathrm{RP}} = G_{\mathbb{M}^\star}\widehat{\tau}(\mathbb{M}^\star) = Q^{-1}G_{\mathbb{M}^\star}G^\top_{\mathbb{M}^\star}\widehat{Y}_{\mathrm{AVG}},\ \widehat{\Gamma}_{K_0} = \left\{ \widehat{Y}^\star_{\mathrm{RP}}(z_1 \cdots z_K) \mid \sum_{k=1}^{K} z_k \leq K_0 \right\}.$$

Under Assumption 2, we have

$$\mathbb{E}\{\widehat{Y}^\star_{\mathrm{RP}}\} = G_{\mathbb{M}^\star}\tau(\mathbb{M}^\star) = G\tau = \overline{Y}.$$

A combinatorial Berry-Esseen bound. We hope to apply Lemma 2 to establish a Berry-Esseen bound for each $\widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z})$. Clearly we have

$$\widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}) = w_{\boldsymbol{z}}^\top \widehat{Y}_{\mathrm{AVG}},\ w_{\boldsymbol{z}}^\top = Q^{-1}G_{\boldsymbol{z},\mathbb{M}^\star}G^\top_{\mathbb{M}^\star}. \tag{6.7}$$

By simple calculation we have

$$\|w_{\boldsymbol{z}}\|_\infty = Q^{-1}|\mathbb{M}^\star|,\ \|w_{\boldsymbol{z}}\|_2 = \sqrt{Q^{-1}|\mathbb{M}^\star|}.$$

Also we can show that

$$\frac{v_N}{v_R} = \sqrt{\frac{w_{\boldsymbol{z}}^\top (N_0 V) w_{\boldsymbol{z}}}{w_{\boldsymbol{z}}^\top (N_0 V_R) w_{\boldsymbol{z}}}} \geq \sqrt{\frac{\underline{\lambda}\|w_{\boldsymbol{z}}\|_2^2}{\overline{c}\overline{s}\|w_{\boldsymbol{z}}\|_2^2}} = \sqrt{\frac{\underline{\lambda}}{\overline{c}\overline{s}}} := \kappa,$$

and obtain

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left\{ \frac{\widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})}{v_N} \leq t \right\} - \Phi(t) \right| \leq \frac{C(\underline{c}, \overline{c}, \kappa) \max_{i\in[N], \boldsymbol{z}\in\mathcal{T}} |Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}} \cdot \sqrt{\frac{|\mathbb{M}^\star|}{N_0 Q}}. \tag{6.8}$$

A probabilistic bound on the ordered statistics. We show a bound on

$$\mathbb{P}\left\{ \max_{\boldsymbol{z}\in\cup_{h'=1}^{h-1}\mathcal{T}_{K_0;h'}} \widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}) < \min_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}} \widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}) \leq \max_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}} \widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}) < \min_{\boldsymbol{z}\in\cup_{h'=h+1}^{H}\mathcal{T}_{K_0;h'}} \widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}) \right\}.$$

It's easy to show that

$$1 - \Phi(x) = \int_x^\infty \phi(t)dt \leq \frac{1}{x}\int_x^\infty t\phi(t)dt \leq \frac{1}{\sqrt{2\pi}x}\left\{ \exp\left(-\frac{x^2}{2}\right) \right\} \leq \frac{1}{\sqrt{2\pi}x}.$$

Hence we know that

$$\mathbb{P}\left\{\sqrt{N_0}\left|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})\right|\geq\sqrt{N_0}d_h^{\star}\right\} \tag{6.9}$$

$$\leq\frac{v_N}{\sqrt{2\pi}d_h^{\star}}\cdot\exp\left(-\frac{d_h^{\star 2}}{2v_N^2}\right)+\frac{C(\underline{c},\overline{c},\kappa)\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}. \tag{6.10}$$

Therefore, for all $\boldsymbol{z}\in\cup_{h'=1}^{h-1}\mathcal{T}_{K_0;h'}$ and $\boldsymbol{z}'\in\mathcal{T}_{K_0;h}$,

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})<0\right\}\\
&=\mathbb{P}\left\{\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}'))-\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}))<\sqrt{N_0}(\overline{Y}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}'))\right\}\\
&\leq\mathbb{P}\left\{\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}'))-\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}))<-2\sqrt{N_0}d_h^{\star}\right\}\\
&=\mathbb{P}\Big\{\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}'))-\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}))<-2\sqrt{N_0}d_h^{\star},\\
&\qquad\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}))<\sqrt{N_0}d_h^{\star}\Big\}\\
&+\mathbb{P}\Big\{\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}'))-\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}))<-2\sqrt{N_0}d_h^{\star},\\
&\qquad\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}))<\sqrt{N_0}d_h^{\star}\Big\}\\
&\leq\mathbb{P}\left\{\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}'))<-\sqrt{N_0}d_h^{\star}\right\}+\mathbb{P}\left\{\sqrt{N_0}(\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}))\geq\sqrt{N_0}d_h^{\star}\right\}.
\end{aligned}$$

Using the earlier results (6.9), we can actually know

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})<0\right\}\\
&\leq\frac{\sqrt{\overline{c}\overline{s}|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0Qd_h^{\star 2}}{2\overline{c}\overline{s}|\mathbb{M}^{\star}|}\right)+\frac{C(\underline{c},\overline{c},\kappa)\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}.
\end{aligned}$$

The same procedure can be repeated for $\boldsymbol{z}\in\cup_{h'=h+1}^{H}\mathcal{T}_{K_0;h'}$ and $\boldsymbol{z}'\in\mathcal{T}_{K_0;h}$. Now a uniou bound gives

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\cup_{h'=1}^{h-1}\mathcal{T}_{K_0;h'}}\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})<\min_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}}\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})\leq\max_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}}\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})<\min_{\boldsymbol{z}\in\cup_{h'=h+1}^{H}\mathcal{T}_{K_0;h'}}\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})\right\}\\
&\geq 1-2|\Gamma_{K_0}||\mathcal{T}_{K_0;h}|\left\{\frac{\sqrt{\overline{c}\overline{s}|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0Qd_h^{\star 2}}{2\overline{c}\overline{s}|\mathbb{M}^{\star}|}\right)+\frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}\right\}.
\end{aligned}$$

Now using that $d_h^{\star}=O(N_0^{\delta_1})$, $N_0d_h^{\star 2}=O(N_0^{1+2\delta_2})$ with $1+2\delta>0$. We know that the exponential term is of lower order compared to the polynomial term. Thus when N is large

enough, we have

$$\mathbb{P}\left\{\max_{\boldsymbol{z}\in\cup_{h'=1}^{h-1}\mathcal{T}_{K_0;h'}}\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})<\min_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}}\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})\leq\max_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}}\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})<\min_{\boldsymbol{z}\in\cup_{h'=h+1}^{H}\mathcal{T}_{K_0;h'}}\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})\right\}$$
$$\geq 1-4|\Gamma_{K_0}||\mathcal{T}_{K_0;h}|\frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}. \tag{6.11}$$

Nice separation. Suppose we are working on a random coordinate $\widetilde{\boldsymbol{z}}$. For $\boldsymbol{z}\notin\mathcal{T}_{K_0;h}$ and any $\bar{\epsilon}>0$,

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h}}|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}})|/\eta_N\geq 2\bar{\epsilon}\right\}\\
&\geq\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h}}|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}})|/\eta_N\geq 2\bar{\epsilon},\widetilde{m}\in\mathcal{T}_{N,h}\right\}\\
&\geq\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h},\boldsymbol{z}'\in\mathcal{T}_{K_0;h}}|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}')|/\eta_N\geq 2\bar{\epsilon},\widetilde{m}\in\mathcal{T}_{K_0;h}\right\}\\
&\geq\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h},\boldsymbol{z}'\in\mathcal{T}_{K_0;h}}|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}')|/\eta_N\geq 2\bar{\epsilon}\right\}+\mathbb{P}\left\{\widetilde{m}\in\mathcal{T}_{K_0;h}\right\}-1\\
&\geq\mathbb{P}\left\{\widetilde{m}\in\mathcal{T}_{K_0;h}\right\}-\sum_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h},\boldsymbol{z}'\in\mathcal{T}_{N,h}}\mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}')|/\eta_N\leq 2\bar{\epsilon}\right\}.
\end{aligned}\tag{6.12}$$

To proceed we have the following tail bound:

$$\begin{aligned}
&\mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')|/\eta_N\leq 2\bar{\epsilon}\right\}\\
=&\mathbb{P}\left\{|\{\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})\}-\{\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}')\}-\{\overline{Y}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}')\}|\leq 2\bar{\epsilon}\eta_N\right\}\\
\leq&\mathbb{P}\left\{|\overline{Y}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}')|-|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|-|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}')|\leq 2\bar{\epsilon}\eta_N\right\}\\
\leq&\mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|+|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}')|\geq 2d^{\star}_h-2\bar{\epsilon}\eta_N\right\}\\
\leq&\mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|\geq d^{\star}_h-\bar{\epsilon}\eta_N\right\}+\mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}')|\geq d^{\star}_h-\bar{\epsilon}\eta_N\right\}\\
&\text{(since } \boldsymbol{z}\notin\mathcal{T}_{K_0;h} \text{ and } \boldsymbol{z}'\in\mathcal{T}_{K_0;h})\\
\leq&4\left\{\frac{\sqrt{\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0Q}(d^{\star}_h-\bar{\epsilon}\eta_N)}\cdot\exp\left(-\frac{N_0Q(d^{\star}_h-\bar{\epsilon}\eta_N)^2}{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)+\frac{\bar{\epsilon}\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}\right\}.\\
&\text{(This is deduced analogously to the proof in the previous part)}
\end{aligned}$$

By the conditions we imposed in the theorem, we know that when N is large enough,

$$d^{\star}_h-\bar{\epsilon}\eta_N>d^{\star}_h/2.$$

Hence, for $N > N(\bar{\epsilon}, \delta_1, \delta_2)$ large enough, we have

$$\sum_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h},\boldsymbol{z}'\in\mathcal{T}_{K_0;h}} \mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')|/\eta_N \le 2\bar{\epsilon}\right\}$$
$$\le 4|\mathcal{T}_{K_0;h}||\Gamma_{K_0}|\left\{\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{\pi N_0 Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0 Q d_h^{\star 2}}{8\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right) + \frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.$$

Combined with (6.12), we have:

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h}}|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}})|/\eta_N \ge 2\bar{\epsilon}\right\}\\
\ge&\mathbb{P}\left\{\widetilde{m}\in\mathcal{T}_{K_0;h}\right\} - \underbrace{4|\mathcal{T}_{K_0;h}||\Gamma_{K_0}|\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{\pi N_0 Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0 Q d_h^{\star 2}}{8\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)}_{\text{Term I}}\\
&-\underbrace{4|\mathcal{T}_{K_0;h}||\Gamma_{K_0}|\frac{\bar{\epsilon}\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{QN_0}}}_{\text{Term II}}.
\end{aligned} \tag{6.13}$$

Now using that $d_h^{\star} = O(N_0^{\delta_1})$, $N_0 d_h^{\star 2} = O(N_0^{1+2\delta_2})$ with $1+2\delta > 0$. The Term I involving the exponential part is always of lower order than Term II. That being said, when N_0 is sufficiently large,

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h}}|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}})|/\eta_N \ge 2\bar{\epsilon}\right\}\\
\ge&\mathbb{P}\left\{\widetilde{m}\in\mathcal{T}_{K_0;h}\right\} - 8|\mathcal{T}_{K_0;h}||\Gamma_{K_0}|\frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}.
\end{aligned}$$

Similarly we can show for any $z\in\mathcal{T}_{K_0;h}$ and $\underline{\epsilon} > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}}|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}})|/\eta_N \le 2\underline{\epsilon}\right\}\\
&\ge \mathbb{P}\left\{\widetilde{\boldsymbol{z}}\in\mathcal{T}_{K_0;h}\right\} - \sum_{\boldsymbol{z}\neq\boldsymbol{z}'\in\mathcal{T}_{K_0;h}}\mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')|/\eta_N > 2\underline{\epsilon}\right\}.
\end{aligned}$$

Then we have for $\boldsymbol{z}\neq\boldsymbol{z}'\in\mathcal{T}_{K_0;h}$,

$$\begin{aligned}
&\mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')|/\eta_N > 2\underline{\epsilon}\right\}\\
&\le \mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})| \ge \underline{\epsilon}\eta_N - d_h\right\} + \mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}') - \overline{Y}(\boldsymbol{z}')| \ge \underline{\epsilon}\eta_N - d_h\right\}\\
&\le 4\left\{\frac{\sqrt{\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0 Q}(\underline{\epsilon}\eta_N - d_h)}\cdot\exp\left(-\frac{N_0 Q(\underline{\epsilon}\eta_N - d_h)^2}{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right) + \frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned}$$

By the scaling of the parameters, when N_0 is large enough $N > N(\delta_2, \delta_3, \underline{\epsilon})$, $\underline{\epsilon}\eta_N - d_h > \underline{\epsilon}\eta_N/2$. That being said,

$$\begin{aligned}
&\mathbb{P}\left\{|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}')|/\eta_N > 2\underline{\epsilon}\right\} \\
\leq & 4\left\{\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{\pi N_0 Q}(\underline{\epsilon}\eta_N)} \cdot \exp\left(-\frac{N_0 Q(\underline{\epsilon}\eta_N)^2}{8\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right) + \frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned}$$

Hence we have:

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}}|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}})|/\eta_N \leq 2\underline{\epsilon}\right\} \\
\geq & \mathbb{P}\left\{\widetilde{\boldsymbol{z}} \in \mathcal{T}_{K_0;h}\right\} - \underbrace{4|\mathcal{T}_{K_0;h}||\Gamma_{K_0}|\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{\pi N_0 Q}(\underline{\epsilon}\eta_N)} \cdot \exp\left(-\frac{N_0 Q(\underline{\epsilon}\eta_N)^{\star 2}}{8\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)}_{\text{Term I}} \\
& - \underbrace{4|\mathcal{T}_{K_0;h}||\Gamma_{K_0}|\frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}}_{\text{Term II}}.
\end{aligned}$$

Again, by the conditions, it is easy to show the term involving the exponential part is of lower order than the polynomial part. That being said,

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}}|\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}})|/\eta_N \leq 2\underline{\epsilon}\right\} \\
\geq & \mathbb{P}\left\{\widetilde{\boldsymbol{z}} \in \mathcal{T}_{K_0;h}\right\} - 8|\mathcal{T}_{K_0;h}||\Gamma_{K_0}|\frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}.
\end{aligned}$$

Specifying the random indices. Introduce the following random indices:

$$\widetilde{\boldsymbol{z}}_h = \mathop{\arg\max}_{\boldsymbol{z}\in\mathcal{T}_{K_0}\setminus\cup_{h'=0}^{h-1}\mathcal{T}_{K_0;h'}} \widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}).$$

Applying (6.11) we know that

$$\mathbb{P}\{\widetilde{\boldsymbol{z}}_h \in \mathcal{T}_{K_0;h}\} \geq 1 - 4|\Gamma_{K_0}||\mathcal{T}_{K_0;h}|\frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}.$$

Aggregating parts. Aggregating all the results above, we can show that, when N_0 is large enough, i.e., $N_0 > n_0(\delta_1, \delta_2, \delta_3, \bar{\epsilon}, \underline{\epsilon})$,

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}} |\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}})| \le \underline{\epsilon}\eta_N, \min_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h}} |\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}})| \ge \bar{\epsilon}\eta_N\right\} \\
\ge& 1 - 2(1 - \mathbb{P}\{\widetilde{\boldsymbol{z}} \in \mathcal{T}_{K_0;h}\}) - 16|\mathcal{T}_{K_0;h}||\Gamma_{K_0}|\frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}} \\
\ge& 1 - 32|\mathcal{T}_{K_0;h}||\Gamma_{K_0}|\frac{C\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}. \qquad (6.14)
\end{aligned}$$

Bounding the factor level combination selection probability. From Step 3 of Algorithm 2, we have

$$\begin{aligned}
&\mathbb{P}\left\{\forall\ h\in[H_0] : \widehat{\mathcal{T}}_{K_0;h} = \mathcal{T}_{K_0;h}\right\} \\
=&\mathbb{P}\Bigg\{\forall\ h\in[H_0] : |\widehat{Y}_{\mathrm{RP}}(\boldsymbol{z}) - \max_{\boldsymbol{z}\in\mathcal{T}_{K_0}\setminus\cup_{h'=0}^{h-1}\mathcal{T}_{K_0;h'}} \widehat{Y}_{\mathrm{RP}}(\boldsymbol{z})| \le \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\in\mathcal{T}_{K_0;h}; \\
&\qquad |\widehat{Y}_{\mathrm{RP}}(\boldsymbol{z}) - \max_{\boldsymbol{z}\in\mathcal{T}_{K_0}\setminus\cup_{h'=0}^{h-1}\mathcal{T}_{K_0;h'}} \widehat{Y}_{\mathrm{RP}}(\boldsymbol{z})| > \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\notin\mathcal{T}_{K_0;h}\Bigg\} \\
\ge&\mathbb{P}\Bigg\{\forall\ h\in[H_0] : |\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \max_{\boldsymbol{z}\in\mathcal{T}_{K_0}\setminus\cup_{h'=0}^{h-1}\mathcal{T}_{K_0;h'}} \widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})| \le \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\in\mathcal{T}_{K_0;h}; \\
&\qquad |\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \max_{\boldsymbol{z}\in\mathcal{T}_{K_0}\setminus\cup_{h'=0}^{h-1}\mathcal{T}_{K_0;h'}} \widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z})| > \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\notin\mathcal{T}_{K_0;h}\Bigg\} - \mathbb{P}\{\widehat{\mathbb{M}} \ne \mathbb{M}^{\star}\} \\
=&\mathbb{P}\Bigg\{\forall h\in[H_0] : |\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}_h)| \le \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\in\mathcal{T}_{K_0;h}; \\
&\qquad |\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}_h)| > \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\notin\mathcal{T}_{K_0;h}\Bigg\} - \mathbb{P}\{\widehat{\mathbb{M}} \ne \mathbb{M}^{\star}\}
\end{aligned}$$

(where we introduce random index $\widetilde{\boldsymbol{z}}_h$ to record the position that achieves maximum)

$$\begin{aligned}
\ge&\mathbb{P}\Bigg\{\forall h\in[H_0] : |\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}_h)| \le \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\in\mathcal{T}_{K_0;h}; \\
&\qquad |\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}_h)| > \bar{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\notin\mathcal{T}_{K_0;h}\Bigg\} - \mathbb{P}\{\widehat{\mathbb{M}} \ne \mathbb{M}^{\star}\}
\end{aligned}$$

(simply using the fact that $\bar{\epsilon} > \underline{\epsilon}$)

$$=\mathbb{P}\left\{\forall h\in[H_0] : \max_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}} |\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}_h)| \le \underline{\epsilon}\eta_N; \min_{\boldsymbol{z}\notin\mathcal{T}_{K_0;h}} |\widehat{Y}^{\star}_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}_h)| > \bar{\epsilon}\eta_N\right\}$$

$$
\begin{aligned}
&- \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^\star\} \\
\geq & 1 - \sum_{h=1}^{H_0} \left(1 - \mathbb{P}\left\{ \max_{\boldsymbol{z} \in \mathcal{T}_{K_0;h}} |\widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}_h)| \leq \underline{\epsilon}\eta_N; \min_{\boldsymbol{z} \notin \mathcal{T}_{K_0;h}} |\widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{RP}}(\widetilde{\boldsymbol{z}}_h)| > \bar{\epsilon}\eta_N \right\}\right) \\
&- \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^\star\} \\
\geq & 1 - \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^\star\} - 32 \left(\sum_{h=1}^{H_0} |\mathcal{T}_{K_0;h}|\right) |\Gamma_{K_0}| \frac{C \max_{i \in [N], \boldsymbol{z} \in \mathcal{T}} |\breve{Y}_i(\boldsymbol{z})|}{\sqrt{\underline{\lambda}}} \cdot \sqrt{\frac{|\mathbb{M}^\star|}{N_0 Q}}.
\end{aligned}
$$

□

6.4 Proof of Theorem 3

Proof. **Asymptotic normality of $\{\widehat{Y}^\star_{(h)}\}_{h=1}^{H_0}$.** First we prove asymptotic normality for the vector of $\{\widehat{Y}^\star_{(h)}\}_{h=1}^{H_0}$ with

$$
\widehat{Y}^\star_{(h)} = \frac{1}{|\mathcal{T}_{K_0;h}|} \sum_{\boldsymbol{z} \in \mathcal{T}_{K_0;h}} \widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}).
$$

By (6.7), we have

$$
\widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}) = w_{\boldsymbol{z}}^\top \widehat{Y}_{\mathrm{AVG}}, \ w_{\boldsymbol{z}} = Q^{-1} G_{\boldsymbol{z},\mathbb{M}^\star} G_{\mathbb{M}^\star}^\top.
$$

Let $c = \{c_h\}_{h=1}^{H_0}$ be any nonzero vector in $\mathbb{R}^{H_0}$. Then

$$
\sum_{h=1}^{H_0} c_h \widehat{Y}^\star_{(h)} = \sum_{h=1}^{H_0} \sum_{\boldsymbol{z} \in \mathcal{T}_{K_0;h}} \frac{c_h}{|\mathcal{T}_{K_0;h}|} \widehat{Y}^\star_{\mathrm{RP}}(\boldsymbol{z}) = \sum_{h=1}^{H_0} \sum_{\boldsymbol{z} \in \mathcal{T}_{K_0;h}} \frac{c_h}{|\mathcal{T}_{K_0;h}|} w_{\boldsymbol{z}}^\top \widehat{Y}_{\mathrm{AVG}}.
$$

Introduce a matrix $W \in \mathbb{R}^{H_0 \times Q}$ that has

$$
W_{\cdot h} = \sum_{\boldsymbol{z} \in \mathcal{T}_{K_0;h}} \frac{w_{\boldsymbol{z}}}{|\mathcal{T}_{K_0;h}|} = (Q|\mathcal{T}_{K_0;h}|)^{-1} \sum_{\boldsymbol{z} \in \mathcal{T}_{K_0;h}} G_{\mathbb{M}^\star} G_{\boldsymbol{z},\mathbb{M}^\star}^\top.
$$

Then clearly we have

$$
\sum_{h=1}^{H_0} c_h \widehat{Y}^\star_{(h)} = c^\top W^\top \widehat{Y}_{\mathrm{AVG}}.
$$

By Lemma 1, $\sqrt{N_0}(\widehat{Y}_{\mathrm{AVG}} - \overline{Y})$ is asymptotically normal with zero mean and asymptotic variance V_{lim}. Then

$$
\sqrt{N_0} \left\{ c^\top W^\top \widehat{Y}_{\mathrm{AVG}} - c^\top W^\top \overline{Y} \right\} \to N(0, c^\top W^\top V_{\mathrm{lim}} W c).
$$

By the definition of tiers (Definition 1) and Assumption 2, we can easily calculate

$$W_{\cdot h}^{\top}\overline{Y} = \sum_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}} \frac{w_{\boldsymbol{z}}^{\top}\overline{Y}}{|\mathcal{T}_{K_0;h}|} = \sum_{\boldsymbol{z}\in\mathcal{T}_{K_0;h}} \frac{\overline{Y}(\boldsymbol{z})}{|\mathcal{T}_{K_0;h}|} = \overline{Y}_{(h)}.$$

Hence by the Cramér-Wold device, it holds that

$$\sqrt{N_0}\left\{\begin{pmatrix}\widehat{Y}^{\star}_{(1)}\\ \vdots \\ \widehat{Y}^{\star}_{(H_0)}\end{pmatrix} - \begin{pmatrix}\overline{Y}_{(1)}\\ \vdots \\ \overline{Y}_{(H_0)}\end{pmatrix}\right\} \to N(0, W^{\top}V_{\lim}W).$$

High probability of perfect selection. By Theorem 1 and 2, it's not hard to see that under the assumed conditions,

$$\mathbb{P}\{\widehat{\mathbb{M}} = \mathbb{M}^{\star}\} \to 1,\ \mathbb{P}\{\forall\ h \in [H_0] : \widehat{\mathcal{T}}_{K_0;h} = \mathcal{T}_{K_0;h}\} \to 1.$$

Probability adjustment for imperfect selection. We show that

$$\sqrt{N_0}\left\{\begin{pmatrix}\widehat{Y}^{\star}_{(1)}\\ \vdots \\ \widehat{Y}^{\star}_{(H_0)}\end{pmatrix} - \begin{pmatrix}\widehat{Y}_{(1)}\\ \vdots \\ \widehat{Y}_{(H_0)}\end{pmatrix}\right\} \xrightarrow{p} 0.$$

This is due to the simple fact that

$$\mathbb{P}\left\{\sqrt{N_0}\begin{pmatrix}\widehat{Y}^{\star}_{(1)}\\ \vdots \\ \widehat{Y}^{\star}_{(H_0)}\end{pmatrix} - \sqrt{N_0}\begin{pmatrix}\widehat{Y}_{(1)}\\ \vdots \\ \widehat{Y}_{(H_0)}\end{pmatrix} \neq 0\right\} \leq \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^{\star}\} + \mathbb{P}\{\exists\ h \in [H_0] : \widehat{\mathcal{T}}_{K_0;h} \neq \mathcal{T}_{K_0;h}\} \to 0.$$

Now by Slutsky's Theorem we conclude

$$\sqrt{N_0}\left\{\begin{pmatrix}\widehat{Y}_{(1)}\\ \vdots \\ \widehat{Y}_{(H_0)}\end{pmatrix} - \begin{pmatrix}\overline{Y}_{(1)}\\ \vdots \\ \overline{Y}_{(H_0)}\end{pmatrix}\right\} \to N(0, W^{\top}V_{\lim}W).$$

□

6.5 Proof of Lemma 2

Proof. We first cite a result from [19], which gives an estimate of the remainder for the combinatorial central limit theorem:

Lemma 3 (Main theorem of [19]). *Let $A_N = \{A_N(i,j)\}$ be a $N \times N$ matrix of real numbers. Let*

$$a_{i\cdot} = N^{-1}\sum_{j=1}^{N} A_N(i,j),\ a_{\cdot j} = N^{-1}\sum_{i=1}^{N} A_N(i,j),\ a_{\cdot\cdot} = N^{-2}\sum_{i,j=1}^{N} A_N(i,j).$$

and

$$\mu = na_{\cdot\cdot},\ \sigma^2 = (N-1)^{-1}\sum_{i,j}\{A_N(i,j) - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}\}^2.$$

Let further $\widehat{A}_N(i,j) = \sigma^{-1}\{A_N(i,j) - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}\}$. Let $\widehat{\tau} = |\mathcal{T}|^{-1}\sum_{i=1}^{N}\{A_N(i,\pi(i))\}$. There is an absolute constant $C > 0$, such that for all A_N with $\sigma^2 > 0$,

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\tau} - \mathbb{E}\widehat{\tau}}{\sigma} \le t\right\} - \Phi(t)\right| \le \frac{C\sum_{i,j}|\widehat{A}_N(i,j)|^3}{N} \le C\max_{i,j}|\widehat{A}_N(i,j)|.$$

Here $\Phi(\cdot)$ is the CDF for the standard normal distribution.

To apply Lemma 3, we build a population matrix in Table 6.1.

Table 6.1: Population matrix A_N

	t_1	$\cdots$	$\cdots$	$\cdots$	t_q	$\cdots$	$\cdots$
1	$c_1 N_1^{-1} Y_1(1)\cdot \mathbf{1}_{N_1}^\top$	$\cdots$	$\cdots$	$\cdots$	$c_q N_q^{-1} Y_1(q)\cdot \mathbf{1}_{N_q}^\top$	$\cdots$	$\cdots$
2	$c_1 N_1^{-1} Y_2(1)\cdot \mathbf{1}_{N_1}^\top$	$\cdots$	$\cdots$	$\cdots$	$c_q N_q^{-1} Y_2(q)\cdot \mathbf{1}_{N_q}^\top$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$

Now simply observe that

$$\widehat{\gamma}_N = \sum_{i=1}^{N} A_N(i,\pi(i)).$$

This leads to

$$\begin{aligned}
a_{i\cdot} &= N^{-1}\sum_{\boldsymbol{z}\in\mathcal{T}} c(\boldsymbol{z})Y_i(\boldsymbol{z}),\\
a_{\cdot j} &= \{NN(\boldsymbol{z})\}^{-1}c(\boldsymbol{z})\sum_{i=1}^{N} Y_i(\boldsymbol{z}) = N(\boldsymbol{z})^{-1}c(\boldsymbol{z})\overline{Y}(\boldsymbol{z}),\\
a_{\cdot\cdot} &= N^{-2}\sum_{i=1}^{N}\sum_{\boldsymbol{z}\in\mathcal{T}} c(\boldsymbol{z})Y_i(\boldsymbol{z}).
\end{aligned}$$

Now it's not hard to verify that

$$\widehat{A}_N(i,j) = \frac{w(\boldsymbol{z})N(\boldsymbol{z})^{-1}\breve{Y}_i(\boldsymbol{z}) - N^{-1}\sum_{i'=1}^N \breve{\tau}_{i'}}{\sqrt{N^{-1}\sum_{i=1}^N \sum_{\boldsymbol{z}\in\mathcal{T}} N(\boldsymbol{z})\{w(\boldsymbol{z})N(\boldsymbol{z})^{-1}\breve{Y}_i(\boldsymbol{z}) - N^{-1}\sum_{i'=1}^N \breve{\tau}_{i'}\}^2}}.$$

By some simple algebra, we have

$$\begin{aligned}\max_{i,j}|\widehat{A}_N(i,j)| &= \frac{\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|w(\boldsymbol{z})N(\boldsymbol{z})^{-1}\breve{Y}_i(\boldsymbol{z}) - N^{-1}\sum_{i'=1}^N \breve{\tau}_{i'}|}{\sqrt{\mathrm{Var}(\widehat{\gamma}_N)}} \\ &\leq \frac{2\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|w(\boldsymbol{z})N(\boldsymbol{z})^{-1}\breve{Y}_i(\boldsymbol{z})|}{v_N}\end{aligned}$$

Hence Lemma 3 implies that

$$\begin{aligned}&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}_N - \mathbb{E}(\widehat{\gamma}_N)}{v_N}\leq t\right\} - \Phi(t)\right| \\ \leq&\frac{C\underline{c}\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|w(\boldsymbol{z})\{Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\}|}{N_0 v_N} \\ \leq&\frac{C\underline{c}\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|w(\boldsymbol{z})\{Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\}|}{\kappa N_0 v_R} \\ \leq&\frac{C\underline{c}\|w\|_\infty\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\{Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\}|}{\kappa N_0 v_R}.\end{aligned}$$

To determine the scale of v_R, we have

$$\begin{aligned}v_R &= \left\{\sum_{\boldsymbol{z}\in\mathcal{T}} w(\boldsymbol{z})^2 N(\boldsymbol{z})^{-1} S(\boldsymbol{z},\boldsymbol{z})\right\}^{1/2} \\ &\geq \overline{c}{N_0}^{-1/2}\|w\|_2\min_{\boldsymbol{z}\in\mathcal{T}} S(\boldsymbol{z},\boldsymbol{z})^{1/2}.\end{aligned}$$

Summarizing all the parts, we have

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}_N - \mathbb{E}(\widehat{\gamma}_N)}{v_N}\leq t\right\} - \Phi(t)\right| \leq \frac{C(\underline{c},\overline{c},\kappa)\|w\|_\infty}{\|w\|_2\sqrt{N_0}}\cdot\frac{\max_{i\in[N],\boldsymbol{z}\in\mathcal{T}}|\{Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\}|}{\min_{\boldsymbol{z}\in\mathcal{T}} S(\boldsymbol{z},\boldsymbol{z})}.$$

□

Chapter 7

Conclusion and Discussion

To handle the new challenges of selecting best combinations in factorial experiments, such as the well-recognized "winner's curse" phenomenon, the overly large number of treatment groups or the methodological concerns of how to proceed under resource constraints, we proposed a general workflow that takes advantage of several crucial components: (i) forward model selection; (ii) factor level combination selection; (iii) statistical inference over the ties. Theoretically speaking, this framework can achieve model selection consistency, factor level combination selection consistency and post-selection inference in an asymptotic perspective. These statistical properties are further elaborated by several numerical experiments.

There are also several generalization worthy of further discussion and exploration. First, it would be interesting to study whether such a framework also works under a more extreme framework where we have a large number of treatment groups but only limited number of replications within each arm. Second, there are many interesting model selection methods or criteria in super population study, such as sure independence screening [20], which can possibly extend the framework to a more high dimensional regime.

Bibliography

1. Wu, C. J. & Hamada, M. S. *Experiments: planning, analysis, and optimization* (John Wiley & Sons, 2011).
2. Zhao, A. & Ding, P. Regression-based causal inference with factorial experiments: estimands, model specifications, and design-based properties. *arXiv preprint arXiv:2101.02400* (2021).
3. Egami, N. & Imai, K. Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association* (2018).
4. Lee, M. R. & Shen, M. *Winner's curse: Bias estimation for total effects of features in online controlled experiments* in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018), 491–499.
5. Andrews, I., Kitagawa, T. & McCloskey, A. *Inference on winners* tech. rep. (National Bureau of Economic Research, 2019).
6. Dasgupta, T., Pillai, N. S. & Rubin, D. B. Causal inference from 2 K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B: Statistical Methodology,* 727–753 (2015).
7. Pashley, N. E. & Bind, M.-A. C. Causal Inference for Multiple Non-Randomized Treatments using Fractional Factorial Designs. *arXiv e-prints,* arXiv–1905 (2019).
8. Zhao, A. & Ding, P. Covariate-adjusted Fisher randomization tests for the average treatment effect. *Journal of Econometrics* **225,** 278–294 (2021).
9. Wang, H. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104,** 1512–1524 (2009).
10. Hao, N. & Zhang, H. H. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **109,** 1285–1301 (2014).
11. Haris, A., Witten, D. & Simon, N. Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics* **25,** 981–1004 (2016).
12. Hao, N., Feng, Y. & Zhang, H. H. Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association* **113,** 615–625 (2018).

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

14. Bien, J., Taylor, J. & Tibshirani, R. A lasso for hierarchical interactions. *Annals of statistics* **41,** 1111 (2013).

15. Neyman, J. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych* **10,** 1–51 (1923).

16. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66,** 688 (1974).

17. Li, X. & Ding, P. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association* **112,** 1759–1769 (2017).

18. Wasserman, L. & Roeder, K. High dimensional variable selection. *Annals of statistics* **37,** 2178 (2009).

19. Bolthausen, E. An estimate of the remainder in a combinatorial central limit theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **66,** 379–386 (1984).

20. Fan, J. & Lv, J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70,** 849–911 (2008).

Appendix A

A discussion with more general centering values

A.1 Unsaturated weighted least square: a closed form expression

In this section we first derive the closed form expression for unsaturated WLS estimation, then verify the nice targeting property we mentioned in the previous section.

First we need to introduce a transformation matrix $\boldsymbol{P}_{\Delta\delta_{[K]}}$, with columns and rows indexed by subsets $\{\mathcal{K} \subset [K]\}$ of the K factors. Generally it is used to reveal the relationship between designs with different configurations of centering factors $\delta_{[K]}$ and $\delta'_{[K]} = \delta_{[K]} + \Delta\delta_{[K]}$. The transformation is actually linear:

$$\left(f_{\delta'_{[K]}}(z^*_{\mathcal{K}})\right)_{\mathcal{K}\subset[K]} = \left(f_{\delta_{[K]}}(z^*_{\mathcal{K}})\right)_{\mathcal{K}\subset[K]} \boldsymbol{P}_{\Delta\delta_{[K]}}. \tag{A.1}$$

The closed form of $\boldsymbol{P}_{\Delta\delta_{[K]}}$ is easy to derive. Note that for all $\mathcal{K}' \subset [K]$, we have

$$f_{\delta'_{[K]}}(z^*_{\mathcal{K}'}) = \sum_{\mathcal{K}\subset\mathcal{K}'} f_{\delta_{[K]}}(z^*_{\mathcal{K}}) \prod_{k\in\mathcal{K}'\setminus\mathcal{K}} (\Delta\delta)_k,$$

which implies the element of $\boldsymbol{P}_{\Delta\delta_{[K]}}$ indexed by $(\mathcal{K}, \mathcal{K}')$ is given by

$$\boldsymbol{P}_{\Delta\delta_{[K]}}(\mathcal{K}, \mathcal{K}') = \begin{cases} \prod_{k\in\mathcal{K}'\setminus\mathcal{K}}(\Delta\delta)_k & , \quad \mathcal{K} \subset \mathcal{K}', \\ 0 & , \quad \mathcal{K} \subsetneq \mathcal{K}'. \end{cases} \tag{A.2}$$

Define $\boldsymbol{Q}_{\Delta\delta_{[K]}} = \boldsymbol{P}^{-1}_{\Delta\delta_{[K]}}$ to be the inverse. Note that $\boldsymbol{Q}_{\Delta\delta_{[K]}}$ is simply taking out a $\Delta\delta_{[K]}$ vector from a group of centering factors, so by symmetry we have

$$\boldsymbol{Q}_{\Delta\delta_{[K]}}(\mathcal{K}, \mathcal{K}') = \begin{cases} (-1)^{|\mathcal{K}'|-|\mathcal{K}|} \prod_{k\in\mathcal{K}'\setminus\mathcal{K}}(\Delta\delta)_k & , \quad \mathcal{K} \subset \mathcal{K}', \\ 0 & , \quad \mathcal{K} \subsetneq \mathcal{K}'. \end{cases} \tag{A.3}$$

We shall give an example of the above matrix in the three-factor case, which appears (incompletely) in the appendix of [2]. Let $A' = A - \delta_A$, $B' = B - \delta_B$, $C' = C - \delta_C$.

$$\begin{pmatrix} 1 \\ A \\ B \\ C \\ AB \\ AC \\ BC \\ ABC \end{pmatrix} = \boldsymbol{P}_{\Delta\delta_{[K]}}^{\top} \begin{pmatrix} 1 \\ A' \\ B' \\ C' \\ A'B' \\ A'C' \\ B'C' \\ A'B'C' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \delta_A & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \delta_B & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \delta_C & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \delta_A\delta_B & \delta_B & \delta_A & 0 & 1 & 0 & 0 & 0 \\ \delta_A\delta_C & \delta_C & 0 & \delta_A & 0 & 1 & 0 & 0 \\ \delta_B\delta_C & 0 & \delta_C & \delta_B & 0 & 0 & 1 & 0 \\ \delta_A\delta_B\delta_C & \delta_B\delta_C & \delta_A\delta_C & \delta_A\delta_B & \delta_C & \delta_B & \delta_A & 1 \end{pmatrix} \begin{pmatrix} 1 \\ A' \\ B' \\ C' \\ A'B' \\ A'C' \\ B'C' \\ A'B'C' \end{pmatrix}.$$

The following theorem shows that $\boldsymbol{P}_{\Delta\delta_{[K]}}$ and $\boldsymbol{Q}_{\Delta\delta_{[K]}}$ totally determines the structure of $\boldsymbol{D}_h$.

Jibberish 4. *Consider weighted least squares with centering factors $\delta_{[K]}$ and weights proportional to size of each stratum. Let $\Delta\delta_{[K]} = \delta_{[K]} - (1/2)_{k=1}^{K}$. The unsaturated regression on up to all m-level main/interactions terms has coefficient vector:*

$$(\widetilde{\tau}_{\mathcal{K}})_{\{|\mathcal{K}|\le m\}} = (\tau_{\mathcal{K}})_{\{|\mathcal{K}|\le m\}} + \boldsymbol{D}_h \cdot (\tau_{\mathcal{K}})_{\{|\mathcal{K}|>m\}}, \tag{A.4}$$

where $\boldsymbol{D}_h$ is given by

$$\boldsymbol{D}_h = \boldsymbol{P}_{\Delta\delta_{[K]}}(\{\mathcal{K}\subset[m]\},\{\mathcal{K}\subset[m]\}) \cdot \boldsymbol{Q}_{\Delta\delta_{[K]}}(\{\mathcal{K}\subset[m]\},\{\mathcal{K}\subset[K]\backslash[m]\}).$$

Corollary 2. *The matrix $\boldsymbol{D}$ has a closed form expression:*

1. *For $\mathcal{K} \subsetneq \mathcal{K}'$,*

$$\boldsymbol{D}_h(\mathcal{K},\mathcal{K}') = 0. \tag{A.5}$$

2. *For $\mathcal{K} \subset \mathcal{K}'$, let $|\mathcal{K}| = k$, $|\mathcal{K}'| = k'$, with $k \le m < k'$,*

$$\boldsymbol{D}_h(\mathcal{K},\mathcal{K}') = \sum_{l=0}^{m-k}(-1)^{k'-k+1-l}\binom{k'-k+1}{l}\prod_{t\in\mathcal{K}'\backslash\mathcal{K}}\left(\delta_t - \frac{1}{2}\right). \tag{A.6}$$

Proof. This result can be derived through careful calculation based on the definition of $\boldsymbol{P}$ and $\boldsymbol{Q}$ from (A.2) and (A.3) along with Theorem 4 thus omitted here. □

A.2 A sufficient condition for sign consistency in population WLS regression

Definition 3 (Active interaction number)**.** *For every z_k of the K factors, there are s_k factors that have nonzero interaction with z_k, where $s_k \in [K-1]$ is a nonnegative integer associated with K. We call s_k the active interaction number of factor z_k. The maximal active interaction number is subsequently defined as $s_K = \max_{k\in[K]} s_k$.*

This definition is mainly devoted to finer technical purposes in Theorem 5.

Jibberish 5. *Assume we run weighted least square under the setting depicted in Theorem 4. Define the maximal decaying rate $c_K = \max_{l\in[K]} c_l$. Recall the predefined maximal active interaction number s_K from Definition 3. If we have*

$$s_K c_K \max_{k=1,\dots,K} |\delta_k - 1/2| < \ln 2, \tag{A.7}$$

then the unsaturated regression coefficients $(\widetilde{\tau}_{\mathcal{K}})_{\{|\mathcal{K}|\le m\}}$ and the corresponding saturated regression coefficients $(\tau_{\mathcal{K}})_{\{|\mathcal{K}|\le m\}}$ from (A.4) *have same signs on every term.*

Condition (A.7) unifies the property of factorial effects and the information of the design pattern(the centering factors $\delta_{[K]}$). The product of s_K and c_K demonstrates a trade-off between the active interaction number and the hierarchy structure. Sparser interactions require slower decaying rate and vice versa. Besides, the product of $s_K c_K$ and $\max_{k=1,\dots,K} |\delta_k - 1/2|$ shows that if δ_k lies more close to $1/2$, less restriction are needed on the effect structure. This aligns with the result in [2]: when $\delta_k = 1/2$ holds for all $k = 1, \dots, K$, $\boldsymbol{D}_h = \boldsymbol{0}$, so that forward selection always works.

A.3 Proofs

Proof of Theorem 4

Proof. Let $\boldsymbol{F}_+$ be the design matrix for up to all m-level terms, and $\boldsymbol{F}_-$ for the remaining ones. The Cochran's Theorem implies

$$\boldsymbol{D}_h = (\boldsymbol{F}_+^\top \boldsymbol{W} \boldsymbol{F}_+)^{-1} (\boldsymbol{F}_+^\top \boldsymbol{W} \boldsymbol{F}_-)$$

Now we can calculate

$$\boldsymbol{F}_+^\top \boldsymbol{W} \boldsymbol{F}_+ = \sum_{i=1}^N w_i f_{+,i} f_{+,i}^\top = \sum_{z\in\mathcal{T}} w(z) N(z) f_+(z) f_+(z)^\top = N \sum_{z\in\mathcal{T}} f_+(z) f_+(z)^\top,$$
$$\boldsymbol{F}_+^\top \boldsymbol{W} \boldsymbol{F}_- = \sum_{i=1}^N w_i f_{+,i} f_{-,i}^\top = \sum_{z\in\mathcal{T}} w(z) N(z) f_+(z) f_-(z)^\top = N \sum_{z\in\mathcal{T}} f_+(z) f_-(z)^\top.$$

Here $f(z) = (f_+(z)^\top, f_-(z)^\top)^\top$ are common vector for the individuals under some treamment $z \in \mathcal{T}$. This suggests for calculating $\boldsymbol{D}_h$ we can need to consider a balanced design where each treatment has only one individual, coded by $f(z)$. Proposition S1 of [2] asserts that when $\delta_k = 1/2$, for $k = 1, \dots, K$, $\boldsymbol{D}_m = \boldsymbol{0}$. Now we use a "0" in the subscript to indicate the designs and quantities in this special case. For example, $\boldsymbol{F}_0 = (\boldsymbol{F}_{+,0}, \boldsymbol{F}_{-,0})$ is the design matrix with centering factors $1/2$, and we have $\boldsymbol{D}_{m,0} = \boldsymbol{0}$.

Let $(\Delta\delta_k)_{[K]} = (\delta_k - 1/2)_{[K]}$, then we have $\boldsymbol{F}_0 = \boldsymbol{F}\boldsymbol{P}_{\Delta\delta_{[K]}}$. Since $(\Delta\delta_k)_{[K]}$ is clearly defined we omit it from the subscript of $\boldsymbol{P}_{\Delta\delta_{[K]}}$ and simply write $\boldsymbol{P}$. $\boldsymbol{P}$(and $\boldsymbol{Q}$ analogously) naturally breaks down to sub-blocks with dimension compatible with $\boldsymbol{F}_{+,0}$ and $\boldsymbol{F}_{-,0}$:

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_+ & \boldsymbol{P}_\pm \\ \boldsymbol{0} & \boldsymbol{P}_- \end{pmatrix}, \ \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_+ & \boldsymbol{Q}_\pm \\ \boldsymbol{0} & \boldsymbol{Q}_- \end{pmatrix}.$$

To obtain $\boldsymbol{D}_h$, we proceed as follows: (i) Regress $\boldsymbol{F}_-$ on $\boldsymbol{F}_{+,0}$ to obtain the coefficient matrix $\widetilde{\boldsymbol{D}}_m$; (ii) Transform $\widetilde{\boldsymbol{D}}_m$ to $\boldsymbol{D}_h$ using $\boldsymbol{P}$.

For Step (i), using $\boldsymbol{Q}$, we have

$$\boldsymbol{F}_- = \boldsymbol{F}_{+,0}\boldsymbol{Q}_\pm + \boldsymbol{F}_{-,0}\boldsymbol{Q}_-.$$

Now by Cochran's Theorem we have

$$\widetilde{\boldsymbol{D}}_m = \boldsymbol{Q}_\pm + \boldsymbol{D}_{m,0}\boldsymbol{Q}_- = \boldsymbol{Q}_\pm.$$

For Step (ii), note $\boldsymbol{F}_{+,0} = \boldsymbol{F}_+\boldsymbol{P}_+$ and $\boldsymbol{P}$ is non-degenerate, we conclude that $\boldsymbol{D}_h = \boldsymbol{P}_+\widetilde{\boldsymbol{D}}_m = \boldsymbol{P}_+\boldsymbol{Q}_\pm$.

□

Proof of Theorem 5

Proof. We proceed in two steps. Recall we run WLS with up to m-level effect/interaction terms.

1. We show that for $m = 1, \ldots, K$, the signs of all the main effect terms are preserved. Without loss of generality our analysis is wrapped for the first factor $\{1\}$.
2. We show that the sign of the higher interaction term is preserved and this can be guaranteed with an elegant adaptation from the main factor analysis.

For Step 1 with a given m, using the classical Cochran's Theorem [2], we have

$$\widetilde{\tau}_{+,1} = \tau_{+,\{1\}} + \boldsymbol{D}_h(\{1\}, :) \cdot \tau_-.$$

Denote $\boldsymbol{D}_h(\{1\}, :) \cdot \tau_-$ as $\mathbf{Bias}_m(\{1\})$. By Theorem 4, this gives

$$\begin{aligned} \mathbf{Bias}_m(\{1\}) &= \sum_{\substack{\mathcal{K}':|\mathcal{K}'|=k'>m, \\ \{1\}\subset\mathcal{K}'\subset[K]}} \boldsymbol{D}_m(\{1\}, \mathcal{K}')\tau_{-,\mathcal{K}'} \\ &= \sum_{\substack{\mathcal{K}':|\mathcal{K}'|=k'>m, \\ \{1\}\subset\mathcal{K}'\subset[K]}} \tau_{-,\mathcal{K}'} \left(\sum_{l=0}^{m-1} (-1)^{k'-l} \binom{k'}{l} \prod_{t\in\mathcal{K}'\setminus\{1\}} \left(\delta_t - \frac{1}{2}\right) \right). \end{aligned}$$

In light of the definition of s_K (Definition 3), at most s_K factors has nonzero interactions with $\{1\}$. Without loss of generality we assume these s_K factors are $\{2\}, \ldots, \{s_K+1\}$. Now instead of taking summation over all $\mathcal{K}' \subset [K]$, we only need those from $[s_K+1]$, since by the hierarchy condition $\tau_{-,\mathcal{K}'} = 0$ for $\mathcal{K}' \subsetneq [s_K+1]$. Now let $\delta_{max} = \max_{t=1,\ldots,K} |\delta_t - 1/2|$. We bound $\widetilde{\tau}_{+,\{1\}}$ as follows:

$$|\mathbf{Bias}_m(\{1\})| \leq \sum_{\substack{\mathcal{K}':|\mathcal{K}'|=k'>m,\\ \{1\}\subset\mathcal{K}'\subset[s_K+1]}} |\tau_{-,\mathcal{K}'}| \delta_{max}^{k'-1} \left| \sum_{l=0}^{m-1} (-1)^{k'-l} \binom{k'}{l} \right| \tag{A.8}$$

$$\leq \sum_{\substack{\mathcal{K}':|\mathcal{K}'|=k'>m,\\ \{1\}\subset\mathcal{K}'\subset[s_K+1]}} c_K^{k'-1} |\tau_{+,\{1\}}| \delta_{max}^{k'-1} \left| \sum_{l=0}^{m-1} (-1)^{k'-l} \binom{k'}{l} \right|. \tag{A.9}$$

(A.8) is derived under the sparsity condition and (A.9) comes from the hierarchy condition. Further, we apply the following facts in combinatorics to (A.9):

$$\left| \sum_{l=0}^{m-1} (-1)^{k'-l} \binom{k'}{l} \right| = \binom{k'-1}{m-1}, \quad \binom{s_K}{k'} \binom{k'-1}{m-1} = \binom{s_K-m+1}{k'-m+1} \binom{s_K}{m-1}.$$

This leads to

$$\begin{aligned}
|\mathbf{Bias}_m(\{1\})| &\leq |\tau_{+,\{1\}}| \sum_{k'=m+1}^{s_K+1} \binom{s_K}{k'-1} c_K^{k'-1} \delta_{max}^{k'-1} \binom{k'-1}{m-1} \\
&\leq |\tau_{+,\{1\}}| \binom{s_K}{m-1} \sum_{k'=m+1}^{s_K+1} \binom{s_K-m+1}{k'-m} c_K^{k'-1} \delta_{max}^{k'-1} \\
&= |\tau_{+,\{1\}}| \binom{s_K}{m-1} c_K^{m-1} \delta_{max}^{m-1} \sum_{t=1}^{s_K-m+1} \binom{s_K-m+1}{t} c_K^t \delta_{max}^t \\
&= |\tau_{+,\{1\}}| \binom{s_K}{m-1} c_K^{m-1} \delta_{max}^{m-1} \left\{ (1+c_K\delta_{max})^{s_K-m+1} - 1 \right\}.
\end{aligned}$$

Now Condition (A.7) ensures $s_K c_K \delta_{max} < \ln 2$, implying

$$\binom{s_K}{m-1} c_K^{m-1} \delta_{max}^{m-1} \leq (s_K c_K \delta_{max})^{m-1} < 1,$$

and

$$(1+c_K\delta_{max})^{s_K-m+1} - 1 = \exp((s_K-m+1)\ln(1+c_K\delta_{max})) - 1 \leq \exp(s_K c_K \delta_{max}) - 1 < 1.$$

Therefore, $|\mathbf{Bias}_m(\{1\})| \leq c|\tau_{+,\{1\}}|$ for some postive constant $c < 1$, suggesting that the induced bias does not twist the sign of $\tau_{+,\{1\}}$.

It remains to complete Step 2. Beyond the main effect terms, the analysis for interactions can be carried out in an exactly same way, due to the the nice targeting property(Theorem 4 and Corollary 2) and the fact that we model the hierarchy structure using one unified decaying rate c_K. For example, in a four-factor case, when calculating the potential bias for AB, the targeting property of $\boldsymbol{D}$ guarantees that we only need to consider contribution from ABC, ABD, and $ABCD$. This works similarly as bounding bias for B from higher-level interactions BC, BD and BCD. Thus without any further technical efforts, we conclude that all signs of the included terms(both main effect and their interactions) are preserved.

□