Forward factorial screening with statistical inference

Abstract

Modern factorial designs may involve a large number of factors and limited number of replications in each treatment level, which leads to a large volume of causal parameters and invalidates classical inference results. We propose a forward factorial screening procedure, which identify salient factorial effects in a forward fashion following the natural heredity structure of factorial experiments. Computationally, the proposed procedure is simple to implement and is compatible with many existing model selection methods. Theoretically, we prove the forward screening procedure selects the non-zero effects consistently in asymptotic regimes. Furthermore, by combining the proposed forward screening framework with factor-based regressions, we construct point estimators for general causal effects in factorial experiments and propose variance estimation schemes, which can incorporate information from multiple treatment levels and improve asymptotic efficiency for certain choices of target effects. Our numerical study illustrates the finite sample performance of the proposed screening and inferential frameworks.

Keywords: causal inference; potential outcomes; design-based inference; heredity; interaction; variable selection

1 Introduction

1.1 Motivation

Factorial designs have been widely utilized in many fields, ranging from agricultural, industrial, to biomedical sciences (Wu and Hamada, 2011; Zhao and Ding, 2021b; Egami and Imai, 2018). The popularity of factorial designs partially ascribed to their ability to simultaneously accommodate multiple factors and provide informative assessments for the magnitudes of main causal effects and their interactions.

Statistical inference in classical factorial experiments has been well understood when the number of factors, denoted as K, is a fixed number. However, in modern factorial designs involving a large number of factors, it is debatable if the fixed K assumption accurately reflects the data collection

mechanism, and classical statistical inference in factorial experiments faces two key challenges. On the one hand, in an ocean of causal parameters (e.g., when $K = 10$, we have $2^K - 1 = 1,023$ factorial effects), practitioners may hope to target on a few salient ones. The choice of these target parameters need to be carefully made not only to guarantee the validity of statistical inference but also to ensure the selected parameters respect the design principles of factorial experiments. On the other hand, it is unknown how the selection step affects estimation and inference for general causal effects.

To address these challenges, we aim to leverage three unique structural principles first capitalized by Wu and Hamada (2011), and propose a forward factor screening procedure that screens out noisy factors in a principled fashion. These three principles are:

- *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.

Forward screening can naturally incorporate these principles and produce a few interpretable parameters that respect the structure of factorial experiments. This procedure will allow people to work on an inherent parameter space with lower dimension, combine information across multiple treatment levels and mitigate the impact of limited replications due to the large Q setup. Moreover, we can take advantage of such benefits to make inference on many general interesting causal quantities.

1.2 Our contribution

This article makes several contributions. First, we propose a forward factorial screening procedure, which identifies salient factorial effects in a forward fashion following the general principles of factorial experiments. The procedure produces a working model with high interpretability because it fully respects the hierarchical structure of factorial experiments. Besides, combined with factor-based regressions, the proposed forward screening framework is easy to implement and demonstrates high computational efficiency. Theoretically, we also justify the screening consistency property of the forward screening procedure in asymptotic regimes. Second, we did thorough efficiency

comparison for inference using factor based regression with/without effect screening, both from theoretical justification and numerical experimentation. We show that effect screening can bring a great efficiency gain for causal effects defined by sparse contrasts while maintaining desirable performance for general contrast. Moreover, we extend the analysis to comparison of multiple causal parameters. Third, we propose strategies that are practically appealing in marginal cases where the effect sizes are small and perfect screening is hard to achieve. We introduce an under-screening strategy that excludes high order interactions and an over-selection strategy that includes high order interactions following the heredity principle. These two strategies can circumvent the potential challenges of under-powered effect size and deliver valid inference.

1.3 Literature review

In the realm of factorial experiments, the factor-based regression typically serves as a dominant strategy for delivering point estimators and confidence regions, due to its simplicity and flexibility in real-life applications. For example, Dasgupta et al. (2015) extended the classical notion of factorial effects to causal counterparts by introducing potential outcome framework and contrast designs. Zhao and Ding (2021b) 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. Pashley and Bind (2019) 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. Zhao and Ding (2021a) 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 screening. 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, Wang (2009) introduced forward regression for main effects screening and proves its screening consistency property. Hao and Zhang (2014) 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. Haris et al. (2016) considered convex modelling of the factorial effects estimation and introduces strong heredity condition to achieve adaptive selection. Hao et al.

(2018) utilized a regularization scheme to tackle the curse of high dimensionality and perform valid variable screening with quadratic regression. Other works including Lim and Hastie (2015); Bien et al. (2013), proposed procedures for learning interactions based on ℓ_1 regularized least squares based on a purely algorithmic perspective without statistical guarantee.

1.4 Notations

We adopt the following notations throughout the manuscript. For asymptotic analyses, $a_N = O(b_N)$ denotes that there exists a positive constant $C > 0$ such that $a_N \leq Cb_N$. $a_N = o(b_N)$ denotes that $a_N/b_N \to 0$ as N goes to infinity. $a_N = \Theta(b_N)$ denotes that there exists positive constants c and C such that $cb_N \leq a_N \leq Cb_N$. For a matrix V, define $\varrho_{\max}(V), \varrho_{\min}(V)$ as the largest and smallest eigenvalue, respectively, and use $\kappa(V) = \varrho_{\max}(V)/\varrho_{\min}(V)$ to denote its condition number. For two positive semi-definite matrices V_1 and V_2, we use $V_1 \preccurlyeq V_2$ to indicate V_1 is dominated by V_2.

For analyzing factorial effects, we work with different levels of sets. For an integer K, let $[K] = \{1, \cdots, K\}$. We use $\mathcal{K}$ in calligraphic to denote a subset of $[K]$. For subsets of the power set of $[K]$, we use blackboard bold font for presentation,. For example, we denote $\mathbb{M} \subset \{\mathcal{K} \mid \mathcal{K} \subset [K]\}$ and denote the power set of $[K]$ as $\mathbb{K}$.

2 Factorial experiment: A general setup

In this section, we introduce four key components and their corresponding notations in factorial experiments, including the general framework of factorial experiments, potential outcomes and treatment assignment mechanisms, the definition of various causal effects, and factor based regressions. To demonstrate how our setup can be conveniently integrated in factorial experiments, we end the section with a concrete example (Example 1) with three factors.

First, we introduce the general framework of a 2^K factorial experimental design, for some $K \geq 2$. In this design, we have K factors each with binary levels, denoted by $z_k \in \{0, 1\}$, $k = 1, \cdots, K$. Let the vector $\boldsymbol{z}_{\mathcal{K}} = (z_k)_{k \in \mathcal{K}}$ denotes the treatment combining the factors in the set $\mathcal{K} \subset [K] = \{1, \ldots, K\}$; let $\boldsymbol{z}$ denotes the treatment combining all K factors (we drop the subscript $[K]$ in $\boldsymbol{z}_{[K]}$ for simplicity). The K factors in total define a collection of $Q = 2^K$ treatment combinations, and we denote the set of of these treatment combinations as

$$\mathcal{T} = \{\boldsymbol{z} = (z_1 \cdots z_K) \mid z_k \in \{0, 1\}, k = 1, \cdots, K\}, \quad |\mathcal{T}| = Q.$$

Second, we introduce potential outcomes and treatment assignment mechanisms in a factorial

experiment. Suppose N units are enrolled in a factorial experiment, with $N(\boldsymbol{z})$ units assigned the treatment $\boldsymbol{z} \in \mathcal{T}$. Unit i has potential outcome $Y_i(\boldsymbol{z})$ if assigned to the treatment $\boldsymbol{z}$. As each unit has Q potential treatment combinations, we aggregate these Q potential outcomes for unit i into a vector $\boldsymbol{Y}_i = \{Y_i(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}}$ using lexicographic order. The potential outcomes have mean $\overline{Y} = (\overline{Y}(\boldsymbol{z}))_{\boldsymbol{z}\in\mathcal{T}}$ and covariance $S = (S(\boldsymbol{z}, \boldsymbol{z}'))_{\boldsymbol{z},\boldsymbol{z}'\in\mathcal{T}}$ with elements defined as follows:

$$\overline{Y}(\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^{N} Y_i(\boldsymbol{z}), \quad S(\boldsymbol{z}, \boldsymbol{z}') = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z}))(Y_i(\boldsymbol{z}') - \overline{Y}(\boldsymbol{z}')).$$

Let Z_i encode the observed treatment for unit i, and the probability of observing this treatment combination is

$$\mathbb{P}\{Z_i = \boldsymbol{z}, i \in [N], \boldsymbol{z} \in \mathcal{T}\} = \frac{\prod_{\boldsymbol{z}\in\mathcal{T}} \big(N(\boldsymbol{z})\big)!}{N!}.$$

The observed outcome is $Y_i = Y_i(Z_i) = \sum_{\boldsymbol{z}\in\mathcal{T}} Y_i(\boldsymbol{z})\mathbf{1}\{Z_i = \boldsymbol{z}\}$. Furthermore, we use N_i to denote the number of units for the treatment group in which the i-th individual is assigned.

Third, we formally define treatment effects frequently used in factorial experiments, and our definition is in line with the existing literature Dasgupta et al. (2015); Zhao and Ding (2021b); Wu and Hamada (2011). For a subset $\mathcal{K} \subset [K] = \{1, \ldots, K\}$ of the K factors, we use the following "contrast vector" notation for better defining the causal effects later. To start with, we define the main causal effect for factor k. To do so, for a treatment level $\boldsymbol{z} = (z_1, \ldots, z_K) \in \mathcal{T}$, we use $g_{\{k\}}(\boldsymbol{z}) = 2z_k - 1$ to denoted the "centered" treatment indicator z_k. We then define a Q-dimensional contrast vector $g_{\{k\}}$ by aggregating these centered treatment variables into a vector using lexicographic order, that is

$$g_{\{k\}} = \{g_{\{k\}}(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}}, \text{ where } g_{\{k\}}(\boldsymbol{z}) = 2z_k - 1. \tag{2.1}$$

Next, for the interaction causal effect of multiple factors in the case with $|\mathcal{K}| \geq 2$, the contrast vector $g_{\mathcal{K}} \in \mathbb{R}^{Q\times 1}$ can be similarly defined as

$$g_{\mathcal{K}} = \{g_{\mathcal{K}}(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}}, \text{ where } g_{\mathcal{K}}(\boldsymbol{z}) = \prod_{k\in\mathcal{K}} g_{\{k\}}(\boldsymbol{z}). \tag{2.2}$$

As a special case, when no factor is considered, we define $g_{\varnothing} = \mathbf{1}_Q$.

Equipped with the contrast vector notation, we are ready to introduce the main effects and the k-way interaction causal effects in factorial experiments. More concretely, the main causal effect of a single factor and the k-way interaction causal effect of multiple factors ($k \geq 2$) are defined by the inner product of the contrast vector $g_{\mathcal{K}}$ and the averaged potential outcome $\overline{Y}$, that is

$$\tau_{\mathcal{K}} = Q^{-1} \cdot g_{\mathcal{K}}^{\top}\overline{Y}, \quad \mathcal{K} \subset [K]. \tag{2.3}$$

We call the effect $\tau_{\mathcal{K}}$ a *parent* of $\tau_{\mathcal{K}'}$ if $\mathcal{K} \subset \mathcal{K}'$ and $|\mathcal{K}| = |\mathcal{K}'| - 1$.

To more conveniently summarize these causal parameters of interest using synthesized notation, we stack the contrast vectors into an orthogonal matrix $G \in \mathbb{R}^{Q \times Q}$ (i.e., $G^\top G = Q \cdot I_Q$), in which the first column is each column is $g_\varnothing$ and the rest of the columns are contrast vectors $g_{\mathcal{K}}$'s. We refer to matrix G as the contrast matrix. The entire collection of causal parameters in factorial experiments can be more handily written as

$$\tau = (\tau_{\mathcal{K}})_{\mathcal{K} \subset [K]} = Q^{-1} \cdot G^\top \overline{Y}, \text{ where } G = (g_{\mathcal{K}})_{\mathcal{K} \subset [K]}. \tag{2.4}$$

As an additional notation which can be useful for our future derivation, we use $\tau_\varnothing = Q^{-1} g_\varnothing^\top \overline{Y}$ to capture the total average of potential outcomes. We realize that the above notations can be rather abstract, and we have provided an illustrative example in Example 1 with three factors $K = 3$.

Fourth, to estimate the above causal parameters, we introduced two types of factor-based regression commonly adopted in the literature Wu and Hamada (2011); Dasgupta et al. (2015); Zhao and Ding (2021a). The first type is the *saturated regression* in which we regress the observed outcome Y_i on the regressor t_i. Here, t_i is one row vector in the contrast matrix G indexed by the unit i's observed treatment status $Z_i = (z_{i,1}, \ldots, z_{i,K})^\top$; or equivalently, we can write $t_i = \{g_{\mathcal{K}}(Z_i)\}_{\mathcal{K} \subset [K]} \in \mathbb{R}^{Q \times 1}$, where $g_{\mathcal{K}}(Z_i) = \prod_{k \in \mathcal{K}} g_{\{k\}}(\boldsymbol{z})$ is defined in (2.2). The second type of the *unsaturated regression* in which we regress the observed outcome Y_i on a subvector of t_i. We denote the subvector of t_i as $t_{i,\mathbb{M}}$, and $\mathbb{M}$ is the subset of the power set of all factor combinations (i.e., $\mathbb{M} \subset \mathbb{K} = \{\mathcal{K} \mid \mathcal{K} \subset [K]\}$). Throughout this manuscript, we refer to any subset of $\mathbb{K}$, $\mathbb{M}$, as a *working model*. The estimated coefficients of the unsaturated regression is denoted as $\widehat{\tau}(\mathbb{M})$, and the collection of true causal effects in the set $\mathbb{M}$ is denoted as $\tau(\mathbb{M})$, i.e., $\tau(\mathbb{M}) = \{\tau_{\mathcal{K}}\}_{\mathcal{K} \in \mathbb{M}}$. Moreover, we use $G(\cdot, \mathbb{M})$ to denote the columns in G indexed by $\mathbb{M}$.

In what follows, we provide a concrete example of our notations with three factors in the uniform factorial experiment. By uniform, we mean that each treatment level contains the same number of units.

Example 1 (An explanation in the uniform 2^3 factorial design)**.** *Suppose we have three binary factors z_1, z_2 and z_3. This three factors amount to 8 treatment combinations, indexed by a triplet $(z_1 z_2 z_3)$ with $z_1, z_2, z_3 \in \{0, 1\}$, in the set*

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

There are $N = \sum_{z_1, z_2, z_3} N(z_1 z_2 z_3) = 2^3 N_0$ units from this design, where $N(z_1 z_2 z_3) = N_0$ denotes the group size under treatment $(z_1 z_2 z_3)$ under the uniform design.

Each unit i corresponds to a potential outcome vector $\mathbf{Y}_i = \{Y_i(z_1z_2z_3)\}_{z_1,z_2,z_3=0,1}^{\top}$. The vector of causal parameters in factorial experiment is

$$\tau = \frac{1}{2^3} G^{\top} \overline{Y} \triangleq \left(\tau_{\varnothing}, \tau_{\{1\}}, \tau_{\{2\}}, \tau_{\{3\}}, \tau_{\{1,2\}}, \tau_{\{1,3\}}, \tau_{\{2,3\}}, \tau_{\{1,2,3\}}\right)^{\top},$$

where G is the contrast matrix

$$G = \begin{array}{c} \\ (000) \\ (001) \\ (010) \\ (011) \\ (100) \\ (101) \\ (110) \\ (111) \end{array} \begin{array}{c} \begin{array}{cccccccc} \tau_{\varnothing} & \tau_{\{1\}} & \tau_{\{2\}} & \tau_{\{3\}} & \tau_{\{1,2\}} & \tau_{\{1,3\}} & \tau_{\{2,3\}} & \tau_{\{1,2,3\}} \end{array} \\ \begin{pmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}.$$

We observe the pair (Y_i, Z_i) for the unit i, where $Z_i = (z_{i,1}, z_{i,2}, z_{i,3})$ is the observed treatment combinations. Let $g_{\{k\}}(Z_i) = 2z_{i,k} - 1$ be the centered version of $z_{i,k}$. For the factor-based regression, the covariate t_i can be constructed from Z_i:

$$t_i = \Big[1,\ g_{\{1\}}(Z_i),\ g_{\{2\}}(Z_i),\ g_{\{3\}}(Z_i),\ g_{\{2,3\}}(Z_i),\ g_{\{1,3\}}(Z_i),\ g_{\{1,2\}}(Z_i),\ g_{\{1,2,3\}}(Z_i)\Big].$$

For instance, when $Z_i = (101)$, t_i corresponds to the row (101) of the contrast matrix G. Then, a saturated regression is to regress Y_i on t_i:

$$Y_i \sim t_i.$$

For the unsaturated regression, if we only include indices $\varnothing$ (the intercept), $\{1\}, \{1,2\}, \{1,3\}, \{123\}$ in our regression, we can form the working model $\mathbb{M} = \{\varnothing, \{1\}, \{1,2\}, \{1,3\}, \{1,2,3\}\}$ and perform

$$Y_i \sim t_{i,\mathbb{M}},\ \textit{where}\ t_{i,\mathbb{M}} = \Big[1,\ g_{\{1\}}(Z_i),\ g_{\{1,3\}}(Z_i),\ g_{\{1,2\}}(Z_i),\ g_{\{1,2,3\}}(Z_i)\Big].$$

3 Forward screening in factorial experiments

In factorial experiments, when the number of factors K is small (hence small Q), classical statistical inference for factorial effects has been well studied Zhao and Ding (2021b). While modern factorial designs often involve a rather large number of factors, we propose a principled procedure to carefully decide an unsaturated working model $\mathbb{M}$ and scientifically screen out potential noisy factors, followed by an investigation of its statistical property.

3.1 Proposed procedure

In this section, we introduce a principle forward screening procedure that not only fully respects the effect hierarchy, sparsity and heredity principles but also results in an interpretable parsimonious model with statistical guarantees. The forward screening procedure is summarized in Algorithm 1 below:

Algorithm 1: Forward factorial screening

Input: Factorial data $\{(Y_i, Z_i)\}_{i=1}^N$; predetermined integer D; initial working model $\widehat{\mathbb{M}} = \{\varnothing\}$; significance level $\{\alpha_d\}_{d=1}^D$.

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

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

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

3 Update intermediate working model to include all the d-order (interaction) terms: $\widehat{\mathbb{M}}' = \widehat{\mathbb{M}} \cup \{\mathcal{K} \mid |\mathcal{K}| = d\} \triangleq \widehat{\mathbb{M}} \cup \mathbb{K}_d$.

4 Screen out indices in $\widehat{\mathbb{M}}'$ according to either the weak/strong heredity principles defined in (3.5) and (3.6), and denote the screened working model as $\widehat{\mathbb{M}}'$.

5 Run the unsaturated factor-based regression on the working model $\widehat{\mathbb{M}}'$:

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

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

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

7 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 $|\mathcal{K}| = d$.

8 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 $\min\{\alpha_d/(|\widehat{\mathbb{M}}'| - |\widehat{\mathbb{M}}|), 1\}$ and remove the non-significant terms from $\widehat{\mathbb{M}}' \backslash \widehat{\mathbb{M}}$.

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

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

In what follows, we illustrate why the proposed procedure in Algorithm 1 respect the three fundamental principles in factorial experiments:

First, our algorithm adheres to the "Effect Hierarchy Principle" as it performs factor screening in a forward style (coded in the global loop from $d = 1$ to $d = D$, Step 2 in particular). More

concretely, we begin with an empty working model. We then select relevant main effects (Steps 4 and 8) and add them into the working model. Once the working model is updated, we continue to select relevant higher order interaction effects in a forward fashion. Such a forward screening procedure is again motivated by the hierarchy principle that lower-order effects are more important than higher-order ones.

Second, our algorithm operates under the "Effect Sparsity Principle" as it removes potentially unimportant effects using marginal t-test with Bonferroni correction (Step 8) This step induces a sparse working model and helps us to identify essential factorial effects, echoing the "Effect Sparsity Principle."

Third, our algorithm incorporates the "Effect Heredity Principle" as it screens out the interaction effects, when either none of their parent effects is included (weak heredity) or some of their patient effects are excluded (strong heredity) in the previous working model (Step 4). Here, by strong and weak heredity (Hao and Zhang, 2014; Lim and Hastie, 2015), mathematically speaking, we mean that:

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

$$\mathcal{K}' \notin \widehat{\mathbb{M}}' \text{ for } \textbf{all } \mathcal{K}' \subset \mathcal{K},\ |\mathcal{K}'| = |\mathcal{K}| - 1. \tag{3.5}$$

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

$$\mathcal{K}' \notin \widehat{\mathbb{M}}' \text{ for } \textbf{some } \mathcal{K}' \subset \mathcal{K},\ |\mathcal{K}'| = |\mathcal{K}| - 1. \tag{3.6}$$

Lastly, we note that our forward screening procedure enhances the interpretability of the selected working model by iterating between the "Sparsity-screening" step (shorted as S-step in the rest of the manuscript), captured by a data-dependent operator $\widehat{\mathtt{S}} = \widehat{\mathtt{S}}(\cdot; \{Y_i, Z_i\}_{i=1}^N)$, and the "Heredity-screening" step (shorted as H-step in the rest of the manuscript), captured by a deterministic operator $\mathtt{H} = \mathtt{H}(\cdot)$. Because the working model is updated in an iterative fashion,

$$\widehat{\mathbb{M}}_1 \xrightarrow{\mathtt{H}} \cdots \xrightarrow{\widehat{\mathtt{S}}} \widehat{\mathbb{M}}_{d-1} \xrightarrow{\mathtt{H}} \widehat{\mathbb{M}}_{d,+} \xrightarrow{\widehat{\mathtt{S}}} \widehat{\mathbb{M}}_d \to \cdots \xrightarrow{\widehat{\mathtt{S}}} \widehat{\mathbb{M}}_D. \tag{3.7}$$

the final working model includes a small number of statistically significant effects that fully respect the heredity principle.

3.2 Forward screening consistency

We are now ready to analyze the screening consistency property of Algorithm 1. To avoid redundancy, we focus on the algorithm that carries out the S-step via the marginal t-test. We shall show

that the proposed algorithm selects the "true working model" up to level D with probability tending to one as the sample size goes to infinity. Here, the true working model at level $k \in [K]$, denoted as $\mathbb{M}_k^\star$, is the collection of $\mathcal{K}$'s where $|\mathcal{K}| = k$ and $\tau_{\mathcal{K}} \neq 0$. While the key technique in our proof is to utilize Berry-Esseen bounds to analyze the statistical property of moment estimators for factorial effects, we defer interested readers for Supplementary Materials for details.

We start by introducing the following condition on *nearly uniform design*:

Condition 1 (Nearly uniform design)**.** *There exists an positive integer $N_0 > 0$ and absolute constants $\underline{c} \leq \bar{c}$, such that*

$$N(\boldsymbol{z}) = c(\boldsymbol{z})N_0 \geq 2, \text{ where } \underline{c} \leq c(\boldsymbol{z}) \leq \bar{c}.$$

This condition generalizes the classical assumption in the fixed Q regime (meaning that Q is fixed and N tends to infinity)

$$Q \text{ is fixed, and } N(\boldsymbol{z})/N \to e(\boldsymbol{z}) \in (0,1), \text{ as } N \to \infty, \tag{3.8}$$

where $e(\boldsymbol{z})$ is a positive constant.

Next, we quantify the order of the size of the true effects $\tau_{\mathcal{K}}$'s and the tuning parameters α_d's adopted in Bonferroni correction, and we allow them to change with the sample size N:

Condition 2 (Order of parameters)**.** *The true parameters and tuning parameters have the following order:*

(i) True parameter: $|\tau_{\mathcal{K}}| = \Theta(N^\delta)$ for some $-1/2 < \delta \leq 0$ and all $\mathcal{K} \in \mathbb{M}^\star$.

(ii) Tuning parameter in Bonferroni correction: $\alpha_d = \Theta(N^{-\delta'})$ for all $d \in [D]$ with some $\delta' > 0$.

Condition 2 specifies the allowable range of the true factorial effects. The tuning parameter α_d converges to zero, which ensures no Type I error asymptotically. Wasserman and Roeder (2009) (Theorem 4.1 and 4.2) assumed similar conditions in high-dimensional model selection settings for linear models.

The next condition specifies a set of regularity assumptions on potential outcomes.

Condition 3 (Regularity conditions on $Y_i(\boldsymbol{z})$'s)**.** *The potential outcomes satisfy*

(i) Nondegenerate correlation matrix. *Let $V^\star$ be the correlation matrix of $\widehat{Y}$. There exists a $\sigma > 0$, such that*

$$\varrho_{\min}(V^\star)/\varrho_{\max}(V^\star) \geq \sigma^2. \tag{3.9}$$

(ii) Bounded fourth central moments. *There exists a universal constant $\Delta > 0$ such that*

$$\max_{\boldsymbol{z}\in[Q]} \frac{1}{N}\sum_{i=1}^{N}\{Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\}^4 \leq \Delta^4. \tag{3.10}$$

(iii) Bounded standardization scales. *Assume M_N is bounded by a constant $M > 0$, where M_N is:*

$$M_N = \frac{\max_{i\in[N],q\in[Q]} |Y_i(q) - \overline{Y}(q)|}{\{\min_{q\in[Q]} S(q,q)\}^{1/2}}.$$

Condition 3(i) requires the correlation matrix of $\widehat{Y}$ to be well behaved. Condition 3(ii) controls the moments of the potential outcomes. Condition 3(iii) imposes a universal bound on the standardization of potential outcomes.

Lastly, we assume the following structural condition on the factorial effects:

Condition 4 (Hierarchical structure in factorial effects)**.** *The nonzero true factorial effects align with the effect heredity principle:*

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

We then present the screening consistency property of our algorithm:

Theorem 1 (Perfect screening property)**.** *Under Conditions 1-4, the working model selected by Algorithm 1 converges to the true model with probability one as the sample size goes to infinity:*

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

4 Inference under perfect screening

In this section, we start with presenting statistical inference framework (including causal parameters, estimators, and confidence interval constructions) when the perfect screening property established in (1) is satisfied. We then study the theoretical properties of the proposed framework in the second subsection.

4.1 Statistical inference

In factorial experiments, a causal parameter of interest can be expressed as a weighted combination of average potential outcomes, that is

$$\gamma = \sum_{\boldsymbol{z} \in \mathcal{T}} \boldsymbol{f}(\boldsymbol{z}) \overline{Y}(\boldsymbol{z}) \triangleq \boldsymbol{f}^\top \overline{Y}, \tag{4.11}$$

where $\boldsymbol{f} = \{\boldsymbol{f}(\boldsymbol{z})\}_{\boldsymbol{z} \in \mathcal{T}}$ is a pre-specified weighting vector. For example, if one is interested in estimating the main factorial effects, $\boldsymbol{f}$ can be taken as the contrast vectors $g_{\{k\}}$ given in (2.1). If one wants to estimate interaction effects, then $\boldsymbol{f}$ can be constructed from (2.2). When we have multiple causal parameters defined through different weighting vectors, we later present a method to handle the multiple comparison issues in this context.

Without factor screening, a well-studied plug-in estimator of γ in the existing literature is to replace with its sample analogue (Li and Ding, 2017; Zhao and Ding, 2021b; Shi and Ding, 2022):

$$\widehat{\gamma} = \sum_{\boldsymbol{z} \in \mathcal{T}} \boldsymbol{f}(\boldsymbol{z}) \widehat{Y}(\boldsymbol{z}) \triangleq \boldsymbol{f}^\top \widehat{Y}, \tag{4.12}$$

where $\widehat{Y}(\boldsymbol{z}) = \{N(\boldsymbol{z})\}^{-1} \sum_{Z_i = \boldsymbol{z}} Y_i$. When $N(z) \geq 2$, its variance can be estimated by:

$$\widehat{v}^2 = \sum_{\boldsymbol{z} \in \mathcal{T}} \boldsymbol{f}(\boldsymbol{z})^2 N(\boldsymbol{z})^{-1} \widehat{S}(\boldsymbol{z}, \boldsymbol{z}) \triangleq \boldsymbol{f}^\top \widehat{V}_{\widehat{Y}} \boldsymbol{f}, \tag{4.13}$$

where $\widehat{S}(\boldsymbol{z}, \boldsymbol{z}) = \{N(\boldsymbol{z}) - 1\}^{-1} \sum_{Z_i = \boldsymbol{z}} (Y_i - \widehat{Y}(\boldsymbol{z}))^2$, and $\widehat{V}_{\widehat{Y}} = \mathrm{Diag}\left\{ N(\boldsymbol{z})^{-1} \widehat{S}(\boldsymbol{z}, \boldsymbol{z}) \right\}_{\boldsymbol{z} \in \mathcal{T}}$.

With the help of factor screening, based on the selected working model $\widehat{\mathbb{M}}$, we consider a potentially more efficient estimator of $\overline{Y}$ via restricted least squares

$$\widehat{Y}_{\mathrm{R}} = \arg\min \left\{ \|\widehat{Y} - \mu\|_2^2 : \ G(\cdot, \widehat{\mathbb{M}}^c)^\top \mu = 0, \ \mu \in \mathbb{R}^Q \right\} = Q^{-1} G(\cdot, \widehat{\mathbb{M}}) G(\cdot, \widehat{\mathbb{M}})^\top \widehat{Y}. \tag{4.14}$$

The closed form solution is obtained by exploiting the orthogonality property of G. We provide a justification of this result in Lemma 6. Under perfect screening, $\widehat{Y}_{\mathrm{R}}$ is also an unbiased estimator for $\overline{Y}$. This allows us to construct a restricted least squares (RLS) based estimator of γ with variance estimation:

$$\widehat{\gamma}_{\mathrm{R}} = \boldsymbol{f}^\top \widehat{Y}_{\mathrm{R}} \triangleq \boldsymbol{f}[\widehat{\mathbb{M}}]^\top \widehat{Y}, \quad \text{and} \quad \widehat{v}_{\mathrm{R}}^2 = \boldsymbol{f}[\widehat{\mathbb{M}}]^\top \widehat{V}_{\widehat{Y}} \boldsymbol{f}[\widehat{\mathbb{M}}], \tag{4.15}$$

where $\boldsymbol{f}[\widehat{\mathbb{M}}] = Q^{-1} G(\cdot, \widehat{\mathbb{M}}) G(\cdot, \widehat{\mathbb{M}})^\top \boldsymbol{f}$. Leverage these two estimators and Theorem 2 to be presented, we are able to construct a level-$(1-\alpha)$ confidence interval for γ:

$$\left[\widehat{\gamma}_{\mathrm{R}} \pm z_{1-\alpha/2} \times \widehat{v}_{\mathrm{R}} \right], \tag{4.16}$$

where $z_{1-\alpha/2}$ is $(1-\alpha/2)$th quantile of a standard normal distribution.

In the following subsection, we provide the theoretical properties of $\widehat{\gamma}_{\mathrm{R}}$ and $\widehat{v}_{\mathrm{R}}^2$, and compare their asymptotic behaviors with the plug-in estimators $\widehat{\gamma}$ and $\widehat{v}^2$ in various settings.

4.2 Theoretical properties under perfect screening

In this section, we first present the asymptotic normality result for $\widehat{\gamma}_{\mathrm{R}}$. To simplify discussion, we denote $\boldsymbol{f}^\star = \boldsymbol{f}[\mathbb{M}^\star] = Q^{-1}G(\cdot,\mathbb{M}^*)G(\cdot,\mathbb{M}^*)^\top \boldsymbol{f}$. Given $\mathbb{M}^*$ is the true sparse model, we have $(\boldsymbol{f}^\star)^\top \overline{Y} = \boldsymbol{f}^\top \overline{Y}$, for all $\boldsymbol{f} \in \mathbb{R}^Q$. We make the following condition for $\boldsymbol{f}^\star$:

Condition 5 (Condition on $\boldsymbol{f}^*$)**.** *The weighting vector $\boldsymbol{f}^\star = Q^{-1}G(\cdot,\mathbb{M}^*)G(\cdot,\mathbb{M}^*)^\top \boldsymbol{f}$ satisfies*

$$N_0^{-1/2} \cdot \frac{\|\boldsymbol{f}^\star\|_\infty}{\|\boldsymbol{f}^\star\|_2} \longrightarrow 0. \tag{4.17}$$

Because $N_0 Q = \Theta(N)$ under Condition 1 and the ratio $\|\boldsymbol{f}^\star\|_\infty / \|\boldsymbol{f}^\star\|_2$ is upper bounded by (see Lemma 1 for a rigorous statement):

$$N_0^{-1/2} \cdot \frac{\|\boldsymbol{f}^\star\|_\infty}{\|\boldsymbol{f}^\star\|_2} \leq \left(\frac{|\mathbb{M}^\star|}{N_0 Q}\right)^{1/2}, \tag{4.18}$$

Condition 5 says that the cardinality of the true working model $|\mathbb{M}^*|$ is asymptotically much smaller compared to the sample size N.

We are now ready to present the asymptotic properties of $\widehat{\gamma}_{\mathrm{R}}$ and $\widehat{v}_{\mathrm{R}}^2$:

Theorem 2 (Statistical properties of $\widehat{\gamma}_{\mathrm{R}}$ and $\widehat{v}_{\mathrm{R}}^2$)**.** *Under Conditions 1-5, we have, as the sample size N tends to infinity,*

$$\frac{\widehat{\gamma}_{\mathrm{R}} - \gamma}{v_{\mathrm{R}}} \rightsquigarrow \mathcal{N}(0,1)$$

where $v_{\mathrm{R}}^2 = \boldsymbol{f}^{\star\top} V_{\widehat{Y}} \boldsymbol{f}^\star$ and $V_{\widehat{Y}} = \mathrm{Diag}\left\{N(\boldsymbol{z})^{-1} S(q,q)\right\} - N^{-1}S$. Furthermore, without loss of generality, assume $\|\boldsymbol{f}^\star\|_\infty = O(Q^{-1})$, the variance estimator $\widehat{v}_{\mathrm{R}}^2$ is consistent and robust:

$$N(\widehat{v}_{\mathrm{R}}^2 - v_{\mathrm{R,lim}}^2) \xrightarrow{\mathbb{P}} 0, \quad v_{\mathrm{R,lim}}^2 \geq v_{\mathrm{R}}^2,$$

where $v_{\mathrm{R,lim}}^2 = \boldsymbol{f}^{\star\top} \mathrm{Diag}\left\{N(\boldsymbol{z})^{-1} S(\boldsymbol{z},\boldsymbol{z})\right\} \boldsymbol{f}^\star$ is the robust asymptotic variance of $\widehat{\gamma}_{\mathrm{R}}$.

The above theorem guarantees that the proposed confidence interval in (4.16) for γ attains the nominal coverage probability asymptotically. We note that the above result applies to any model selection procedure with the oracle property (Fan and Li, 2001) and does not restrict to the marginal t-test with Bonferroni correction adopted in Algorithm 1. Furthermore, it allows us to compare the conditions for reaching asymptotic normality of $\widehat{\gamma}$, which we formalize in the following remark:

Remark 1 (Comparison of conditions for asymptotic normality)**.** *To establish the central limit theorem of the classical plug-un estimator* $\widehat{\gamma}$ (4.12)*, Shi and Ding (2022) provided the following condition similar to our Condition 5*

$$N_0^{-1/2} \cdot \frac{\|\boldsymbol{f}\|_\infty}{\|\boldsymbol{f}\|_2} \to 0. \tag{4.19}$$

In a case where N_0 *is small (see Condition 1 for the definition of* N_0*) and both* $\boldsymbol{f}$ *and* $\mathbb{M}^*$ *are sparse, Condition* (4.19) *fails while Condition* (4.17) *continues to hold, suggesting that our proposed estimator* $\widehat{\gamma}_{\mathrm{R}}$ *can be more robust than* $\widehat{\gamma}$ *in practice.*

To elaborate the benefits of conducting forward factorial screening in terms of asymptotic efficiency, we make a simple comparison in Proposition 1 based on a concrete choice of $\boldsymbol{f}$. Based on the result established in Shi and Ding (2022), the plug-in estimator $\widehat{\gamma}$ satisfies

$$\frac{\widehat{\gamma} - \gamma}{v} \rightsquigarrow \mathcal{N}(0,1), \quad \text{where } v^2 = \boldsymbol{f}^\top V_{\widehat{Y}} \boldsymbol{f},$$

under appropriate conditions.

Proposition 1 (Asymptotic relative efficiency comparison between $\widehat{\gamma}$ and $\widehat{\gamma}_{\mathrm{R}}$ with sparse $\boldsymbol{f}$)**.** *Define the condition number of* $V_{\widehat{Y}}$ *as* $\kappa(V_{\widehat{Y}})$*, and let* $s^\star$ *denote the number of nonzero elements in* $\boldsymbol{f}$*. Under the condition that both* $\widehat{\gamma}$ *and* $\widehat{\gamma}_{\mathrm{R}}$ *converge to a normal distribution as the sample size tends to infinity, the asymptotic relative efficiency between* $\widehat{\gamma}$ *and* $\widehat{\gamma}_{\mathrm{R}}$ *is upper bounded by*

$$\frac{v_{\mathrm{R}}^2}{v^2} \leq \kappa(V_{\widehat{Y}}) \cdot \frac{s^\star |\mathbb{M}^\star|}{Q},$$

provided that $s^\star |\mathbb{M}^\star| < Q$.

Now we add some interpretation for Proposition 1. The condition number $\kappa(V_{\widehat{Y}})$ captures the variability of the variances of $\widehat{Y}(\boldsymbol{z}) = \{N(\boldsymbol{z})\}^{-1} \sum_{Z_i = \boldsymbol{z}} Y_i$ across multiple treatment combination groups in $\mathcal{T}$. The higher the variability is, the larger the condition number becomes. When the variance of $\widehat{Y}(\boldsymbol{z})$ does not change with its group membership $\boldsymbol{z}$, the condition number $\kappa(V_{\widehat{Y}}) = 1$. This suggests that the proposed RLS-based estimator $\widehat{\gamma}_{\mathrm{R}}$ is more efficient than the plug-in estimator $\widehat{\gamma}$. When the variance of $\widehat{Y}(\boldsymbol{z})$ does change with its group membership $\boldsymbol{z}$, but the variability of such changes is limited in the sense that $\kappa(V_{\widehat{Y}}) < Q/s^\star |\mathbb{M}^\star|$, the proposed estimator is also more efficient than $\widehat{\gamma}$. The above result can be extend to compare the length of the confidence intervals as well; see Proposition 2 in the Supplementary Material for details.

5 Inference under imperfect screening

In factorial experiments, the perfect screening property can be too much to hope for, this is partially because the factorial experimental data are driven by the Effect Hierarchy Principle (Wu and Hamada, 2011). Implied by this principle, the main factorial effects and lower-order factorial effects are more likely to be non-negligible than the higher-order factorial effects. Once the higher factorial effects are marginal, the perfect screening property may no longer hold. Take our selection procedure in Algorithm 1 (marginal t-test with Bonferroni correction) for example, whenever Condition 2(i) is violated, we may lose the perfect screening property established in Theorem 6 with high chance.

In response to this challenge, we propose two alternative strategies along with proposed causal effects estimators (summarized in Figure 1) in the first subsection. We then present their theoretical guarantees in the subsequent section.

5.1 Statistical inference with two alternative strategies

The two proposed strategies are built on the belief that perfect screening is more plausible for selecting the main factorial effects and lower-order factorial effects up to level $d^\star$.

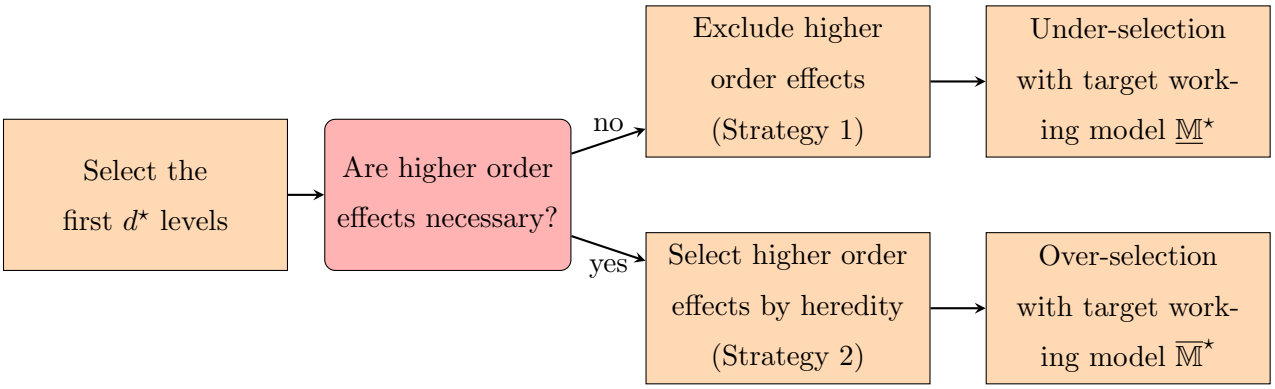

Figure 1: General strategy for factorial screening

For the first strategy, when the higher order factorial effects beyond level $d^\star$ are believed to be negligible, we may stop our forward screening procedure in Algorithm 1 at $d = d^*$ (instead of $d = D$). Similar strategies have been applied in the existing hierarchical model selection literature. For example, Hao and Zhang (2014); Lim and Hastie (2015); Yuan et al. (2007), among others, propose screening procedures that only include main effect selection and two-way interaction screening, assuming that the higher order interactions (beyond level two) are negligible. As this strategy

targets a true working model $\underline{\mathbb{M}}^\star$ up to level $d^\star$, that is,

$$\underline{\mathbb{M}}^\star = \cup_{d=1}^{d^\star} \mathbb{M}_d^\star \subseteq \mathbb{M}^\star,$$

this strategy leads to an under-selected parsimonious working model. We summarize this strategy in as follows:

Strategy 1 (Under selection by excluding higher order interactions)**.** *In Algorithm 1, we stop the screening procedure at $d = d^*$. Or equivalently, we set $\alpha_d = \infty$ for $d \geq d^\star + 1$ so that no effects beyond level $d^\star$ will be selected and $\widehat{\underline{\mathbb{M}}} = \cup_{d=1}^{d^\star} \widehat{\mathbb{M}}_d$.*

Given the selected working model $\widehat{\underline{\mathbb{M}}}$, we can then again construct an estimator of $\gamma = \boldsymbol{f}^\top \overline{Y}$ (also defined in Section 4.1) based on the restricted least squares:

$$\widehat{\gamma}_{\text{RU}} = \boldsymbol{f}[\widehat{\underline{\mathbb{M}}}]^\top \widehat{Y}, \quad \text{and} \quad \widehat{v}_{\text{RU}}^2 = \boldsymbol{f}[\widehat{\underline{\mathbb{M}}}]^\top \widehat{V}_{\widehat{Y}} \boldsymbol{f}[\widehat{\underline{\mathbb{M}}}].$$

For the second strategy, rather than excluding all higher order interactions with negligible effects, we may further leverage the Effect Heredity Principle and continue our screening procedure beyond level $d^\star$. This means that instead of selecting the higher order interactions via marginal t-test and Bonferroni correction, we select the higher order interaction terms whenever either all of their parent effects are selected (strong heredity) or one of their parent effects is selected (weak heredity). While such a strategy takes higher order factorial effects into account, it often targets a true working model $\overline{\mathbb{M}}^\star$ that includes the true model $\mathbb{M}^\star$, that is,

$$\mathbb{M}^\star \subseteq \overline{\mathbb{M}}^\star = \bigcup_{d=1}^{D} \overline{\mathbb{M}}_d^\star, \text{ where } \overline{\mathbb{M}}_d^\star = \begin{cases} \mathbb{M}_d^\star, & d \leq d^\star; \\ \mathrm{H}^{(d-d^\star)}(\mathbb{M}_{d^\star}^\star), & d^\star + 1 \leq d \leq D. \end{cases}$$

The selected model by this strategy is expected to introduce an over-selected model that includes $\mathbb{M}^\star$ as well. We summarize this strategy as follows:

Strategy 2 (Over selection by including higher order interactions through the Effect Heredity Principle)**.** *In Algorithm 1, set $\alpha_d = 0, d \geq d^\star + 1$ and apply a heredity principle (either weak or strong, depending on people's knowledge on the structure of the effects). Then the high order effects beyond level $d^\star$ are selected merely by heredity principle and*

$$\widehat{\overline{\mathbb{M}}} = \cup_{d=1}^{D} \widehat{\mathbb{M}}_d; \ \widehat{\mathbb{M}}_d = \mathrm{H}^{(d-d^\star)}(\widehat{\mathbb{M}}_{d^\star}), d \geq d^\star + 1.$$

Here $\mathrm{H}^{(d-d^\star)}$ is the $(d-d^\star)$-order composition of H*, meaning applying* H *for $(d-d^\star)$ times.*

Given the selected working model $\widehat{\overline{\mathbb{M}}}$, similarly, we can construct an estimator of $\gamma = \boldsymbol{f}^\top \overline{Y}$ based on the restricted least squares:

$$\widehat{\gamma}_{\text{RO}} = \boldsymbol{f}[\widehat{\overline{\mathbb{M}}}]^\top \widehat{Y}, \quad \text{and} \quad \widehat{v}^2_{\text{RO}} = \boldsymbol{f}[\widehat{\overline{\mathbb{M}}}]^\top \widehat{V}_{\widehat{Y}} \boldsymbol{f}[\widehat{\overline{\mathbb{M}}}].$$

In the following subsection, we study the theoretical properties of $\widehat{\gamma}_{\text{RO}}$ and $\widehat{\gamma}_{\text{RU}}$ and demonstrate their trade-offs for statistical inference.

5.2 Theoretical properties under imperfect screening

Throughout this section, as we are in a scenario where the higher order factorial effects beyond level $d^\star$ are negligible, we work under a relaxed condition of Condition 2 as follows:

Condition 6 (Order of parameters up to level $d^\star$)**.** *The true parameters and tuning parameters have the following order:*

(i) True parameter: $|\tau_{\mathcal{K}}| = \Theta(N^\delta)$ *for some* $-1/2 < \delta \leq 0$ *and all* $\mathcal{K} \in \underline{\mathbb{M}}^\star$.

(ii) Tuning parameter in Bonferroni correction: $\alpha_d = \Theta(N^{-\delta'})$ *for all* $d \leq d^\star$ *with some* $\delta' > 0$.

Condition 6 no longer impose any restriction on the order of the parameters beyond level $d^\star$. By Theorem 6, Condition 6 guarantees that Algorithm 1 perfectly screens the first $d^\star$ levels of factorial effects in the sense that

$$\mathbb{P}\left\{\widehat{\mathbb{M}}_d = \mathbb{M}^\star_d,\ d = 1, \cdots d^\star\right\} \to 1.$$

We start by analyzing the statistical property of $\widehat{\gamma}_{\text{RU}}$ with $\widehat{\underline{\mathbb{M}}}$ obtained from the under selection Strategy 1. Because the selected working model might deviate from the truth beyond level $d^\star$, $\widehat{\gamma}_{\text{RU}}$ may not be an unbiased estimator of γ. Therefore, we focus on weighting vectors $\boldsymbol{f}$ that satisfy certain orthogonality conditions as introduced in Theorem 3 below:

Theorem 3 (Guarantee for Strategy 1)**.** *Define* $\underline{\boldsymbol{f}}^\star = Q^{-1} G(\cdot, \underline{\mathbb{M}}^\star) G(\cdot, \underline{\mathbb{M}}^\star)^\top \boldsymbol{f}$. *Assume Conditions 1, 3, 4 and 6. Also assume* $\boldsymbol{f}$ *satisfies the following orthogonality condition:*

$$G(\cdot, \mathbb{M}^\star_d)^\top \boldsymbol{f} = 0,\ d^\star + 1 \leq d \leq D^\star. \tag{5.20}$$

If as N *tends to infinity,*

$$N_0{}^{-1/2} \cdot \frac{\|\underline{\boldsymbol{f}}^\star\|_\infty}{\|\underline{\boldsymbol{f}}^\star\|_2} \longrightarrow 0,$$

then

$$\frac{\widehat{\gamma}_{\mathrm{RU}} - \gamma}{v_{\mathrm{RU}}} \rightsquigarrow \mathcal{N}(0,1).$$

where $v_{\mathrm{RU}}^2 = \underline{\boldsymbol{f}}^{\star\top} V_{\widehat{Y}} \underline{\boldsymbol{f}}^{\star}$. *Furthermore, without loss of generality, assume* $\|\overline{\boldsymbol{f}}^{\star}\|_\infty = O(Q^{-1})$, *the variance estimator* $\widehat{v}_{\mathrm{RU}}^2$ *is consistent and robust:*

$$N(\widehat{v}_{\mathrm{RU}}^2 - v_{\mathrm{RU,lim}}^2) \xrightarrow{\mathbb{P}} 0, \quad v_{\mathrm{RU,lim}}^2 \geq v_{\mathrm{RU}}^2,$$

where $v_{\mathrm{RU,lim}}^2 = \underline{\boldsymbol{f}}^{\star\top} \mathrm{Diag}\left\{N(\boldsymbol{z})^{-1} S(\boldsymbol{z},\boldsymbol{z})\right\} \underline{\boldsymbol{f}}^{\star}$ *is the robust asymptotic variance of* $\widehat{\gamma}_{\mathrm{RU}}$.

The orthogonality condition presented in (5.20) restricts the weighting vector $\boldsymbol{f}$ to be orthogonal to the higher order contrasts, because the higher order interactions are excluded from the model and making inference on a weighted combination of those excluded interactions is infeasible. One example satisfying (5.20) is an arbitrary nonzero linear combination of lower-order contrasts, given by $\boldsymbol{f} = G(,\cup_{d=1}^{d^\star}\mathbb{M}_d^\star)\boldsymbol{b}$, for any real $\boldsymbol{b}$ with dimension $\sum_{d=1}^{d^\star}|\mathbb{M}_d^\star|$. When $\boldsymbol{b}$ is one of canonical bases, γ is the corresponding non-zero factorial effects.

As for Strategy 2, similarly, we have the following results:

Theorem 4 (Guarantee for Strategy 2)**.** *Assume Conditions 1, 3, 4 and 6. Define the weighting vector* $\overline{\boldsymbol{f}}^{\star} = Q^{-1}G(\cdot,\overline{\mathbb{M}}^{\star})G(\cdot,\overline{\mathbb{M}}^{\star})^\top \boldsymbol{f}$. *If as* N *tends to infinity,*

$$N_0^{-1/2} \cdot \frac{\left\|\overline{\boldsymbol{f}}^{\star}\right\|_\infty}{\left\|\overline{\boldsymbol{f}}^{\star}\right\|_2} \longrightarrow 0,$$

then

$$\frac{\widehat{\gamma}_{\mathrm{RO}} - \gamma}{v_{\mathrm{RO}}} \rightsquigarrow \mathcal{N}(0,1),$$

where $v_{\mathrm{RO}}^2 = \overline{\boldsymbol{f}}^{\star\top} V_{\widehat{Y}} \overline{\boldsymbol{f}}^{\star}$. *Furthermore, without loss of generality, assume* $\|\overline{\boldsymbol{f}}^{\star}\|_\infty = O(Q^{-1})$, *the variance estimator* $\widehat{v}_{\mathrm{RO}}^2$ *is consistent and robust:*

$$N(\widehat{v}_{\mathrm{RO}}^2 - v_{\mathrm{RO,lim}}^2) \xrightarrow{\mathbb{P}} 0, \quad v_{\mathrm{RO,lim}}^2 \geq v_{\mathrm{RO}}^2,$$

where $v_{\mathrm{RO,lim}}^2 = \overline{\boldsymbol{f}}^{\star\top} \mathrm{Diag}\left\{N(\boldsymbol{z})^{-1} S(\boldsymbol{z},\boldsymbol{z})\right\} \overline{\boldsymbol{f}}^{\star}$ *is the robust asymptotic variance of* $\widehat{\gamma}_{\mathrm{RO}}$.

When analyzing Strategy 1 and 2, Algorithm 1 recovers a target model with high probability. Both strategies have advantages and disadvantages. Under-selection reflects bias-variance trade-off: it can induce more bias for certain weighting vectors, but the constructed estimator typically enjoys smaller variance. Over-selection can reduce bias for estimation, but may lose efficiency because

one might include too many redundant terms into the selected model. In practice, if higher order interactions are not crucial for study, Strategy 1 should be applied. If high order interactions are of interest and hard to select, one could pursue Strategy 2 as a practically useful and interpretable solution.

Remark 2. *Under homoscedasticity, we can prove* $v_{\mathrm{RU}}^2 \leq v_{\mathrm{RO}}^2$. *Therefore, by excluding higher order terms and pursuing under-selection, we can obtain an equal or smaller asymptotic variance compared with over-selection. In general, due to heteroskedasticity, the order of* v_{RU}^2 *and* v_{RO}^2 *depends on the choice of target weighing vector* $\boldsymbol{f}$. *Here we take a sparse* $\boldsymbol{f} = \boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top$ *as an example. We can derive*

$$\frac{v_{\mathrm{RU}}^2}{v_{\mathrm{RO}}^2} \leq \kappa(V_{\widehat{Y}}) \cdot \frac{\left|\underline{\mathbb{M}}^\star\right|}{\left|\overline{\mathbb{M}}^\star\right|}.$$

When the variability of $V_{\widehat{Y}}$ *between treatment arms is small in the sense that* $\kappa(V_{\widehat{Y}}) < \left|\overline{\mathbb{M}}^\star\right| / \left|\underline{\mathbb{M}}^\star\right|$, *under-selection leads to smaller asymptotic variance for inferring* $\boldsymbol{e}_1^\top \overline{Y}$.

6 An extension to comparing multiple causal effects

In the previous sections, we consider the problem on making inference on a single factorial causal effect $\gamma = \boldsymbol{f}^\top \overline{Y}$. We now provide an extension of our framework that delivers valid inference on multiple causal parameters.

In this section, we study the problem of "inference on best causal effects" as an application of forward factorial screening and inference based on restricted least squares. The goal of the problem is to compare a set of candidate causal effects in the form of (4.11) and identify the maximal effect. More concretely, suppose we have a set of effects defined by pre-specified weighting vectors:

$$\Gamma = \{\gamma_1, \ldots, \gamma_L\}, \quad \gamma_l = \boldsymbol{f}_l^\top \overline{Y}.$$

We hope to perform statistical inference on the maximal effects in Γ:

$$\gamma_{(1)} = \max_{l \in [L]} \gamma_l.$$

As one specific example, if we choose $\{\boldsymbol{f}_l\}_{l \in [L]} = \{\boldsymbol{e}(\boldsymbol{z})\}_{\boldsymbol{z} \in \mathcal{T}}$ to be a subset of the canonical bases $\{\boldsymbol{e}(\boldsymbol{z})\}_{\boldsymbol{z} \in \mathcal{T}}$, then our inferential target is

$$\overline{Y}_{(1)} = \max_{z \in \mathcal{T}'} \overline{Y}(\boldsymbol{z}). \tag{6.21}$$

In this case, the goal is to identify the treatment combination that demonstrates the "best" performance measured by level of average potential outcome and do inference on the average potential outcome for the selected arms.

The challenge in identifying $\gamma_{(1)}$ lies in that similar effects might not be statistically distinguishable with limited sample size. We introduce the following notion $\mathcal{L}_1$ to include all effects that achieve or are statistically "close" to $\gamma_{(1)}$:

$$\mathcal{L}_1 = \left\{ l \in [L] \mid |\gamma_l - \gamma_{(1)}| = O(N^{-\delta_3}) \right\}, \text{ for some } \delta_3 > 0. \tag{6.22}$$

A well-known fact is that the naive estimator $\max_{\boldsymbol{z}\in[Q]} \widehat{Y}(\boldsymbol{z})$ can be problematic for inferring $\gamma_{(1)}$, especially if $\mathcal{L}_1$ contains more than one element (Andrews et al., 2019; Wei et al., 2022).

With the above basic setup, we introduce the following Algorithm 2 for selecting the most effective effects. The algorithm consists of three major components:

(i) Construct $\widehat{\gamma}_l = \boldsymbol{f}_l^\top \widehat{Y}_{\mathrm{R}}$ with feature screening (Step 1-2). These RLS based estimators enjoy great benefits for large Q and small N_0 regimes.

(ii) Construct $\widehat{\mathcal{L}}_1$ as an estimator for $\mathcal{L}_1$ (Step 3). $\widehat{\mathcal{L}}_1$ aggregates the arms whose estimated averages are close to the empirical maximum. We will show that with proper tuning, this procedure recovers $\mathcal{L}_1$ with high probability.

(iii) Construct estimators by averaging over $\widehat{\mathcal{L}}_1$ (Step 4). The difference in values across arms in $\mathcal{L}_1$ are mainly due to randomness of the study. Therefore, we average the estimates over the selected $\widehat{\mathcal{L}}_1$ to alleviate the impact of randomness and obtain accurate estimates.

We provide some theoretical analysis of Algorithm 2. Define

$$d_h = \max_{\boldsymbol{z}\in\mathcal{L}_1} |\gamma_l - \gamma_{(1)}|, \quad d_h^\star = \min_{\boldsymbol{z}\notin\mathcal{L}_1} |\gamma_l - \gamma_{(1)}|.$$

which we refer to as within-group diameter and between-group distance respectively. Condition 7 below characterizes the order of the involved quantities:

Condition 7 (Order of d_h, $d_h^\star$ and η_N)**.** *Assume the following scaling of parameters:*

$$d_h^\star = \Theta(N^{\delta_1}), \quad \eta_N = \Theta(N^{\delta_2}), \quad d_h = \Theta(N^{\delta_3}).$$

with $\delta_3 \le -1/2 < \delta_2 < \delta_1 \le 0$.

Define the population counterpart of $\boldsymbol{f}_{(1)}$:

$$\boldsymbol{f}_{(1)}^\star = (Q|\mathcal{L}_1|)^{-1} \sum_{l\in\mathcal{L}_1} G(\cdot, \mathbb{M}^\star) G(\cdot, \mathbb{M}^\star)^\top \boldsymbol{f}_l.$$

Algorithm 2: Select the most effective arms

Input: Factorial data (Y_i, Z_i); predetermined integer D; initial model for factorial effects $\widehat{\mathbb{M}} = \{\varnothing\}$; significance level $\{\alpha_d\}_{d=1}^{D}$; set of weighting vectors $\{\boldsymbol{f}_l\}_{l\in[L]}$; thresholds η_N.

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

1 Perform effects screening with Algorithm 1 and obtain working model $\widehat{\mathbb{M}}$.

2 Obtain RLS based estimates: let $\widehat{Y}_{\mathrm{R}}$ be defined as (4.14), and compute

$$\widehat{\gamma}_l = \boldsymbol{f}_l^\top \widehat{Y}_{\mathrm{R}} = G(\cdot, \widehat{\mathbb{M}})\widehat{\tau}(\widehat{\mathbb{M}}), \quad l \in [L].$$

3 Select the best factor level combinations:

$$\widehat{\mathcal{L}}_1 = \left\{ l \in [L] \mid |\widehat{\gamma}_l - \max_{l\in[L]} \widehat{\gamma}_l| \le \eta_N \right\}.$$

4 Define

$$\boldsymbol{f}_{(1)} = (Q|\widehat{\mathcal{L}}_1|)^{-1} \sum_{l\in\widehat{\mathcal{L}}_1} G(\cdot, \widehat{\mathbb{M}}) G(\cdot, \widehat{\mathbb{M}})^\top \boldsymbol{f}_l.$$

Then generate point estimates and variance estimator for the effect size over the selected tie:

$$\widehat{Y}_{(1)} = \frac{1}{|\widehat{\mathcal{L}}_1|} \sum_{l\in\widehat{\mathcal{L}}_1} \widehat{\gamma}_l = \boldsymbol{f}_{(1)}^\top \widehat{Y},$$
$$\widehat{v}_{(1)} = \boldsymbol{f}_{(1)}^\top \widehat{V}_Y \boldsymbol{f}_{(1)}.$$

5 **return** $\widehat{\mathcal{L}}_1, \widehat{Y}_{(1)}, \widehat{v}_{(1)}$

Theorem 5 (Asymptotic results on the estimated effects with screening)**.** *Assume Condition 1, 3 and 7. Further assume that the perfect screening holds with probability tending to one (i.e.,* $\mathbb{P}\{\widehat{\mathbb{M}} = \mathbb{M}^\star\} \to 1$*). Assume* $|\mathbb{M}^\star| = \Theta(N^{\delta_4})$ *for some* $\delta_4 \ge 0$ *and*

$$N^{-(1+2\delta_2-\delta_4)} \to 0, \quad L \cdot |\mathcal{L}_1| \cdot N^{-\frac{1-\delta_4}{2}} \to 0.$$

Then the point estimates are asymptotically jointly normal:

$$\frac{\widehat{\gamma}_{(1)} - \gamma_{(1)}}{v_{(1)}} \rightsquigarrow \mathcal{N}(0,1),$$

where $v^2_{(1)} = \boldsymbol{f}^{\star\top}_{(1)} V_Y \boldsymbol{f}^{\star}_{(1)}$. *Moreover,* $\widehat{v}^2_{(1)}$ *is consistent and robust:*

$$N(\widehat{v}^2_{(1)} - v^2_{(1),\lim}) \xrightarrow{\mathbb{P}} 0,\ v^2_{(1),\lim} \geq v^2_{(1)},$$

where $v^2_{(1),\lim} = \boldsymbol{f}^{\star\top}_{(1)} \mathrm{Diag}\left\{N(\boldsymbol{z})^{-1} S(\boldsymbol{z},\boldsymbol{z})\right\} \boldsymbol{f}^{\star}_{(1)}$.

From Theorem 5, with factorial screening, when the size of the true working model is small (say $\delta_4 = 0$) and screening is consistent, $N^{-(1+2\delta_2)}$ always converge to 0. Thus one only needs

$$L|\mathcal{L}_1|\left(|\mathbb{M}^\star|/N\right)^{1/2} \to 0.$$

The condition relies on the scaling of N instead of a particular set of $N(\boldsymbol{z})$'s. In other words, we can incorporate information from other treatment arms to facilitate inference. On the contrary, without screening, one requires Q to be small compared to N or $\{\boldsymbol{f}_l\}_{l\in[L]}$ are dense, which might not be true in large Q setups and practical scenarios such as (6.21).

7 Simulation

In this section, we use numerical experiments to study the finite sample performance of the proposed forward screening framework and the inferential properties of the RLS based estimator.

7.1 Varying sample size

In the first setup, we evaluate the performance of different screening procedure under varying sample sizes. We set up a 2^8 uniform factorial experiment ($K = 8$). There are N_0 units in each treatment arm where N_0 is set to be a varying number. We generate potential outcomes independently from a shifted exponential distribution:

$$Y_i(\boldsymbol{z}) \sim \mathrm{EXP}(1) - 1 + \mu(\boldsymbol{z}).$$

Here $\mu(\boldsymbol{z})$ are super population means of potential outcomes under treatment $\boldsymbol{z}$. $\mu(\boldsymbol{z})$ are generated such that the factorial effects satisfy the following structure:

- Main effects: the main effects corresponding to the first five factors, $\tau_{\{1\}}, \ldots, \tau_{\{5\}}$, are nonzero; the rest three main effects, $\tau_{\{6\}}, \ldots, \tau_{\{8\}}$, are zero.
- Two-way interactions: the two-way interactions associated with the first five factors are nonzero, i.e., $\tau_{\{kl\}} \neq 0$ for $k \neq l, k, l \in [5]$. All the rest two-way interactions are zero.

- Higher order interactions: all the higher order interactions $\tau_{\mathcal{K}}$ where $|\mathcal{K}| \geq 3$ are zero.

The above rules guarantee that the factorial effects follow the strong heredity principle and demonstrate a sparse pattern. More details can be found in the `R` code attached in the support materials.

To perform effect screening, we compare four methods in our study: (i) *Forward Bonferroni.* screening in a forward fashion based on Bonferroni corrected margin t tests; (ii) *Forward Lasso.* screening in a forward fashion based on Lasso; (iii) *Naive Bonferroni.* screening with the full working model based on Bonferroni corrected margin t tests; (iv) *Naive Lasso.* screening with the full working model based on Lasso. For each of the screening methods, we use several criteria to compare the performance of the methods. (i) *Perfect selection probability.* We report the perfect selection probability $\mathbb{P}\{\widehat{\mathbb{M}} = \mathbb{M}^\star\}$. (ii) *Power of* $\widehat{\gamma}_{\mathrm{R}}$. We evaluate the power of the RLS based estimators $\widehat{\gamma}_{\mathrm{R}}$ and $\widehat{v}_{\mathrm{R}}$ for testing a causal effect γ_{target} specified by a sparse vector:

$$\gamma_{\mathrm{target}} = \boldsymbol{f}^\top \overline{Y}, \quad \boldsymbol{f}(\boldsymbol{z}) = \begin{cases} -1, & z_1 = \cdots = z_8 = 0; \\ 1, & z_1 = 0, z_2 = \cdots = z_8 = 1; \\ 0, & \text{for other } \boldsymbol{z}. \end{cases}$$

Intuitively speaking, γ_{target} reflects the causal effects of setting all the factors as 1 except for the first factor compared to the baseline treatment level. (iii) *Coverage probability of the RLS based confidence interval.* We report the coverage probability for γ_{target} defined above with the RLS based confidence interval (4.16) with confidence level 0.95. The results are average over 500 Monte Carlo runs and are presented in Figure 2.

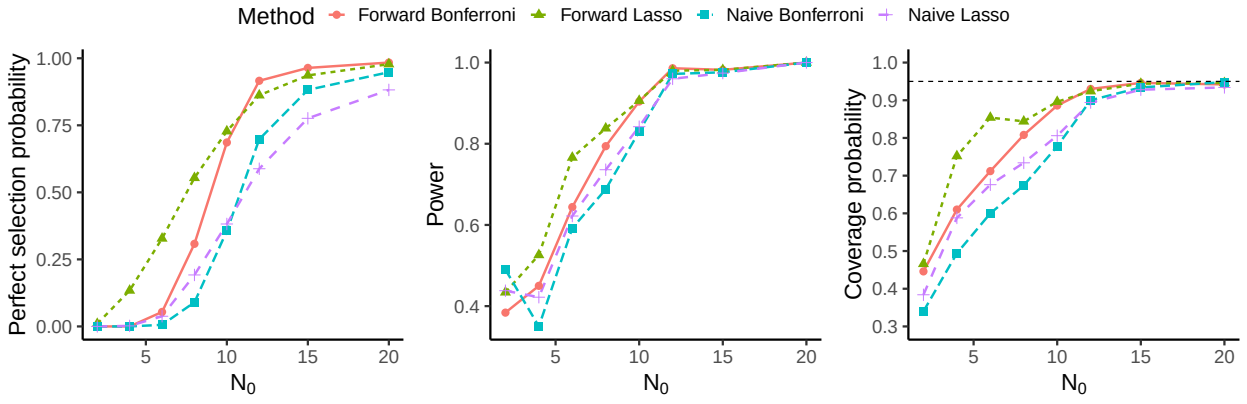

Figure 2: Performance evaluation of effect screening under varying sample size. The left panel presents the probability of perfect model selection. The middle panel shows the power curve for testing $\mathrm{H}_0 : \gamma_{\mathrm{target}} = 0$ with the RLS based estimator $\widehat{\gamma}_{\mathrm{R}}$. The right panel presents the coverage probability of $\widehat{\gamma}_{\mathrm{R}}$ for the target causal effect γ_{target}.

From the left panel of Figure 2, all four effect screening methods lead to perfect selection with probability tending to 1 as the number of replications N_0 increases. Nevertheless, with the forward procedure, the probability of perfect screening are higher than naive screening procedure for all tested N_0. Besides, forward screening keeps the heredity structure and demonstrates higher interpretability than the naive screening methods.

In terms of the power of $\widehat{\gamma}_{\mathrm{R}}$ and $\widehat{v}_{\mathrm{R}}$ for testing $\mathrm{H}_0 : \gamma_{\mathrm{target}} = 0$, the middle panel of Figure 2 shows that all four methods have asymptotic rejection probability 1, while forward screening possesses higher power when the number of replications are limited. It is worthy of mentioning that in our numerical study we have also compared the behavior of the naive moment estimator $\widehat{\gamma}$ and $\widehat{v}$, whose power is much lower than the the RLS based estimators and thus omitted in the figure. This echoes with our conclusion that effect screening can incorporate information across treatment levels and improve the power for testing general causal effects. Analogously, from the right panel of Figure 2, we can witness an improvement in coverage probability of the RLS based confidence intervals with the forward screening procedure.

7.2 Varying effect size

In this section, we study the finite sample performance of effect screening under varying effect sizes. The data generating mechanism is similar to the previous section, except that we fix N_0 to be 2 but vary the size of the factorial effects. Analogously, we still use the four effect screening procedure (Forward Bonferroni, Forward Lasso, Naive Bonferroni and Naive Lasso) and the three criteria (perfect selection probability, power of $\widehat{\gamma}_{\mathrm{R}}$ and coverage probability of the RLS based confidence interval) for comparison. The results are reported in Figure 3.

From Figure 3, we can see that perfect effect screening and valid inference is more plausible when the effect size is large enough. Small effect size is a challenging setting, especially when coupled with limited number of replications (recall $N_0 = 2$ in this simulation setting). Nevertheless, forward screening can increase the perfect selection probability and improve the performance of the RLS based inference even for small effect size.

8 Discussion

In this paper we focus on randomized factorial experiments with binary-level factors. We will leave possible extensions to multi-level factors as future endeavor. Besides, it is of interest to discuss effect screening in observational factorial experiments where the data generating mechanism

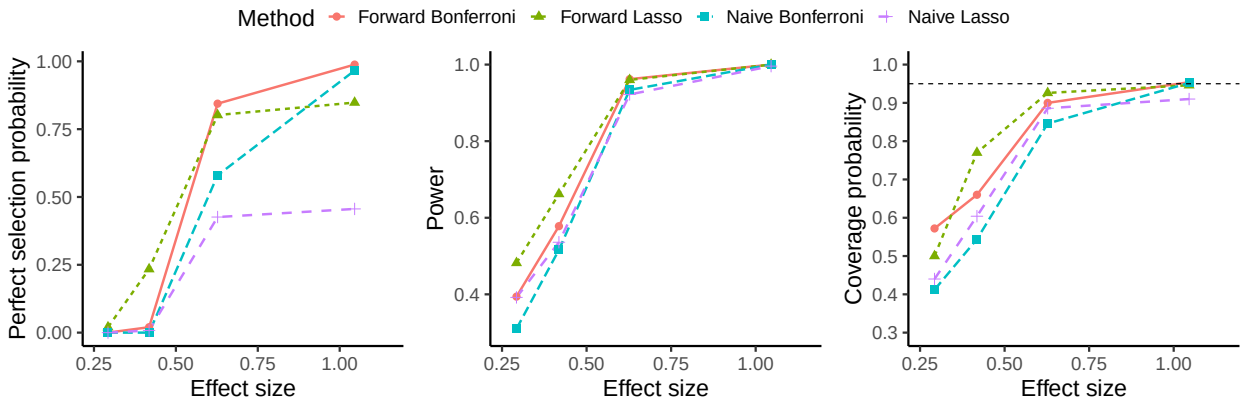

Figure 3: Performance evaluation of effect screening under varying effect sizes. The left panel presents the probability of perfect model selection. The middle panel shows the power curve for testing $\mathrm{H}_0 : \gamma_{\text{target}} = 0$ with the RLS based estimator $\widehat{\gamma}_{\mathrm{R}}$. The right panel presents the coverage probability of $\widehat{\gamma}_{\mathrm{R}}$ for the target causal effect γ_{target}.

is typically more complicated. Moreover, many classical super population literature have propose various post-selection inference strategies, such as sample splitting and selective inference. It is an open research question whether these strategies work under complete randomization. We leave it to future research.

References

Andrews, I., Kitagawa, T., and McCloskey, A. (2019). Inference on winners. Technical report, National Bureau of Economic Research.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111.

Dasgupta, T., Pillai, N. S., and Rubin, D. B. (2015). Causal inference from 2 k factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 727–753.

Egami, N. and Imai, K. (2018). Causal interaction in factorial experiments: Application to conjoint analysis. *Journal of the American Statistical Association*.

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

Hao, N., Feng, Y., and Zhang, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 113(522):615–625.

Hao, N. and Zhang, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301.

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

Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769.

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

Pashley, N. E. and Bind, M.-A. C. (2019). Causal inference for multiple non-randomized treatments using fractional factorial designs. *arXiv e-prints*, pages arXiv–1905.

Shi, L. and Ding, P. (2022). Berry–esseen bounds for design-based causal inference with possibly diverging treatment levels and varying group sizes. *arXiv preprint arXiv:2209.12345*.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

Wei, W., Zhou, Y., Zheng, Z., and Wang, J. (2022). Inference on the best policies with many covariates. *arXiv preprint arXiv:2206.11868*.

Wu, C. J. and Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons.

Yuan, M., Joseph, V. R., and Lin, Y. (2007). An efficient variable selection approach for analyzing designed experiments. *Technometrics*, 49(4):430–439.

Zhao, A. and Ding, P. (2021a). Covariate-adjusted fisher randomization tests for the average treatment effect. *Journal of Econometrics*, 225(2):278–294.

Zhao, A. and Ding, P. (2021b). Regression-based causal inference with factorial experiments: estimands, model specifications, and design-based properties. *arXiv preprint arXiv:2101.02400*.

A General results on consistency of forward screening

We are now ready to show the guarantee for our procedure. Our presentation starts from a theorem (Theorem 6) quantifying the performance of the forward procedure (3.7) under general assumptions on the S-step and P-step. Then we derive a corollary (Theorem 1) by specifying the S-step as Bonferroni corrected marginal t tests and the P-step as screening based on heredity. The high level of the proof proceeds through mathematical induction:

1. (Base case) Show that forward screening selects the correct main effects with probability tending to one.
2. (Induction step) Show that if we correctly screen the effects up to k-way interactions (main effects if $k = 1$), then we are able to correctly detect the non-nulls among all $(k + 1)$-way interactions.

In order to achieve satisfactory screening results, some regularization conditions need to be imposed to characterize a "good" layer-wise S-step, and the P-step should ensure that the procedure progress in a way that is compatible with the structure of the true factorial effects. In light of this, we use $\mathbb{M}^\star_{d,+}$ to denote the pruned set of effects on the d-th layer based on the true model $\mathbb{M}^\star_{d-1}$ on the previous layer; that is,

$$\mathbb{M}^\star_{d,+} = \mathtt{H}(\mathbb{M}^\star_{d-1}).$$

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

Assumption 1 (Validity and consistency of the selection operator)**.** *We denote*

$$\widetilde{\mathbb{M}}_d = \widehat{\mathtt{S}}(\mathbb{M}^\star_{d,+}; \{Y_i, Z_i\}_{i=1}^N),$$

where $\mathbb{M}^\star_{d,+} = \mathtt{H}(\mathbb{M}^\star_{d-1})$ *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*

$$\textit{Validity: } \limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d \cap \mathbb{M}^{\star c}_d \neq \varnothing\right\} \leq \alpha_d, \tag{1.23}$$

$$\textit{Consistency: } \limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}^c_d \cap \mathbb{M}^\star_d \neq \varnothing\right\} = 0. \tag{1.24}$$

This assumption can be verified for many screening procedures. In Theorem 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, a combination of data splitting, adaptation of ℓ_1 regularization and marginal t tests can also fulfill such a requirement (Wasserman and Roeder, 2009).

Besides, we assume the $\mathtt{H}(\cdot)$ operator respects the structure of the nonzero factorial effects:

Assumption 2 (H-heredity)**.** *For $d = 1, \cdots, D-1$, it holds*

$$\mathbb{M}_{d+1}^{\star} \subset \mathtt{P}(\mathbb{M}_d^{\star}).$$

One special case of $\mathtt{H}(\cdot)$ operator satisfying Assumption 2 is naively adding all the the higher order interactions regardless of the lower-order screening results. Besides, if we have evidence that the effects have particular hierarchical structure, applying the corresponding heredity principle such as (3.5) or (3.6) can improve screening accuracy as well as interpretability of the screening results.

Theorem 6 (Screening consistency)**.** *Assume $\mathbb{M}^{\star} \neq \varnothing$. Assume Assumption 1 and 2. Then the forward screening procedure* (3.7) *has the following properties:*

(i) Type I error control. *Forward screening controls the Type I error rate, in the sense that*

$$\limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} \neq \varnothing \text{ for some } d \in [D]\right) \leq \alpha = \sum_{d=1}^{D} \alpha_d. \tag{1.25}$$

(ii) Screening consistency. *Further assume $\alpha = \alpha_N \to 0$. The forward procedure consistently selects all the nonzero effects up to D levels with probability tending to 1:*

$$\limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^{\star} \text{ for all } d \in [D]\right) = 1. \tag{1.26}$$

Theorem 6 consists of two parts. First, one can control the type I error rate, which is defined as the probability of over-selects at least one zero effect. The definition is introduced and elaborated detailedly in Wasserman and Roeder (2009) for model selection. Second, if the tuning parameter $\alpha = \sum_{d=1}^{D} \alpha_d$ vanish asymptotically, one can actually achieve perfect screening up to D levels of effects. To apply Theorem 6 to specific procedures, the key step is to verify Assumption 1 and justify Assumption 2, which we will do for Bonferroni corrected marginal t tests as an example in the next section.

Moreover, the scaling of α_N plays an important role in theoretical discussion. To achieve perfect selection, we hope α_N decays as fast as possible; ideally if α_N equals zero then we do not commit any type I error (or equivalently, we will never select redundant effects). However, for many data-dependent selection procedure α can only decay at certain rates, because a fast decaying α means

higher possibility of rejection, thus can lead to strict under-selection. Therefore, in the tuning process, α_d should be scaled properly if one wants to pursue perfect selection. Nevertheless, even if the tuning is hard and perfect model selection can not be achieved, we still have many strategies to exploit the advantage of the forward screening procedure. We will have more discussions in later sections.

B More results on inference under perfect screening

Lemma 1. *For $\boldsymbol{f}^\star \neq 0$, we have*

$$\frac{\|\boldsymbol{f}^\star\|_\infty}{\|\boldsymbol{f}^\star\|_2} \leq \left(\frac{|\mathbb{M}^\star|}{Q}\right)^{1/2}.$$

Proof of Lemma 1. WLOG assume $\|\boldsymbol{f}\|_2 = 1$. Due to the orthogonality of G, we can decompose $\boldsymbol{f}$ as

$$\boldsymbol{f} = \frac{1}{\sqrt{Q}} G(\cdot, \mathbb{M}^\star) b_1 + \frac{1}{\sqrt{Q}} G(\cdot, \mathbb{M}^{\star c}) b_2,$$

where $b_1 \in \mathbb{R}^{|\mathbb{M}^\star|}$ and $b_2 \in \mathbb{R}^{|\mathbb{M}^{\star c}|}$ and $\|(b_1^\top, b_2^\top)^\top\|_2 = 1$.

Then

$$\boldsymbol{f}^\star = Q^{-1} G(\cdot, \mathbb{M}^\star) G(\cdot, \mathbb{M}^\star)^\top \boldsymbol{f} = \frac{1}{\sqrt{Q}} G(\cdot, \mathbb{M}^\star) b_1.$$

Hence

$$\|\boldsymbol{f}^\star\|_\infty \leq \frac{1}{\sqrt{Q}} \|b_1\|_1, \quad \|\boldsymbol{f}^\star\|_2 = \|b_1\|_2, \quad \frac{\|\boldsymbol{f}^\star\|_\infty}{\|\boldsymbol{f}^\star\|_2} \leq \frac{1}{\sqrt{Q}} \cdot \frac{\|b_1\|_1}{\|b_1\|_2} \leq \left(\frac{|\mathbb{M}^\star|}{Q}\right)^{1/2}.$$

□

Proposition 2 (Asymptotic length of confidence comparison with sparse $\boldsymbol{f}$)**.** *Define the condition number of $V_{\widehat{Y}}$ as $\kappa(V_{\widehat{Y}})$. Let $s^\star$ be the number of nonzero elements in $\boldsymbol{f}$. We have the following upper bound:*

$$\frac{v_{\mathrm{R,lim}}^2}{v_{\mathrm{lim}}^2} \leq \kappa(V_{\widehat{Y}}) \cdot \min\{\frac{s^\star |\mathbb{M}^\star|}{Q}, 1\}.$$

C Technical proofs

C.1 Preliminaries: some important probabilistic results in randomized experiments

Consider an estimator of the form

$$\widehat{\gamma} = Q^{-1} \sum_{\boldsymbol{z} \in \mathcal{T}} w(\boldsymbol{z}) \widehat{Y}(\boldsymbol{z}),$$

with variance estimation

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

It is known that (Li and Ding, 2017)

$$\mathbb{E}\{\widehat{Y}\} = \overline{Y},\ V_{\widehat{Y}} = \mathrm{Var}\left\{\widehat{Y}\right\} = \mathrm{Diag}\left\{N(\boldsymbol{z})^{-1} S(\boldsymbol{z}, \boldsymbol{z})\right\} - N^{-1} S, \tag{3.27}$$

where $S \in \mathbb{R}^{Q \times Q}$ is the covariance matrix for potential outcomes. Then (3.27) further leads to the following facts:

$$\mathbb{E}\{\widehat{\gamma}\} = \sum_{\boldsymbol{z} \in \mathcal{T}} \boldsymbol{f}(\boldsymbol{z}) \overline{Y}(\boldsymbol{z}) = \gamma, \tag{3.28}$$

$$\mathrm{Var}\{\widehat{\gamma}\} = \sum_{\boldsymbol{z} \in \mathcal{T}} \boldsymbol{f}(\boldsymbol{z})^2 N(\boldsymbol{z})^{-1} S(\boldsymbol{z}, \boldsymbol{z}) - N^{-1} \boldsymbol{f}^\top S \boldsymbol{f}, \tag{3.29}$$

$$\mathbb{E}\{\widehat{v}^2\} = \sum_{\boldsymbol{z} \in \mathcal{T}} \boldsymbol{f}(\boldsymbol{z})^2 N(\boldsymbol{z})^{-1} S(\boldsymbol{z}, \boldsymbol{z}). \tag{3.30}$$

We have the following variance estimation results and Berry-Esseen bounds:

Lemma 2 (Variance concentration and Berry-Esseen bounds in finite population)**.** *Denote* $\gamma = \mathbb{E}\{\widehat{\gamma}\}$*,* $v^2 = \mathrm{Var}(\widehat{\gamma})$ *and* $v_R^2 = \mathbb{E}\{\widehat{v}_R^2\}$*. Suppose the following conditions hold:*

- *Nondegenerate variance. There exists a* $\sigma_w > 0$*, such that*

$$Q^{-2} \sum_{\boldsymbol{z}=1}^{Q} w(\boldsymbol{z})^2 N_{\boldsymbol{z}}^{-1} S(\boldsymbol{z}, \boldsymbol{z}) \leq \sigma_w^2 v^2. \tag{3.31}$$

- *Bounded fourth moments. There exists a* $\delta > 0$ *such that*

$$\max_{\boldsymbol{z} \in [Q]} \frac{1}{N} \sum_{i=1}^{N} \{Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\}^4 \leq \Delta^4. \tag{3.32}$$

1. *The variance estimator is robust for the true variance: $v_R \geq v$. Besides, the following tail bound holds:*

$$\mathbb{P}\left\{N|\widehat{v}_R - v_R| > t\right\} \leq \frac{C\overline{c}^3\underline{c}^{-4}\|w\|_\infty^2\Delta^4}{QN_0} \cdot \frac{1}{t^2}. \tag{3.33}$$

2. *We have a Berry-Esseen bound with the true variance:*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v} \leq t\right\} - \Phi(t)\right| \leq 2C\sigma_w \frac{\underline{c}^{-1}\|w\|_\infty \max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\|w\|_2\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}. \tag{3.34}$$

3. *We have a Berry-Esseen bound with the estimated variance: for any $\epsilon_N \in (0, 1/2]$,*

$$\begin{aligned}\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{\widehat{v}_R} \leq t\right\} - \Phi\left(\frac{v_R}{v}t\right)\right| \leq \epsilon_N &+ \frac{C\overline{c}^3\underline{c}^{-4}\|w\|_\infty^2\Delta^4}{QN_0}\cdot\frac{1}{(Nv^2\epsilon_N)^2} \\ &+ 2C\sigma_w\frac{\underline{c}^{-1}\|w\|_\infty \max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\|w\|_2\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}.\end{aligned}$$

Proof of Lemma 2. 1. See Lemma S13 of Shi and Ding (2022).

2. See Theorem 1 of Shi and Ding (2022).

3. Proof of the third part. First we show a useful result: for $|a| \leq 1/2$ and any $b \in \mathbb{R}$,

$$\sup_{t\in\mathbb{R}}|\Phi\{(1+a)t+b\} - \Phi\{t\}| \leq |a| + |b|. \tag{3.35}$$

This can be proved by a simple step of intermediate value theorem: for any $t \in \mathbb{R}$,

$$\begin{aligned}&|\Phi\{(1+a)t+b\} - \Phi\{t\}| \\ =&|\phi(\xi_{t,(1+a)t})\cdot(at+b)| \\ =&|\phi(\xi_{t,(1+a)t})\cdot at| + |\phi(\xi_{t,(1+a)t})\cdot b| \\ =&|a|\cdot|\phi(\xi_{t,(1+a)t})\cdot t|\cdot\mathbf{1}\{|t|\leq 1\} + |a|\cdot|\phi(\xi_{t,(1+a)t})\cdot t|\cdot\mathbf{1}\{|t|>1\} + |\phi(\xi_{t,(1+a)t})\cdot b| \\ \leq&\frac{1}{\sqrt{2\pi}}|a|\cdot\mathbf{1}\{|t|\leq 1\} + \frac{1}{\sqrt{2\pi}}|a||t|\cdot\exp(-t^2/8)\cdot\mathbf{1}\{|t|>1\} + \frac{1}{\sqrt{2\pi}}|b| \\ \leq&|a|+|b|.\end{aligned}$$

WLOG we consider $t \geq 0$ because $t < 0$ can be handled similarly. For any $\epsilon_N > 0$, We have

$$\begin{aligned}\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{\widehat{v}_R} \leq t\right\} &= \mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v} \leq \frac{\widehat{v}_R}{v}t\right\} \\ &= \mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v} \leq \frac{\widehat{v}_R}{v}t, \left|\frac{\widehat{v}_R - v_R}{v}\right| \leq \epsilon_N\right\} + \mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v} \leq \frac{\widehat{v}_R}{v}t, \left|\frac{\widehat{v}_R - v_R}{v}\right| > \epsilon_N\right\}.\end{aligned}$$

Then we can show that

$$\begin{aligned}\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{\widehat{v}_R}\le t\right\} &\le \mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v}\le \frac{\widehat{v}_R}{v}t, \left|\frac{\widehat{v}_R-v_R}{v}\right|\le \epsilon_N\right\}+\mathbb{P}\left\{\left|\frac{\widehat{v}_R-v_R}{v}\right|>\epsilon_N\right\}\\ &\le \mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v}\le \left(\frac{v_R}{v}+\epsilon_N\right)t\right\}+\mathbb{P}\left\{\left|\frac{\widehat{v}_R-v_R}{v}\right|>\epsilon_N\right\}.\end{aligned}$$

For the first term, we have

$$\begin{aligned}&\sup_{t\ge 0}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v}\le\left(\frac{v_R}{v}+\epsilon_N\right)t\right\}-\Phi\left\{\left(\frac{v_R}{v}+\epsilon_N\right)t\right\}\right|\\ &\le 2C\sigma_w\frac{\underline{c}^{-1}\|w\|_\infty\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\|w\|_2\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}.\end{aligned}$$

For the second term, using the variance estimation results in Part 1 we have

$$\begin{aligned}\mathbb{P}\left\{\left|\frac{\widehat{v}_R-v_R}{v}\right|\ge\epsilon_N\right\} &\le \mathbb{P}\left\{\left|\frac{\widehat{v}_R-v_R}{v}\right|\cdot\left|\frac{\widehat{v}_R+v_R}{v}\right|\ge\epsilon_N\right\}\ \text{(because } v_R \text{ is robust)}\\ &=\mathbb{P}\left\{\left|\frac{N\widehat{v}_R^2-Nv_R^2}{Nv^2}\right|\ge\epsilon_N\right\}\\ &\le\frac{C\overline{c}^3\underline{c}^{-4}\|w\|_\infty^2\Delta^4}{QN_0}\cdot\frac{1}{(Nv^2\epsilon_N)^2}.\end{aligned}$$

Besides, by (3.35), when $\epsilon_N\le 1/2$, we also have

$$\sup_{t\in\mathbb{R}}\left|\Phi\left\{\left(\frac{v_R}{v}+\epsilon_N\right)t\right\}-\Phi\left(\frac{v_R}{v}t\right)\right|\le\frac{v\epsilon_N}{v_R}\le\epsilon_N.$$

Aggregating all the parts above, we can show that for any $t\ge 0$,

$$\begin{aligned}\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{\widehat{v}_R}\le t\right\} &\le \Phi\left(\frac{v_R}{v}t\right)+\epsilon_N+\frac{C\overline{c}^3\underline{c}^{-4}\|w\|_\infty^2\Delta^4}{QN_0}\cdot\frac{1}{(Nv^2\epsilon_N)^2}\\ &\quad+2C\sigma_w\frac{\underline{c}^{-1}\|w\|_\infty\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\|w\|_2\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}.\end{aligned}$$

On the other hand, we can show that

$$\begin{aligned}\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{\widehat{v}_R}\le t\right\} &\ge \mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v}\le\frac{\widehat{v}_R}{v}t,\left|\frac{\widehat{v}_R-v_R}{v}\right|\le\epsilon_N\right\}\\ &\ge\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v}\le\left(\frac{v_R}{v}-\epsilon_N\right)t\right\}-\mathbb{P}\left\{\left|\frac{\widehat{v}_R-v_R}{v}\right|\ge\epsilon_N\right\}.\end{aligned}\tag{3.36}$$

By (3.35), when $\epsilon_N\le 1/2$, we also have

$$\sup_{t\in\mathbb{R}}\left|\Phi\left\{\left(\frac{v_R}{v}-\epsilon_N\right)t\right\}-\Phi\left(\frac{v_R}{v}t\right)\right|\le\epsilon_N.$$

So we can derive a lower bound analogous to (3.36). Note that the results can be analogously generalized to $t \leq 0$. Putting pieces together, we can show that for any $t \geq 0$,

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{\widehat{v}_R}\leq t\right\}-\Phi\left(\frac{v_R}{v}t\right)\right| \leq \epsilon_N + \frac{C\overline{c}^3\underline{c}^{-4}\|w\|_\infty^2\Delta^4}{QN_0}\cdot\frac{1}{(Nv^2\epsilon_N)^2} + 2C\sigma_w\frac{\underline{c}^{-1}\|w\|_\infty \max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\|w\|_2\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}.$$

□

The following corollary shows the studentized Berry-Esseen bounds in the special case where $w=(w(\boldsymbol{z}))_{\boldsymbol{z}\in[Q]}$ is a contrast vector for factorial effects. That is, $w=g_{\mathcal{K}}$ for some $\mathcal{K}\in\mathbb{K}$.

Corollary 1. *Assume Condition* (3.31) *and* (3.32) *hold. Let $w=g_{\mathcal{K}}$ for some $\mathcal{K}\in\mathbb{K}$. Then we have a Berry-Esseen bound with the estimated variance:*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\tau}_{\mathcal{K}}-\tau_{\mathcal{K}}}{\widehat{v}_R}\leq t\right\}-\Phi\left(\frac{v_R}{v}t\right)\right| \leq 2\left(\frac{C\sigma_w^4\overline{c}^5\underline{c}^{-6}\Delta^4}{\{\min_{\boldsymbol{z}\in\mathcal{T}}S(\boldsymbol{z},\boldsymbol{z})\}^2}\right)^{1/3}\cdot\frac{1}{(QN_0)^{1/3}} + 2C\sigma_w\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}}\cdot\frac{1}{(QN_0)^{1/2}}.$$

Proof of Corollary 1. **Lower bound for Nv^2.** Note that $\|w\|_2^2=Q, \|w\|_\infty=1$. Using Condition (3.31), we have

$$\begin{aligned} Nv^2 &\geq N\sigma_w^{-2}Q^{-2}\sum_{\boldsymbol{z}=1}^{Q}w(\boldsymbol{z})^2N_{\boldsymbol{z}}^{-1}S(\boldsymbol{z},\boldsymbol{z}) \\ &\geq (\underline{c}QN_0)\cdot\sigma_w^{-2}\overline{c}^{-1}Q^{-1}N_0^{-1}\min_{\boldsymbol{z}\in\mathcal{T}}S(\boldsymbol{z},\boldsymbol{z})\cdot(Q^{-1}\|w\|_2^2) \\ &= \sigma_w^{-2}\underline{c}\overline{c}^{-1}\min_{\boldsymbol{z}\in\mathcal{T}}S(\boldsymbol{z},\boldsymbol{z}). \end{aligned}$$

Therefore, the Berry-Esseen bound becomes

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\tau}_{\mathcal{K}}-\tau_{\mathcal{K}}}{\widehat{v}_R}\leq t\right\}-\Phi\left(\frac{v_R}{v}t\right)\right| \leq \epsilon_N + \frac{C\sigma_w^4\overline{c}^5\underline{c}^{-6}\Delta^4}{(QN_0)\{\min_{\boldsymbol{z}\in\mathcal{T}}S(\boldsymbol{z},\boldsymbol{z})\}^2}\cdot\frac{1}{\epsilon_N^2} + 2C\sigma_w\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{QN_0}}.$$

Optimize the summation of the first and second term. By taking derivative over ϵ_N on the upper bound and solving for the zero point, we know that when

$$\epsilon_N=\left(\frac{2C\sigma_w^4\overline{c}^5\underline{c}^{-6}\Delta^4}{(QN_0)\{\min_{\boldsymbol{z}\in\mathcal{T}}S(\boldsymbol{z},\boldsymbol{z})\}^2}\right)^{1/3},$$

the upper bound is minimized and

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\tau}_{\mathcal{K}}-\tau_{\mathcal{K}}}{\widehat{v}_R}\le t\right\}-\Phi\left(\frac{v_R}{v}t\right)\right|\le 2\left(\frac{C\sigma_w^4\overline{c}^5\underline{c}^{-6}\Delta^4}{\{\min_{\boldsymbol{z}\in\mathcal{T}}S(\boldsymbol{z},\boldsymbol{z})\}^2}\right)^{1/3}\cdot\frac{1}{(QN_0)^{1/3}}$$
$$+2C\sigma_w\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}}\cdot\frac{1}{(QN_0)^{1/2}}.$$

□

Additionally, we have a Berry-Esseen bounds after screening the effects:

Lemma 3 (Berry Esseen bound with screening)**.** *Let*

$$\boldsymbol{f}[\mathbb{M}]=Q^{-1}G(\cdot,\mathbb{M})G(\cdot,\mathbb{M})^\top\boldsymbol{f}. \tag{3.37}$$

Assume there exists $\sigma_w>0$ such that

$$\sum_{\boldsymbol{z}=1}^{Q}\boldsymbol{f}[\mathbb{M}](\boldsymbol{z})^2N_{\boldsymbol{z}}^{-1}S(\boldsymbol{z},\boldsymbol{z})\le\sigma_w^2v^2(\mathbb{M}). \tag{3.38}$$

Then it holds

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma[\mathbb{M}]}{v(\mathbb{M})}\le t\right\}-\Phi(t)\right|$$
$$\le 2\mathbb{P}\left\{\widehat{\mathbb{M}}\neq\mathbb{M}\right\}+2C\sigma_w\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}\cdot\frac{\|\boldsymbol{f}[\mathbb{M}]\|_\infty}{\|\boldsymbol{f}[\mathbb{M}]\|_2}. \tag{3.39}$$

Proof of Lemma 3. With the selected working model we have

$$\begin{aligned}
&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma[\mathbb{M}]}{v(\mathbb{M})}\le t\right\}-\Phi(t)\right|\\
=&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma[\mathbb{M}]}{v(\mathbb{M})}\le t,\widehat{\mathbb{M}}=\mathbb{M}\right\}-\Phi(t)+\mathbb{P}\left\{\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma[\mathbb{M}]}{v(\mathbb{M})}\le t,\widehat{\mathbb{M}}\neq\mathbb{M}\right\}\right|\\
\le&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma[\mathbb{M}]}{v(\mathbb{M})}\le t,\widehat{\mathbb{M}}=\mathbb{M}\right\}-\Phi(t)\right|+\mathbb{P}\left\{\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma[\mathbb{M}]}{v(\mathbb{M})}\le t,\widehat{\mathbb{M}}\neq\mathbb{M}\right\}\\
=&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}[\mathbb{M}]-\gamma[\mathbb{M}]}{v(\mathbb{M})}\le t,\widehat{\mathbb{M}}=\mathbb{M}\right\}-\Phi(t)\right|+\mathbb{P}\left\{\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma[\mathbb{M}]}{v(\mathbb{M})}\le t,\widehat{\mathbb{M}}\neq\mathbb{M}\right\}\\
\le&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}[\mathbb{M}]-\gamma[\mathbb{M}]}{v(\mathbb{M})}\le t\right\}-\Phi(t)\right|+2\mathbb{P}\left\{\widehat{\mathbb{M}}\neq\mathbb{M}\right\}.
\end{aligned}$$

Now we have

$$\begin{aligned}
\widehat{\gamma}(\mathbb{M})&=\boldsymbol{f}^\top G(\cdot,\mathbb{M})\widehat{\tau}(\mathbb{M})\\
&=\boldsymbol{f}^\top G(\cdot,\mathbb{M})G(\cdot,\mathbb{M})^\top\widehat{Y}\\
&=\boldsymbol{f}[\mathbb{M}]^\top\widehat{Y}.
\end{aligned}$$

By Theorem 1 of Shi and Ding (2022), we have a Berry-Esseen bound with the true variance:

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}(\mathbb{M})-\gamma[\mathbb{M}]}{v}\leq t\right\}-\Phi(t)\right|\leq 2C\sigma_w\frac{\|\boldsymbol{f}[\mathbb{M}]\|_\infty \underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\|\boldsymbol{f}[\mathbb{M}]\|_2\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}.$$

□

C.2 Proof of Theorem 6

Proof of Theorem 6. **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

$$\left|\mathbb{P}\left(\widehat{\mathbb{M}}_d\subset\mathbb{M}_d^\star,d=1,\cdots,D_0\right)-\mathbb{P}\left(\widehat{\mathbb{M}}_d=\mathbb{M}_d^\star,d=1,\cdots,D_0\right)\right|\to 0. \tag{3.40}$$

Because for any $D_0\in[D]$, we always have:

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

(3.40) is equivalent to: for all $D_0\leq D$,

$$\mathbb{P}\left(\text{for all } d\in[D_0],\widehat{\mathbb{M}}_d\subset\mathbb{M}_d^\star;\ \text{there exists } d\in[D_0],\widehat{\mathbb{M}}_d\subsetneq\mathbb{M}_d^\star\right)\to 0. \tag{3.41}$$

1. **Main effects.** First, because we assume the tests are consistent (Assumption 1), meaning asymptotically no false negatives:

$$\mathbb{P}\left(\widehat{\mathbb{M}}_1^c\cap\mathbb{M}_1^\star\neq\varnothing\right)\to 0\Leftrightarrow\mathbb{P}\left(\mathbb{M}_1^\star\subset\widehat{\mathbb{M}}_1\right)\to 1.$$

Therefore,

$$\mathbb{P}\left(\widehat{\mathbb{M}}_1\subsetneq\mathbb{M}_1^\star\right)\to 0.\ \text{(no under selection for main effects)} \tag{3.42}$$

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 (3.40) or (3.41) for some $D_0 \geq 1$. For $D_0 + 1$, the following holds:

$$\begin{aligned}
0 &\leq \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0 + 1\right\}\right) - \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right) \\
&= \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0 + 1\right\} - \left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right) \\
&\leq \mathbb{P}\left(\forall d \in [D_0 + 1], \widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star; \exists d \in [D_0], \widehat{\mathbb{M}}_d \subsetneq \mathbb{M}_d^\star\right) \\
&\leq \mathbb{P}\left(\forall d \in [D_0], \widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star; \exists d \in [D_0], \widehat{\mathbb{M}}_d \subsetneq \mathbb{M}_d^\star\right) \to 0. \text{ (by (3.41))}
\end{aligned}$$

Hence

$$\left|\mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0 + 1\right\}\right) - \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right)\right| \to 0. \quad (3.43)$$

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 $\widehat{\mathbb{M}}_{D_0+1,+}$. Under Assumption 2, 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

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

(using (3.40) and the fact that D is a fixed integer)

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

(on the given event, $\widehat{\mathbb{M}}_{D_0,+} = \mathcal{P}(\widehat{\mathbb{M}}_{D_0-1}) = \mathcal{P}(\mathbb{M}_{D_0-1}^\star) = \mathbb{M}_{D_0,+}^\star$ and $\widehat{\mathbb{M}}_{D_0} = \mathcal{S}_N(\widehat{\mathbb{M}}_{D_0,+}) = \widetilde{\mathbb{M}}_{D_0}$)

$$\leq \limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_1 \cap \mathbb{M}_1^{\star c} \neq \varnothing\right) + \sum_{D_0=2}^{D} \mathbb{P}\left(\widetilde{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star\ c} \neq \varnothing\right) \leq \sum_{D_0=1}^{D} \alpha_{D_0} = \alpha. \tag{3.44}$$

Therefore the target probability gets controlled under α.

Proof of the second result. Under $\alpha = \alpha_N \to 0$, (3.44) implies that with probability tending to one,

$$\widehat{\mathbb{M}}_d \cap (\mathbb{M}_d^\star)^c = \varnothing, \text{ for } d = 1, \cdots, D \Leftrightarrow \widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star = \varnothing, \text{ for } d = 1, \cdots, D.$$

Now apply (3.40), 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.

□

C.3 Proof of Theorem 1

We state and prove a more general version of Theorem 1:

Theorem 7 (Bonferroni corrected marginal t test)**.** *Let* $\widetilde{\mathbb{M}}_d = \widehat{\mathtt{S}}(\mathbb{M}_{d,+}^\star)$ *where* $\mathbb{M}_{d,+}^\star = \mathtt{P}(\mathbb{M}_{d-1}^\star)$. *Assume Conditions 1, 2, 3 and 4. Then we have the following results for the screening procedure based on Bonferroni corrected marginal t-test:*

(i) (Validity) $\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} \neq \varnothing\right\} \leq \alpha_d$ *for all* $d = 1, \cdots, D$.

(ii) (Consistency) $\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\} = 0$ *for all* $d = 1, \cdots, D$.

(iii) (Type I error control) Overall the procedure achieves type I error rate control:

$$\limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}} \cap (\cup_{d=1}^D \mathbb{M}_d^\star)^c \neq \varnothing\right) \leq \alpha.$$

(iv) (Perfect selection) When δ' *is strictly positive, we have* $\max_{d\in[D]} \alpha_d \to 0$ *and*

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

Part (i) and ii of Theorem 1 justified Assumption 1 and 2 respectively, which build up the basis for applying Theorem 6. Part (iii) guarantees type I error control under the significance level α. When we let α decay to zero, Part (iii) implies that we will not include redundant terms into the selected working model. Part (iv) further states a stronger result with vanishing α - perfect selection can be achieved asymptotically.

Proof of Theorem 1. 1. First, we show validity:

$$\begin{aligned}
\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} \neq \varnothing\right\} &= \limsup_{N\to\infty} \mathbb{P}\left\{\exists \mathcal{K} \in \mathbb{M}_{d,+}^{\star}\backslash\mathbb{M}_d^{\star}, \left|\frac{\widehat{\tau}_{\mathcal{K}}}{\widehat{\sigma}_{\mathcal{K}}}\right| \geq \Phi^{-1}\left(1-\frac{\alpha_d}{2|\mathbb{M}_{d,+}^{\star}|}\right)\right\} \\
&\leq \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_{d,+}^{\star}\backslash\mathbb{M}_d^{\star}} \mathbb{P}\left\{\left|\frac{\widehat{\tau}_{\mathcal{K}}}{\widehat{\sigma}_{\mathcal{K}}}\right| \geq \Phi^{-1}\left(1-\frac{\alpha_d}{2|\mathbb{M}_{d,+}^{\star}|}\right)\right\} \\
&\leq \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_{d,+}^{\star}\backslash\mathbb{M}_d^{\star}} \left(\frac{\alpha_d}{|\mathbb{M}_{d,+}^{\star}|} + \frac{\widetilde{C}}{(QN_0)^{1/3}}\right) \leq \alpha_d.
\end{aligned}$$

2. Second, we show consistency.

$$\begin{aligned}
&\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^{\star} \neq \varnothing\right\} \\
&= \limsup_{N\to\infty} \mathbb{P}\left\{\exists \mathcal{K} \in \mathbb{M}_d^{\star}, \left|\frac{\widehat{\tau}_{\mathcal{K}}}{\widehat{\sigma}_{\mathcal{K}}}\right| \leq \Phi^{-1}\left(1-\frac{\alpha_d}{2|\mathbb{M}_{d,+}^{\star}|}\right)\right\} \\
&\leq \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^{\star}} \mathbb{P}\left\{\left|\frac{\widehat{\tau}_{\mathcal{K}}}{\widehat{\sigma}_{\mathcal{K}}}\right| \leq \Phi^{-1}\left(1-\frac{\alpha_d}{2|\mathbb{M}_{d,+}^{\star}|}\right)\right\} \\
&\leq \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^{\star}} \mathbb{P}\left\{\left|\frac{\widehat{\tau}_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right| \leq \frac{\widehat{\sigma}_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\Phi^{-1}\left(1-\frac{\alpha_d}{2|\mathbb{M}_{d,+}^{\star}|}\right)\right\} \\
&\leq \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^{\star}} \mathbb{P}\left\{\left|\frac{\widehat{\tau}_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right| \leq \left\{1+\frac{\widetilde{C}}{(QN_0)^{1/3}}\right\}\Phi^{-1}\left(1-\frac{\alpha_d}{2|\mathbb{M}_{d,+}^{\star}|}\right)\right\} + \mathbb{P}\left\{\frac{\widehat{\sigma}_{\mathcal{K}}}{\sigma_{\mathcal{K}}} > 1+\frac{\widetilde{C}}{(QN_0)^{1/3}}\right\}.
\end{aligned}$$

For simplicity, let

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

Then

$$\begin{aligned}
&\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_k^{\star} \neq \varnothing\right\} \\
&\leq \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^{\star}} \left(\mathbb{P}\left\{-Z_d^{\star} - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}} \leq \frac{\widehat{\tau}_{\mathcal{K}}}{\sigma_{\mathcal{K}}} - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}} \leq Z_d^{\star} - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right\} + \frac{\widetilde{C}}{(QN_0)^{1/3}}\right) \\
&= \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^{\star}} \Phi\left\{r_{\mathcal{K}}^{-1}\left(Z_d^{\star} - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\} - \Phi\left\{r_{\mathcal{K}}^{-1}\left(-Z_d^{\star} - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\}. \qquad (3.45)
\end{aligned}$$

With Condition 2, we have

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

Because $\delta > -1/2$ and $\delta' \geq 0$, 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 (3.45) converges to zero. This concludes the proof.

3. Based on the above two parts and Theorem 6, 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 (3.44), we directly compute

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

4. The perfect selection result follows from Part 1,2 and Theorem 6.

□

C.4 Proof of Theorem 2

Theorem 2 is a direct result of Lemma 2 and the following Berry Esseen result:

Lemma 4 (Berry-Esseen bound under perfect screening)**.** *Let* $\boldsymbol{f}[\mathbb{M}]$ *be given by* (3.37)*. Assume* (3.38)*. Then it holds*

$$\begin{aligned}
&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}(\widehat{\mathbb{M}})-\gamma}{v(\mathbb{M}^\star)} \leq t\right\} - \Phi(t)\right| \\
&\leq 2\mathbb{P}\left\{\widehat{\mathbb{M}} \neq \mathbb{M}^\star\right\} + 2C\sigma_w\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\cdot\sqrt{N_0}}}\cdot\frac{\|\boldsymbol{f}[\mathbb{M}^\star]\|_\infty}{\|\boldsymbol{f}[\mathbb{M}^\star]\|_2}.
\end{aligned} \tag{3.46}$$

Proof of Lemma 4. This lemma is a direct application of Lemma 3. First we check that

$$\gamma(\mathbb{M}^\star) = \gamma.$$

From the definition of γ (3.28), we have

$$\begin{aligned}\gamma &= \boldsymbol{f}^\top \overline{Y} \\ &= \boldsymbol{f}^\top G\tau = \boldsymbol{f}^\top G(\cdot, \mathbb{M}^\star)\tau(\mathbb{M}^\star) \\ &= Q^{-1}\boldsymbol{f}^\top G(\cdot, \mathbb{M}^\star)G(\cdot, \mathbb{M}^\star)^\top \overline{Y} = \gamma(\mathbb{M}^\star).\end{aligned}$$

Now apply Lemma 3 with $\mathbb{M} = \mathbb{M}^\star$ gives the result. □

C.5 Proof of Theorem 3

Proof of Theorem 3. According to Condition 6 and Theorem 1, with Strategy 1,

$$\mathbb{P}\left\{\widehat{\mathbb{M}} = \cup_{d=1}^{d^\star}\mathbb{M}_d^\star\right\} \to 1.$$

We will apply Lemma 3 with

$$\underline{\mathbb{M}}^\star = \cup_{d=1}^{d^\star}\mathbb{M}_d^\star.$$

We only need to verify $\gamma = \gamma[\mathbb{M}]$ in this case.

$$\begin{aligned}\gamma &= \boldsymbol{f}^\top \overline{Y} \\ &= \boldsymbol{f}^\top G\tau = \boldsymbol{f}^\top G(\cdot, \underline{\mathbb{M}}^\star)\tau(\underline{\mathbb{M}}^\star) + \boldsymbol{f}^\top G(\cdot, \underline{\mathbb{M}}^{\star c})\tau(\underline{\mathbb{M}}^{\star c}).\end{aligned}$$

Now by (5.20), $\boldsymbol{f}^\top G(\cdot, \mathbb{M}^c) = 0$. Hence

$$\gamma = Q^{-1}\boldsymbol{f}^\top G(\cdot, \cup_{d=1}^{d^\star}\mathbb{M}_d^\star)G(\cdot, \cup_{d=1}^{d^\star}\mathbb{M}_d^\star)^\top \overline{Y} = \gamma.$$

□

C.6 Proof of Theorem 4

Proof of Theorem 4. This proof can be finished by applying Lemma 3 with $\mathbb{M} = \overline{\mathbb{M}}^\star$ and checking $\gamma[\overline{\mathbb{M}}^\star] = \gamma$, which is omitted here. □

C.7 Proof of Proposition 1

Proof of Proposition 1. WLOG we assume only the first $s^\star$ elements of $\boldsymbol{f}$ are nonzero. That is,

$$\boldsymbol{f} = f_1\boldsymbol{e}_1 + \cdots + f_{s^\star}\boldsymbol{e}_{s^\star}. \tag{3.47}$$

We can verify that

$$\|Q^{-1}G(\cdot,\mathbb{M}^\star)G(\cdot,\mathbb{M}^\star)^\top \boldsymbol{e}_k\|_2 = \frac{|\mathbb{M}^\star|}{Q}, \quad \forall\ k \in [Q]. \tag{3.48}$$

Therefore,

$$\frac{v_{\mathrm{R}}^2}{v^2} = \frac{\boldsymbol{f}^\star V_{\widehat{Y}} \boldsymbol{f}^\star}{\boldsymbol{f} V_{\widehat{Y}} \boldsymbol{f}} \leq \frac{\varrho_{\max}(V_{\widehat{Y}})\|\boldsymbol{f}^\star\|_2^2}{\varrho_{\min}(V_{\widehat{Y}})\|\boldsymbol{f}\|_2^2} = \kappa(V_{\widehat{Y}}) \cdot \frac{\|\boldsymbol{f}^\star\|_2^2}{\|\boldsymbol{f}\|_2^2}.$$

On one hand, using $Q^{-1}G(\cdot,\mathbb{M}^\star)G(\cdot,\mathbb{M}^\star)^\top \preccurlyeq I_Q$, we immediately have

$$\frac{\|\boldsymbol{f}^\star\|_2^2}{\|\boldsymbol{f}\|_2^2} \leq 1. \tag{3.49}$$

On the other hand, using (3.47) and (3.48), we have

$$\frac{\|\boldsymbol{f}^\star\|_2^2}{\|\boldsymbol{f}\|_2^2} \leq \frac{\|\boldsymbol{f}\|_1^2}{\|\boldsymbol{f}\|_2^2} \cdot \frac{|\mathbb{M}^\star|}{Q} \leq \frac{s^\star|\mathbb{M}^\star|}{Q}. \tag{3.50}$$

Combining (3.49) and (3.50) concludes the proof. □

C.8 Proof of Theorem 5

For simplicity, we focus on the case given by (6.21). The general proof can be completed similarly. We begin by the following lemma:

Lemma 5 (Consistency of the selected tie sets)**.** *Assume Conditions 1, 3 and 7. There exists universal constants $C, C' > 0$, such that when $N > n(\delta_1, \delta_2, \delta_3)$,*

$$\begin{aligned}\mathbb{P}\left\{\widehat{\mathcal{T}}_1 = \mathcal{T}_1\right\} \geq & 1 - \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^\star\} \\ & - C|\mathcal{T}'||\mathcal{T}_1|\left\{\sqrt{\frac{\bar{c}\Delta|\mathbb{M}^\star|}{N^{1+2\delta_2}}}\exp\left(-\frac{C'N^{1+2\delta_2}}{\bar{c}\Delta|\mathbb{M}^\star|}\right) + \sigma\frac{\underline{c}^{-1/2}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}} \cdot \sqrt{\frac{|\mathbb{M}^\star|}{N}}\right\}.\end{aligned}$$

Lemma 5 establishes a finite sample bound to quantify the performance of the tie set selection step in Algorithm 2. The tail bound implies that the performance of tie selection depends on several elements:

- Quality of effect screening. Ideally we hope perfect screening can be achieved. In other words, the misspecification probability $\mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^\star\}$ is small in an asymptotic sense.
- Size of the tie $|\mathcal{T}_1|$ and the number of factor combinations considered $|\mathcal{T}'|$. These two quantities play a natural role because one can expect the difficulty of selection will increase if there are too many combinations present in the first tie or involved in comparison.

- Size of between-group distance $d_h^\star$. If the gap between $\overline{Y}_{(1)}$ and the remaining order values are large, $\eta_N = \Theta(N^{\delta_2})$ is allowed to take larger values and the term

$$\sqrt{\frac{\bar{c}\Delta|\mathbb{M}^\star|}{N^{1+2\delta_2}}}\exp\left(-\frac{C'N^{1+2\delta_2}}{\bar{c}\Delta|\mathbb{M}^\star|}\right)$$

  can become smaller in magnitude.

- Population level property of potential outcomes. The scale of the centered potential outcomes $|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|$ should be controlled, and the population variance $S(\boldsymbol{z},\boldsymbol{z})$ should be non-degenerate.

- The relative scale between number of nonzero effects $|\mathbb{M}^\star|$ and the total number of units N. The larger N is compared to $|\mathbb{M}^\star|$, the easier for us to draw valid asymptotic conclusions.

Proof of Lemma 5. The high level idea of the proof is: we first prove the non-asymptotic bounds over the random event $\widehat{\mathbb{M}} = \mathbb{M}^\star$, then make up for the cost of $\widehat{\mathbb{M}} \neq \mathbb{M}^\star$. Over $\widehat{\mathbb{M}} = \mathbb{M}^\star$, we have

$$\widehat{Y}_{\mathrm{r}} = \widehat{Y}_{\mathrm{r}}^\star = G(\cdot,\mathbb{M}^\star)\widehat{\tau}(\mathbb{M}^\star) = Q^{-1}G(\cdot,\mathbb{M}^\star)G(\cdot,\mathbb{M}^\star)^\top\widehat{Y}.$$

We apply Lemma 3 to establish a Berry-Esseen bound for each $\widehat{Y}_{\mathrm{r}}^\star(\boldsymbol{z})$. Note that

$$\widehat{Y}_{\mathrm{r}}^\star(\boldsymbol{z}) = \boldsymbol{f}_{\boldsymbol{z}}^\top\widehat{Y},\ \boldsymbol{f}_{\boldsymbol{z}}^\top = Q^{-1}G(\boldsymbol{z},\mathbb{M}^\star)G(\cdot,\mathbb{M}^\star)^\top. \tag{3.51}$$

By calculation we have

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

Also we can show that

$$\sum_{\boldsymbol{z}'=1}^{Q}\boldsymbol{f}_{\boldsymbol{z}}(\boldsymbol{z}')^2N_{\boldsymbol{z}'}^{-1}S(\boldsymbol{z}',\boldsymbol{z}') \leq \sigma^2v^2(\mathbb{M}).$$

and obtain

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{Y}_{\mathrm{r}}^\star(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})}{v_N} \leq t\right\} - \Phi(t)\right| \leq 2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\sqrt{\frac{|\mathbb{M}^\star|}{QN_0}}. \tag{3.52}$$

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

$$\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}'\setminus\mathcal{T}_1}\widehat{Y}_{\mathrm{r}}^\star(\boldsymbol{z}) < \min_{\boldsymbol{z}\in\mathcal{T}_1}\widehat{Y}_{\mathrm{r}}^\star(\boldsymbol{z}) \leq \max_{\boldsymbol{z}\in\mathcal{T}_1}\widehat{Y}_{\mathrm{r}}^\star(\boldsymbol{z})\right\}.$$

We can show that

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

Hence we know that

$$\begin{aligned}
&\mathbb{P}\left\{\sqrt{N}\left|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})\right|\ge\sqrt{N}d_h^{\star}\right\}\\
\le&\frac{v_N}{\sqrt{2\pi}d_h^{\star}}\cdot\exp\left(-\frac{d_h^{\star 2}}{2v_N^2}\right)+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}. \qquad (3.53)
\end{aligned}$$

Therefore, for all $\boldsymbol{z}\in\mathcal{T}'\backslash\mathcal{T}$ and $\boldsymbol{z}'\in\mathcal{T}_1$,

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

Using (3.53) we have

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}')-\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})<0\right\}\\
\le&\frac{\sqrt{\overline{c}\Delta|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0Qd_h^{\star 2}}{2\overline{c}\overline{s}|\mathbb{M}^{\star}|}\right)+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}.
\end{aligned}$$

Now a union bound gives

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}')-\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})<0\right\}\\
&\ge 1-|\mathcal{T}_1||\mathcal{T}'|\left\{\frac{\sqrt{\overline{c}\overline{s}|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0Qd_h^{\star 2}}{2\overline{c}\overline{s}|\mathbb{M}^{\star}|}\right)+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}\right\}.
\end{aligned}$$

Now using that $d_h^{\star}=\Theta(N^{\delta_1})$, $Nd_h^{\star 2}=\Theta(N^{1+2\delta_1})$ with $1+2\delta_1>0$. The first term in the bracket has the following order

$$\frac{\sqrt{\overline{c}\overline{s}|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0Qd_h^{\star 2}}{2\overline{c}\overline{s}|\mathbb{M}^{\star}|}\right)=\Theta\left(\sqrt{\frac{\overline{c}\overline{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_1}}}\exp\left\{-\frac{C'N^{1+2\delta_1}}{\overline{c}\overline{s}|\mathbb{M}^{\star}|}\right\}\right)$$

where $C' > 0$ is a universal constant due to Condition 2.Note that $\delta_2 > \delta_1$. Thus when N is large enough, we have

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}') - \widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) < 0\right\} \\
\geq& 1 - C|\mathcal{T}_1||\mathcal{T}'|\left\{\sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_1}}}\exp\left\{-\frac{C'N^{1+2\delta_1}}{\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right\} + \sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned} \tag{3.54}$$

Nice separation. Suppose we are working on a random coordinate $\widetilde{\boldsymbol{z}}$. For $\boldsymbol{z} \notin \mathcal{T}_1$ and any $\bar{\epsilon} > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \widehat{Y}_{\mathrm{r}}^{\star}(\widetilde{\boldsymbol{z}})|/\eta_N \geq 2\bar{\epsilon}\right\} \\
&\geq \mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1}|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}')|/\eta_N \geq 2\bar{\epsilon}, \widetilde{m}\in\mathcal{T}_1\right\} \\
&\geq \mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1}|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}')|/\eta_N \geq 2\bar{\epsilon}\right\} + \mathbb{P}\left\{\widetilde{m}\in\mathcal{T}_1\right\} - 1 \\
&\geq \mathbb{P}\left\{\widetilde{m}\in\mathcal{T}_1\right\} - \sum_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1}\mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \widehat{Y}_{\mathrm{r}}^{\star}(\widetilde{\boldsymbol{z}}')|/\eta_N \leq 2\bar{\epsilon}\right\}.
\end{aligned} \tag{3.55}$$

To proceed we have the following tail bound:

$$\begin{aligned}
&\mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}')|/\eta_N \leq 2\bar{\epsilon}\right\} \\
=&\mathbb{P}\left\{|\{\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\} - \{\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}') - \overline{Y}(\boldsymbol{z}')\} - \{\overline{Y}(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z}')\}| \leq 2\bar{\epsilon}\eta_N\right\} \\
\leq&\mathbb{P}\left\{|\overline{Y}(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z}')| - |\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})| - |\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}') - \overline{Y}(\boldsymbol{z}')| \leq 2\bar{\epsilon}\eta_N\right\} \\
\leq&\mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})| + |\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}') - \overline{Y}(\boldsymbol{z}')| \geq 2d_h^{\star} - 2\bar{\epsilon}\eta_N\right\} \\
\leq&\mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})| \geq d_h^{\star} - \bar{\epsilon}\eta_N\right\} + \mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}') - \overline{Y}(\boldsymbol{z}')| \geq d_h^{\star} - \bar{\epsilon}\eta_N\right\} \\
&\text{(because } \boldsymbol{z}\notin\mathcal{T}_1 \text{ and } \boldsymbol{z}'\in\mathcal{T}_1\text{)} \\
\leq&4\left\{\frac{\sqrt{\bar{c}\Delta|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0 Q}(d_h^{\star} - \bar{\epsilon}\eta_N)}\cdot\exp\left(-\frac{N_0 Q(d_h^{\star} - \bar{\epsilon}\eta_N)^2}{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)\right. \\
&\left.+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned}$$

(This is deduced analogously to the proof in the previous part)

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

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

Hence, for $N > N(\delta_1, \delta_2)$, we have

$$\begin{aligned}
&\sum_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1} \mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}')|/\eta_N \le 2\bar{\epsilon}\right\}\\
\le& 4|\mathcal{T}_1||\mathcal{T}'|\left\{\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{\pi N_0 Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0 Q d_h^{\star 2}}{8\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned}$$

Combined with (3.55), we have:

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{r}}^{\star}(\widetilde{\boldsymbol{z}})|/\eta_N \ge 2\bar{\epsilon}\right\}\\
\ge&\mathbb{P}\{\widetilde{m}\in\mathcal{T}_1\}-\underbrace{4|\mathcal{T}_1||\mathcal{T}'|\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{\pi N_0 Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0 Q d_h^{\star 2}}{8\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)}_{\text{Term I}}\\
&\underbrace{-4|\mathcal{T}_1||\mathcal{T}'|2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}}_{\text{Term II}}. \qquad (3.56)
\end{aligned}$$

Analogous to the discussion in the previous part, when N is sufficiently large, we can show

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{r}}^{\star}(\widetilde{\boldsymbol{z}})|/\eta_N \ge 2\bar{\epsilon}\right\}\\
\ge&\mathbb{P}\{\widetilde{m}\in\mathcal{T}_1\}-C|\mathcal{T}_1||\mathcal{T}'|\left\{\sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_2}}}\exp\left\{-\frac{C'N^{1+2\delta_2}}{\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right\}+\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned}$$

Similarly we can show for any $z\in\mathcal{T}_1$ and $\underline{\epsilon}>0$,

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_1}|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{r}}^{\star}(\widetilde{\boldsymbol{z}})|/\eta_N \le 2\underline{\epsilon}\right\}\\
&\ge\mathbb{P}\{\widetilde{\boldsymbol{z}}\in\mathcal{T}_1\}-\sum_{\boldsymbol{z}\neq\boldsymbol{z}'\in\mathcal{T}_1}\mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}')|/\eta_N > 2\underline{\epsilon}\right\}.
\end{aligned}$$

Then we have for $\boldsymbol{z}\neq\boldsymbol{z}'\in\mathcal{T}_1$,

$$\begin{aligned}
&\mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}')|/\eta_N > 2\underline{\epsilon}\right\}\\
&\le\mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|\ge\underline{\epsilon}\eta_N-d_h\right\}+\mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}')|\ge\underline{\epsilon}\eta_N-d_h\right\}\\
&\le 4\Bigg\{\frac{\sqrt{\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0 Q}(\underline{\epsilon}\eta_N-d_h)}\cdot\exp\left(-\frac{N_0 Q(\underline{\epsilon}\eta_N-d_h)^2}{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)\\
&+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\Bigg\}.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\left\{|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}')|/\eta_N > 2\underline{\epsilon}\right\} \\
\leq & 4\left\{\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{\pi N_0 Q(\underline{\epsilon}\eta_N)}} \cdot \exp\left(-\frac{N_0 Q(\underline{\epsilon}\eta_N)^2}{8\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right) + 2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned}$$

Hence we have:

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_1}|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \widehat{Y}_{\mathrm{r}}^{\star}(\widetilde{\boldsymbol{z}})|/\eta_N \leq 2\underline{\epsilon}\right\} \\
\geq & \mathbb{P}\left\{\widetilde{\boldsymbol{z}} \in \mathcal{T}_1\right\} - \underbrace{4|\mathcal{T}_1||\mathcal{T}'|\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{\pi N_0 Q(\underline{\epsilon}\eta_N)}} \cdot \exp\left(-\frac{N_0 Q(\underline{\epsilon}\eta_N)^{\star 2}}{8\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)}_{\text{Term I}} \\
& - \underbrace{4|\mathcal{T}_1||\mathcal{T}'|2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}}_{\text{Term II}}.
\end{aligned}$$

Again, by the conditions, we can show

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_1}|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \widehat{Y}_{\mathrm{r}}^{\star}(\widetilde{\boldsymbol{z}})|/\eta_N \leq 2\underline{\epsilon}\right\} \\
\geq & \mathbb{P}\left\{\widetilde{\boldsymbol{z}} \in \mathcal{T}_1\right\} - C|\mathcal{T}_1||\mathcal{T}'|\left\{\sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_2}}}\exp\left\{-\frac{C'N^{1+2\delta_2}}{\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right\} + \sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned}$$

Specifying the random indices. Introduce the following random indices:

$$\widetilde{\boldsymbol{z}}_h = \arg\max_{\boldsymbol{z}\in\mathcal{T}'} \widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}).$$

Applying (3.54) we know that

$$\begin{aligned}
&\mathbb{P}\{\widetilde{\boldsymbol{z}}_h \in \mathcal{T}_1\} \\
\geq & 1 - C|\mathcal{T}'||\mathcal{T}_1|\left\{\sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_2}}}\exp\left(-\frac{C'N^{1+2\delta_2}}{\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right) + \sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned}$$

Aggregating parts. Aggregating all the results above, we can show that, when N is large enough, i.e., $N > n(\delta_1, \delta_2, \delta_3)$,

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_1}|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \widehat{Y}_{\mathrm{r}}^{\star}(\widetilde{\boldsymbol{z}})| \leq \underline{\epsilon}\eta_N, \min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z}) - \widehat{Y}_{\mathrm{r}}^{\star}(\widetilde{\boldsymbol{z}})| \geq \bar{\epsilon}\eta_N\right\} \\
\geq & 1 - C|\mathcal{T}'||\mathcal{T}_1|\left\{\sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_2}}}\exp\left(-\frac{C'N^{1+2\delta_2}}{\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right) + \sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned} \tag{3.57}$$

Bounding the factor level combination selection probability. From the formulated procedure, we have

$$
\begin{aligned}
&\mathbb{P}\left\{\widehat{\mathcal{T}}_1 = \mathcal{T}_1\right\} \\
=&\mathbb{P}\Bigg\{|\widehat{Y}_{\mathrm{R}}(\boldsymbol{z}) - \max_{\boldsymbol{z}\in\mathcal{T}'}\widehat{Y}_{\mathrm{R}}(\boldsymbol{z})| \le \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\in\mathcal{T}_1; \\
&\quad |\widehat{Y}_{\mathrm{R}}(\boldsymbol{z}) - \max_{\boldsymbol{z}\in\mathcal{T}'}\widehat{Y}_{\mathrm{R}}(\boldsymbol{z})| > \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\notin\mathcal{T}_1\Bigg\} \\
\ge&\mathbb{P}\Bigg\{|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \max_{\boldsymbol{z}\in\mathcal{T}'}\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z})| \le \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\in\mathcal{T}_1; \\
&\quad |\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \max_{\boldsymbol{z}\in\mathcal{T}'}\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z})| > \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\notin\mathcal{T}_1\Bigg\} - \mathbb{P}\{\widehat{\mathbb{M}}\neq\mathbb{M}^\star\} \\
=&\mathbb{P}\Bigg\{|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| \le \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\in\mathcal{T}_1; \\
&\quad |\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| > \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\notin\mathcal{T}_1\Bigg\} - \mathbb{P}\{\widehat{\mathbb{M}}\neq\mathbb{M}^\star\}
\end{aligned}
$$

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

$$
\begin{aligned}
\ge&\mathbb{P}\Bigg\{|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| \le \underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\in\mathcal{T}_1; \\
&\quad |\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| > \overline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\notin\mathcal{T}_1\Bigg\} - \mathbb{P}\{\widehat{\mathbb{M}}\neq\mathbb{M}^\star\}
\end{aligned}
$$

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

$$
\begin{aligned}
=&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| \le \underline{\epsilon}\eta_N; \min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| > \overline{\epsilon}\eta_N\right\} \\
&- \mathbb{P}\{\widehat{\mathbb{M}}\neq\mathbb{M}^\star\} \\
\ge&1 - \sum_{h=1}^{H_0}\left(1 - \mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| \le \underline{\epsilon}\eta_N; \min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| > \overline{\epsilon}\eta_N\right\}\right) \\
&- \mathbb{P}\{\widehat{\mathbb{M}}\neq\mathbb{M}^\star\} \\
\ge&1 - \mathbb{P}\{\widehat{\mathbb{M}}\neq\mathbb{M}^\star\} \\
&-C|\mathcal{T}'||\mathcal{T}_1|\left\{\sqrt{\frac{\overline{c}\overline{s}|\mathbb{M}^\star|}{N^{1+2\delta_2}}}\exp\left(-\frac{C'N^{1+2\delta_2}}{\overline{c}\overline{s}|\mathbb{M}^\star|}\right) + \sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^\star|}{N_0Q}}\right\}.
\end{aligned}
$$

□

Lemma 5 suggests that, under the conditions assumed in Theorem 5, we select the first tie set

consistently as $N \to \infty$. Now Theorem 5 is a direct result of Lemma 2 and Lemma 3.

D More discussions

D.1 Statement and proof of Lemma 6

Lemma 6. *$\widehat{Y}_{\mathrm{R}}$ from (4.14) can be expressed as:*

$$\widehat{Y}_{\mathrm{R}} = Q^{-1} G(\cdot, \widehat{\mathbb{M}}) G(\cdot, \widehat{\mathbb{M}})^{\top} \widehat{Y}. \tag{4.58}$$

If $\widehat{\mathbb{M}} = \mathbb{M}^{\star}$, $\mathbb{E}\left\{\widehat{Y}_{\mathrm{R}}\right\} = \overline{Y}$.

Proof of Lemma 6. Due to the orthogonality of G, we have the following decomposition:

$$\widehat{Y} = Q^{-1} G(\cdot, \widehat{\mathbb{M}}) G(\cdot, \widehat{\mathbb{M}})^{\top} \widehat{Y} + Q^{-1} G(\cdot, \widehat{\mathbb{M}}^{c}) G(\cdot, \widehat{\mathbb{M}}^{c})^{\top} \widehat{Y}.$$

By the constraint in (4.14), we have

$$\|\widehat{Y} - \mu\|^2 = \|Q^{-1} G(\cdot, \widehat{\mathbb{M}}^{c}) G(\cdot, \widehat{\mathbb{M}}^{c})^{\top} \widehat{Y}\|^2 + \|Q^{-1} G(\cdot, \widehat{\mathbb{M}}) G(\cdot, \widehat{\mathbb{M}})^{\top} \widehat{Y} - \mu\|^2,$$

which is minimized at

$$\widehat{\mu} = \widehat{Y}_{\mathrm{R}} = Q^{-1} G(\cdot, \widehat{\mathbb{M}}) G(\cdot, \widehat{\mathbb{M}})^{\top} \widehat{Y}.$$

Besides, $\widehat{\mu}$ satisfies the constraint in (4.14).

Next we verify $\mathbb{E}\left\{\widehat{Y}_{\mathrm{R}}\right\} = \overline{Y}$ if $\widehat{\mathbb{M}} = \mathbb{M}^{\star}$. Utilizing the orthogonality of G again, we have

$$\overline{Y} = Q^{-1} G(\cdot, \mathbb{M}^{\star}) G(\cdot, \mathbb{M}^{\star})^{\top} \overline{Y} + Q^{-1} G(\cdot, \mathbb{M}^{\star c}) G(\cdot, \mathbb{M}^{\star c})^{\top} \overline{Y}$$

□

D.2 Discussion of several screening methods in finite population factorial experiments

- Bonferroni correction:
- LASSO:

$$\min_{\tau \in \mathbb{R}^H} \sum_{i=1}^{N} (y_i - f_i^{\top} \tau) + \lambda \|\tau\|_1,$$

- AIC:

$$\min_{\tau\in\mathbb{R}^H,\mathbb{M}\subset[K]} \sum_{i=1}^{N}(y_i - f_i^\top \tau_{\mathbb{M}}) + \lambda|\mathbb{M}|,$$

- BIC:

$$\min_{\tau\in\mathbb{R}^H,\mathbb{M}\subset[K]} \sum_{i=1}^{N}(y_i - f_{i,\mathbb{M}}^\top \tau_{\mathbb{M}}) + \lambda|\mathbb{M}|\log(N),$$

D.3 Discussion: more general centering values

D.3.1 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:

$$\left(f_{\delta'_{[K]}}(z^*_{\mathcal{K}})\right)_{\mathcal{K}\subset[K]} = \left(f_{\delta_{[K]}}(z^*_{\mathcal{K}})\right)_{\mathcal{K}\subset[K]} \boldsymbol{P}_{\Delta\delta_{[K]}}. \tag{4.59}$$

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

$$\boldsymbol{P}_{\Delta\delta_{[K]}}(\mathcal{K},\mathcal{K}') = \begin{cases} \prod_{k\in\mathcal{K}'\setminus\mathcal{K}}(\Delta\delta)_k & , \quad \mathcal{K}\subset\mathcal{K}', \\ 0 & , \quad \mathcal{K}\subsetneq\mathcal{K}'. \end{cases} \tag{4.60}$$

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

$$\boldsymbol{Q}_{\Delta\delta_{[K]}}(\mathcal{K},\mathcal{K}') = \begin{cases} (-1)^{|\mathcal{K}'|-|\mathcal{K}|}\prod_{k\in\mathcal{K}'\setminus\mathcal{K}}(\Delta\delta)_k & , \quad \mathcal{K}\subset\mathcal{K}', \\ 0 & , \quad \mathcal{K}\subsetneq\mathcal{K}'. \end{cases} \tag{4.61}$$

We shall give an example of the above matrix in the three-factor case, which appears (incompletely) in the appendix of Zhao and Ding (2021b). 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 $\boldsymbol{D}_h$.

Theorem 8. *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:*

$$(\widetilde{\tau}_{\mathcal{K}})_{\{|\mathcal{K}|\leq m\}} = (\tau_{\mathcal{K}})_{\{|\mathcal{K}|\leq m\}} + \boldsymbol{D}_h \cdot (\tau_{\mathcal{K}})_{\{|\mathcal{K}|>m\}}, \tag{4.62}$$

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

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

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

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

$$\boldsymbol{D}_h(\mathcal{K},\mathcal{K}') = 0. \tag{4.63}$$

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

$$\boldsymbol{D}_h(\mathcal{K},\mathcal{K}') = \sum_{l=0}^{m-k}(-1)^{k'-k+1-l}\binom{k'-k+1}{l}\prod_{t\in\mathcal{K}'\backslash\mathcal{K}}\left(\delta_t-\frac{1}{2}\right). \tag{4.64}$$

Proof. This result can be derived through careful calculation based on the definition of $\boldsymbol{P}$ and $\boldsymbol{Q}$ from (4.60) and (4.61) along with Theorem 8 thus omitted here. □

D.3.2 A sufficient condition for sign consistency in population WLS regression

Definition 1 (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 9.

Theorem 9. *Assume we run weighted least square under the setting depicted in Theorem 8. Define the maximal decaying rate $c_K = \max_{l\in[K]} c_l$. Recall the predefined maximal active interaction number s_K from Definition 1. If we have*

$$s_K c_K \max_{k=1,\ldots,K} |\delta_k - 1/2| < \ln 2, \tag{4.65}$$

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 (4.62) *have same signs on every term.*

Condition (4.65) 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 $\max_{k=1,\ldots,K} |\delta_k - 1/2|$ 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 Zhao and Ding (2021b): when $\delta_k = 1/2$ holds for all $k = 1, \ldots, K$, $\boldsymbol{D}_h = \mathbf{0}$, so that forward selection always works.