Forward screening and post-screening inference in factorial designs

Lei Shi* Jingshen Wang† Peng Ding ‡

Abstract

Ever since the seminal work of F. Yates, factorial designs have been an important experimental tool to simultaneously estimate the treatment effects of multiple factors. In factorial designs, the number of treatment levels may grow exponentially with the number of factors, which motivates the forward screening strategy based on the sparsity, hierarchy, and heredity principles for factorial effects. Although this strategy is intuitive and has been widely used in practice, its rigorous statistical theory has not been formally established. To fill this gap, we establish a design-based theory for forward factor screening in factorial designs based on the potential outcome framework. We not only prove its consistency property but also discuss statistical inference after factor screening. In particular, with perfect screening, we quantify the advantages of forward screening based on asymptotic efficiency gain in estimating factorial effects. With imperfect screening in higher-order interactions, we propose two novel strategies and investigate their impact on subsequent inference. Our formulation differs from the existing literature on variable selection and post-selection inference because our theory is based solely on the physical randomization of the factorial design and does not rely on a correctly-specified outcome model.

Keywords: Causal inference; Design-based inference; Forward selection; Post-selection inference.

*Division of Biostatistics, University of California, Berkeley. leishi@berkeley.edu

†Division of Biostatistics, University of California Berkeley. jingshenwang@berkeley.edu. Corresponding author.

‡Peng Ding, Department of Statistics, University of California, Berkeley. pengdingpku@berkeley.edu

1 Introduction

1.1 Factorial experiments: opportunities and challenges

Ever since the seminal work of Yates (1937), factorial designs have been widely used in many fields, including agricultural, industrial, and biomedical sciences (e.g., Box et al., 2005; Wu and Hamada, 2011; Gerber and Green, 2012). For example, in social science, one government funded research by Zhang (2022) studied the social construction of hate crime in the U.S. using factorial experiments based on three factors: race, sexual orientation, and religious affiliation. As another example, in ecology, Rillig et al. (2019) studied multiple global change factors in driving soil functions and microbial biodiversity with factorial designs involving up to ten factors, such as XXXXX. Factorial experiments are popular partially because they can simultaneously accommodate multiple factors and offer opportunities to estimate not only the main causal effects of factors but also their interactions.

We focus on the 2^K factorial design in which K binary factors are randomly assigned to N experimental units. With a small K, we can simultaneously estimate the $2^K - 1$ main effects and interactions. Nevertheless, when K is large, the number of factorial effects grows exponentially with K. This motivates us to conduct factor screening based on sparsity, hierarchy, and heredity principles for factorial effects. More precisely, Wu and Hamada (2011) summarized these three principles as below:

(a) (sparsity) The number of important factorial effects is small.

(b) (hierarchy) Lower-order effects are more important than higher-order effects, and effects of the same order are equally important.

(c) (heredity) Higher-order effects are significant only if their corresponding lower-order effects are significant.

The sparsity principle motivates conducting factor screening in factorial designs. The hierarchy principle motivates the forward screening strategy that starts from lower-order effects and then moves on to higher-order effects. The heredity principle motivates using structural restrictions on higher-order effects based on the selected lower-order effects. Due to its simplicity and computational efficiency, while the forward screening strategy has been widely used in data analysis (Wu and Hamada, 2011; Espinosa et al., 2016), its design-based theory under the potential outcome framework has not been formally established. Moreover, it is often challenging to understand the

impact of factor screening on the subsequent statistical inference. The overarching goal of this manuscript is to fill these gaps.

1.2 Our contributions and literature review

We summarize our contribution from three perspectives:

First, our study adds to the growing literature of factorial designs with growing number of factors under the potential outcome framework (Dasgupta et al., 2015; Branson et al., 2016; Lu, 2016b; Espinosa et al., 2016; Egami and Imai, 2019; Blackwell and Pashley, 2021; Zhao and Ding, 2021; Pashley and Bind, 2023; Wu et al., 2022). To deal with a large number of factors, Espinosa et al. (2016) and Egami and Imai (2019) informally used factor screening without studying its statistical properties, whereas Zhao and Ding (2021) discussed parsimonious model specifications that are chosen a priori and independent of data. The rigorous theory for factor screening is generally missing in this literature, let alone the theory for statistical inference after factor screening. At a high level, our contributions fill the gaps.

Second, we formalize forward factor screening and establish its consistency under the design-based framework under few outcome modeling assumptions; see Section 3. Factor screening in factorial design sounds like a familiar statistical task if we formulate it as a variable selection problem in a linear model. Thus, forward screening is reminiscent of the vast literature on forward selection. Wang (2009) and Wieczorek and Lei (2022) proved the consistency of forward selection for the main effects in a linear model, whereas Hao and Zhang (2014) and Hao et al. (2018) moved further to allow for second-order interactions. Other researchers proposed various penalized regressions to encode the sparsity, hierarchy, and heredity principles (e.g., Yuan et al., 2007; Zhao et al., 2009; Bickel et al., 2010; Bien et al., 2013; Lim and Hastie, 2015; Haris et al., 2016), without formally studying the statistical properties of the selected model. Our design-based framework departs from the literature without assuming a correctly-specified linear outcome model. This framework is classic in experimental design and causal inference with randomness coming solely from the design of experiments rather than the error terms in a linear model (Neyman, 1923/1990; Kempthorne, 1952; Freedman, 2008; Lin, 2013; Dasgupta et al., 2015). This framework invokes fewer outcome modeling assumptions but consequently imposes technical challenges for developing the theory. Bloniarz et al. (2016) discussed the design-based theory for covariate selection in treatment-control experiments, but the corresponding theory for factorial designs is largely unexplored.

Third, we discuss statistical inference after forward factor screening with (Sections 4 and 6) or without perfect screening (Section 5). On the one hand, we prove the screening consistency of the

forward screening procedure, which ensures that the selected factorial effects are the true, non-zero ones. With this perfect screening property, we can then proceed as if the selected working model is the true model. This allows us to ignore the impact of forward screening on the subsequent inference, which is similar to the proposal of Zhao et al. (2021) for statistical inference after Lasso Tibshirani (1996). In particular, we quantify the advantages of conducting forward screening based on the asymptotic efficiency gain for estimating factorial effects. As an application under perfect screening, we discuss statistical inference for the mean outcome under the best factorial combination (Andrews et al., 2019; Guo et al., 2021; Wei et al., 2022). On the other hand, we acknowledge that perfect screening can be too much to hope for in practice as it requires strong regularity conditions on factorial effects. As a remedy, we propose two strategies to deal with imperfect screening in higher-order interactions, and we study their impacts on post-screening inference. A key motivation for our strategies is to ensure that the parameters of interest after forward factorial screening are not data-dependent, avoiding philosophical debates in the current literature of post-selection inference (Fithian et al., 2014; Kuchibhotla et al., 2022).

1.3 Notation

We will use the following notation throughout. 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 matrix V, define $\varrho_{\max}(V)$ and $\varrho_{\min}(V)$ as the largest and smallest eigen-values, respectively, and define $\kappa(V) = \varrho_{\max}(V)/\varrho_{\min}(V)$ as its condition number. For two positive semi-definite matrices V_1 and V_2, we write $V_1 \preccurlyeq V_2$ or $V_2 \succcurlyeq V_1$ if $V_2 - V_1$ is positive semi-definite.

We will use different levels of sets. For an integer K, let $[K] = \{1, \ldots, K\}$. We use $\mathcal{K}$ in calligraphic to denote a subset of $[K]$. Let $\mathbb{K} = \{\mathcal{K} \mid \mathcal{K} \subset [K]\}$ denote the power set of $[K]$. We also use blackboard bold font to denote subsets of $\mathbb{K}$. For example, $\mathbb{M} \subset \mathbb{K}$ denotes that $\mathbb{M}$ is a subset of $\mathbb{K}$.

We will use $A_i \sim B_i$ to denote the least-squares fit of A_i's on B_i's, which is purely a numerical procedure without assuming a linear model. Let $\xrightarrow{\mathbb{P}}$ denote convergence in probability and $\rightsquigarrow$ denote convergence in distribution.

2 Setup of factorial designs

This section introduces the key mathematical components of factorial experiments. Section 2.1 introduces the notation of potential outcomes and the definitions of the factorial effects. Section 2.2 introduces the treatment assignment mechanism, the observed data, and regression analysis of the data. Section 2.3 uses a concrete example of a 2^3 factorial experiment to illustrate the key concepts.

2.1 Potential outcomes and factorial effects

We first introduce the general framework of a 2^K factorial design, with $K \geq 2$ being an integer. This design has K binary factors, and factor k can take value $z_k \in \{0, 1\}$ for $k = 1, \ldots, K$. Let $\boldsymbol{z} = (z_1, \ldots, z_K)$ denote the treatment combining all K factors. The K factors in total define $Q = 2^K$ treatment combinations, collected in the set below:

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

We follow the potential outcome notation of Dasgupta et al. (2015) for 2^K factorial designs. Unit i has potential outcome $Y_i(\boldsymbol{z})$ under each treatment level $\boldsymbol{z}$. Corresponding to the $Q = 2^K$ treatment levels, each unit i has Q potential outcomes, vectorized as $\boldsymbol{Y}_i = \{Y_i(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}}$ using the lexicographic order. Over units $i = 1, \ldots, N$, the potential outcomes have finite-population mean vector $\overline{Y} = (\overline{Y}(\boldsymbol{z}))_{\boldsymbol{z}\in\mathcal{T}}$ and covariance matrix $S = (S(\boldsymbol{z}, \boldsymbol{z}'))_{\boldsymbol{z},\boldsymbol{z}'\in\mathcal{T}}$, with elements defined as follows:

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

We then use the potential outcomes to define factorial effects. For a subset $\mathcal{K} \subset [K]$ of the K factors, we introduce the following "contrast vector" notation to facilitate the presentation. To start with, we define the main causal effect for factor k. 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 the 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 interactions of multiple factors with $|\mathcal{K}| \geq 2$, we define the contrast vector $g_{\mathcal{K}} \in \mathbb{R}^Q$ 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 $g_\varnothing = \mathbf{1}_Q$. Stack the contrast vectors into a $Q \times Q$ matrix

$$G = (g_\varnothing, g_{\{1\}}, \ldots, g_{\{K\}}, g_{\{1,2\}}, \ldots, g_{\{K-1,K\}}, \ldots, g_{[K]}),$$

which has orthogonal columns with $G^\top G = Q \cdot I_Q$. We refer to G as the contrast matrix.

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

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

For convenience in description, we use $\tau_\varnothing = Q^{-1} g_\varnothing^\top \overline{Y}$ to denote the average of potential outcomes. We call the effect $\tau_\mathcal{K}$ a *parent* of $\tau_{\mathcal{K}'}$ if $\mathcal{K} \subset \mathcal{K}'$ and $|\mathcal{K}| = |\mathcal{K}'| - 1$. More compactly, we summarize the entire collection of causal parameters in factorial experiments as

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

2.2 Treatment assignment, observed data, and regression analysis

Under the design-based framework, the treatment assignment mechanism characterizes the completely randomized factorial design. In other words, the experimenter randomly assigns $N(\boldsymbol{z})$ units to treatment level $\boldsymbol{z} \in \mathcal{T}$, with $\sum_{\boldsymbol{z} \in \mathcal{T}} N(\boldsymbol{z}) = N$. Assume $N(\boldsymbol{z}) \geq 2$ to allow for variance estimation within each treatment level. Let $Z_i \in \mathcal{T}$ denote the treatment level for unit i. The treatment vector $(Z_1, \ldots, Z_N)$ is a random permutation of a vector with prespecified number $N(\boldsymbol{z})$ of the corresponding treatment level $\boldsymbol{z}$, for $\boldsymbol{z} \in \mathcal{T}$.

For each unit i, the treatment level Z_i only reveals one potential outcome. We use $Y_i = Y_i(Z_i) = \sum_{\boldsymbol{z} \in \mathcal{T}} Y_i(\boldsymbol{z}) \mathbf{1}\{Z_i = \boldsymbol{z}\}$ to denote the observed outcome. We also use $N_i = N(Z_i)$ to denote the number of units for the treatment group in which unit i is assigned to. The central task of causal inference in factorial designs is to use the observed data $(Z_i, Y_i)_{i=1}^N$ to estimate factorial effects. Define

$$\widehat{Y}(\boldsymbol{z}) = N(\boldsymbol{z})^{-1} \sum_{i=1}^N \mathbf{1}\{Z_i = \boldsymbol{z}\} Y_i, \quad \widehat{S}(\boldsymbol{z}, \boldsymbol{z}) = \{N(\boldsymbol{z}) - 1\}^{-1} \sum_{i=1}^N \mathbf{1}\{Z_i = \boldsymbol{z}\} (Y_i - \widehat{Y}(\boldsymbol{z}))^2$$

as the sample mean and variance of the observed outcomes under treatment $\boldsymbol{z}$. Vectorize the sample means as $\widehat{Y} = (\widehat{Y}(\boldsymbol{z}))_{\boldsymbol{z} \in \mathcal{T}}$, which has mean $\overline{Y}$ and covariance matrix $V_{\widehat{Y}} = D_{\widehat{Y}} - N^{-1} S$ (Li and Ding, 2017), where

$$D_{\widehat{Y}} = \text{Diag}\left\{N(\boldsymbol{z})^{-1} S(\boldsymbol{z}, \boldsymbol{z})\right\}_{\boldsymbol{z} \in \mathcal{T}}.$$

An unbiased estimator for $D_{\widehat{Y}}$ is

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

whereas S does not have an unbiased sample analogue because the potential outcomes across treatment levels are never jointly observed for the same units. Therefore, $\widehat{V}_{\widehat{Y}}$ is a conservative estimator of the covariance matrix in the sense that $\mathbb{E}\{\widehat{V}_{\widehat{Y}}\} = D_{\widehat{Y}} \succcurlyeq V_{\widehat{Y}}$.

A dominant approach to estimate factorial effects from factorial designs is through estimating least-squares coefficients based on appropriate model specifications. Let g_i denote the row vector in the contrast matrix G corresponding to unit i's treatment level Z_i, that is, $g_i = \{g_{\mathcal{K}}(Z_i)\}_{\mathcal{K}\subset[K]} \in \mathbb{R}^Q$ with $g_{\mathcal{K}}(\boldsymbol{z})$ defined in (2.2). For a set of target effects $\{\tau_{\mathcal{K}}\}_{\mathcal{K}\in\mathbb{M}}$ indexed by $\mathbb{M}$, we can run weighted least squares (WLS) to obtain unbiased estimates:

$$\widehat{\tau} = \arg\min_{\tau} \sum_{i=1}^{N} w_i(Y_i - g_i^\top \tau)^2 \text{ with } w_i = 1/N_i.$$

With a small K, we can simply fit the *saturated regression* by regressing the observed outcome Y_i on the regressor g_i. The saturated regression involves $Q = 2^K$ coefficients without any restrictions on the targeted factorial effects.

In contrast, an *unsaturated regression* involves fewer coefficients by regressing the observed outcome Y_i on $g_{i,\mathbb{M}}$, a subvector of g_i, where $\mathbb{M} \subset \mathbb{K}$ is a subset of the power set of all factors. That is,

$$\widehat{\tau} = \arg\min_{\tau} \sum_{i=1}^{N} w_i(Y_i - g_{i,\mathbb{M}}^\top \tau)^2 \text{ with } w_i = 1/N_i. \tag{2.3}$$

For the convenience of description, we will call $\mathbb{M}$ a *working model*. We use a working model to generate estimates based on least squares without assuming its correctness. When $\mathbb{M} = \mathbb{K}$, (2.3) incorporates the saturated regression 2.2. Based on the unsaturated regression with working model $\mathbb{M}$, let

$$\widehat{\tau}(\mathbb{M}) = \{\widehat{\tau}_{\mathcal{K}}\}_{\mathcal{K}\in\mathbb{M}} \quad \text{and} \quad \tau(\mathbb{M}) = \{\tau_{\mathcal{K}}\}_{\mathcal{K}\in\mathbb{M}}$$

denote the vectors of estimated and true coefficients, respectively. Zhao and Ding (2021) showed that if we run unsaturated regressions with weights $1/N_i$ for unit i, then the obtained estimated coefficients are unbiased for the true factorial effects within the working model $\mathbb{M}$. More precisely, $\widehat{\tau}(\mathbb{M}) = Q^{-1}G(\cdot,\mathbb{M})\widehat{Y}$, where $G(\cdot,\mathbb{M})$ to denote the columns in G indexed by $\mathbb{M}$. Because $\widehat{\tau}(\mathbb{M})$ is a linear transformation of $\widehat{Y}$, we can use the following estimator for its covariance matrix:

$$\widehat{\boldsymbol{\Sigma}}(\mathbb{M}) = \frac{1}{Q^2}G(\cdot,\mathbb{M})^\top \widehat{V}_{\widehat{Y}} G(\cdot,\mathbb{M}). \tag{2.4}$$

See Lemma S1 in Section A.1 of the supplementary material for more discussions on the above algebraic results for unsaturated regressions.

2.3 An illustrating example of a 2^3 factorial design

We realize that the above notation can be rather abstract. In what follows, we provide an illustrative Example 1 below with $K = 3$ factors.

Example 1 (2^3 factorial design)**.** *Suppose we have three binary factors z_1, z_2, and z_3. These three factors generate 8 treatment combinations, indexed by a triplet $(z_1z_2z_3)$ with $z_1, z_2, z_3 \in \{0, 1\}$, in the set*

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

Each unit i has a potential outcome vector $\boldsymbol{Y}_i = \{Y_i(z_1z_2z_3)\}^\top_{z_1,z_2,z_3=0,1}$. The vector of causal parameters in this 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 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 regressor g_i corresponding to the treatment level Z_i equals

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

For instance, when $Z_i = (101)$, the regressor g_i corresponds to the row (101) of the contrast matrix G. Then, a saturated regression is to regress Y_i on g_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 weighted least squares* $Y_i \sim t_{i,\mathbb{M}}$*, where*

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

and the weight for unit i *equals* $1/N_i = 1/N(Z_i)$*.*

3 Forward screening in factorial experiments

In factorial designs with small K, we can simply run the saturated regression to estimate all factorial effects simultaneously. However, when K is large, saturated regression can be computationally unwieldy and scientifically unreasonable by delivering potentially noisy estimates of all higher-order interactions. As a remedy, forward screening is a popular strategy frequently adopted in practice to analyze data collected from factorial experiments, due to its clear benefits in screening out a large number of zero nuisance factorial effects. In this section, we formalize forward screening as a principled procedure to carefully decide an unsaturated working model $\widehat{\mathbb{M}}$. We first present a formal version of forward screening and then demonstrate its consistency property.

3.1 A formal forward screening procedure

In this subsection, we introduce a principled 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. More concretely, the algorithm starts by performing factor screening over lower-order effects, then move forward to select the significant higher order effects following the heredity principle. Algorithm 1 summarizes the forward screening procedure.

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

First, Algorithm 1 obeys the 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.

Algorithm 1: Forward factorial screening

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

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

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

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

3 Update the 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, and renew the screened working model as $\widehat{\mathbb{M}}'$.

5 Run the unsaturated regression with the working model $\widehat{\mathbb{M}}'$:

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

6 Obtain coefficients $\widehat{\tau}(\widehat{\mathbb{M}}')$ and robust covariance estimation $\widehat{\mathbf{\Sigma}}(\widehat{\mathbb{M}}')$ defined in (2.4).

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-tests using the above $\widehat{\tau}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ and $\widehat{\sigma}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ under the 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** $\widehat{\mathbb{M}}$

Second, Algorithm 1 operates under the sparsity principle as it removes potentially unimportant effects using marginal t-tests with the Bonferroni correction (Step 8). This step induces a sparse working model and helps us to identify essential factorial effects. The sparsity-inducing step can incorporate many popular selection frameworks, such as marginal t-tests, Lasso (Tibshirani, 1996), sure independence screening (Fan and Lv, 2008), etc. For simplicity, we present Algorithm 1 with marginal t-tests and relegate more general discussions to Section B of the supplementary material.

Third, Algorithm 1 incorporates the heredity principle as it screens out the interaction effects (Wu and Hamada, 2011; Hao and Zhang, 2014; Lim and Hastie, 2015) 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).

Lastly, we note that our forward screening procedure enhances the interpretability of the selected working model by iterating between the "Sparsity-screening" step (called the 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 (called the 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,

$$\widehat{\mathbb{M}}_1 \xrightarrow{\mathtt{H}} \widehat{\mathbb{M}}_{2,+} \xrightarrow{\widehat{\mathtt{S}}} \widehat{\mathbb{M}}_2 \cdots \xrightarrow{\widehat{\mathtt{S}}} \widehat{\mathbb{M}}_{d-1} \xrightarrow{\mathtt{H}} \widehat{\mathbb{M}}_{d,+} \xrightarrow{\widehat{\mathtt{S}}} \widehat{\mathbb{M}}_d \to \cdots \xrightarrow{\widehat{\mathtt{S}}} \widehat{\mathbb{M}}_D. \tag{3.5}$$

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

3.2 Consistency of forward screening

We are now ready to analyze the screening consistency property of Algorithm 1. We shall show that the proposed algorithm selects the targeted working model up to level D with probability tending to one as the sample size goes to infinity. Here, the targeted 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 $\tau_{\mathcal{K}} \neq 0$. Define the full targeted working model up to level D as

$$\mathbb{M}_{1:D}^\star = \bigcup_{d=1}^{D} \mathbb{M}_d^\star.$$

In particular, when $D = K$, we omit the subscript and simply denote $\mathbb{M}^\star = \mathbb{M}_{1:K}^\star$.

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

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

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

Condition 1 allows for diverging Q and bounded $N(\boldsymbol{z})$'s across all treatment levels (Shi and Ding, 2022). It generalizes the classical assumption in the fixed Q regime where Q is fixed, and each treatment arm contains a sufficiently large number of replications (Li and Ding, 2017).

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

Condition 2 (Order of parameters)**.** *The true factorial effects $\tau_{\mathcal{K}}$'s and tuning parameters α_d's have the following orders:*

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

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

(iii) *Size of the targeted working model:* $\sum_{d=1}^{D} |\mathbb{M}_d^\star| = \Theta(N^{\delta''})$ *for some* $0 \leq \delta'' < 1/3$.

Condition 2(i) specifies the allowable order of the true factorial effects. If this condition fails, the effect size is of the same order as the statistical error and thus is too small to be detected by marginal t-test. Similar conditions are also adopted in model selection literature, including Zhao and Yu (2006) and Wieczorek and Lei (2022). Condition 2(ii) requires the tuning parameter α_d to converge to zero, which ensures that there is no Type I error in our procedure as N goes to infinity and hence guarantees the selection consistency. Wasserman and Roeder (2009, Theorems 4.1 and 4.2) assumed similar conditions in high-dimensional model selection settings for linear models. Condition (iii) restricts the size of the targeted working model. The rate is due to our technical analysis. Similar conditions also appeared in Zhao and Yu (2006), Wieczorek and Lei (2022) and Wasserman and Roeder (2009).

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

Condition 3 (Regularity conditions on the potential outcomes)**.** *The potential outcomes satisfy the following conditions:*

(i) Nondegenerate correlation matrix. *Let* $V^\star$ *be the correlation matrix of* $\widehat{Y}$. *There exists* $\sigma > 0$ *such that the condition number of* $V^\star$ *is smaller than or equal to* σ^2.

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

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

(iii) Bounded standardization scales. *There exists a constant* $M > 0$ *such that* $M_N \leq M$ *where*

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

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, which is required by Shi and Ding (2022) to prove the Berry–Esseen bound based on Stein's method.

Lastly, we impose the following structural conditions 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* $\mathcal{K}' \subset \mathcal{K}$ *with* $|\mathcal{K}'| = |\mathcal{K}| - 1$ *such that* $\tau_{\mathcal{K}'} \neq 0$.
- *Strong heredity:* $\tau_{\mathcal{K}} \neq 0$ *only if* $\tau_{\mathcal{K}'} \neq 0$ *for all* $\mathcal{K}' \subset \mathcal{K}$ *with* $|\mathcal{K}'| = |\mathcal{K}| - 1$.

Finally, we present the screening consistency property of Algorithm 1:

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

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

4 Inference under perfect screening

Statistical inference is relatively straightforward under the perfect screening of the factorial effects. If forward screening correctly identifies the true, nonzero factorial effects with probability approaching one, we can proceed as if the selected working model is predetermined. In Section 4.1, we present the point estimators and confidence intervals for general causal parameters. In Section 4.2, we study the advantages of forward screening in terms of asymptotic efficiency in estimating general causal parameters, compared with the corresponding estimators without forward screening. We relegate the extensions to vector parameters to Section A.2 of the supplementary material since it is conceptually straightforward.

4.1 Post-screening inference for general causal parameters

Define a general causal parameter of interest as a weighted combination of average potential outcomes:

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

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). However, we allow $\boldsymbol{f}$ to be different from the contrast vectors $g_{\mathcal{K}}$. For instance, if one wants to focus on the first two arms in factorial experiments and estimate the average treatment effect, we shall choose

$$\boldsymbol{f} = (1, -1, 0, \ldots, 0)^{\top}.$$

In general, researchers may tailor the choice of $\boldsymbol{f}$ to the specific research questions of interest.

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

$$\widehat{\gamma} = \boldsymbol{f}^\top \widehat{Y} = \sum_{\boldsymbol{z}\in\mathcal{T}} \boldsymbol{f}(\boldsymbol{z})\widehat{Y}(\boldsymbol{z}). \tag{4.6}$$

Under regularity conditions in Shi and Ding (2022), the plug-in estimator $\widehat{\gamma}$ satisfies a central limit theorem $(\widehat{\gamma}-\gamma)/v \rightsquigarrow \mathcal{N}(0,1)$ with the variance $v^2 = \boldsymbol{f}^\top V_{\widehat{Y}} \boldsymbol{f}$. When $N(z) \geq 2$, its variance can be estimated by:

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

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 the restricted least squares (RLS)

$$\widehat{Y}_{\mathrm{R}} = \arg\min_{\mu\in\mathbb{R}^Q} \Big\{ \|\widehat{Y}-\mu\|_2^2 : \; G(\cdot,\widehat{\mathbb{M}}^c)^\top \mu = 0 \Big\}, \tag{4.7}$$

which leverages the information that the nuisance effects $G(\cdot,\widehat{\mathbb{M}}^c)^\top \overline{Y}$ are all zero. The $\widehat{Y}_{\mathrm{R}}$ in (4.7) has a closed form solution (see Lemma S6 in the supplementary material):

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

Under perfect screening, $\widehat{Y}_{\mathrm{R}}$ is also a consistent estimator for $\overline{Y}$, so $\widehat{\gamma}_{\mathrm{R}} = \boldsymbol{f}^\top \widehat{Y}_{\mathrm{R}}$ is also consistent for γ. Introduce the following notation

$$\boldsymbol{f}[\mathbb{M}] = Q^{-1} G(\cdot,\mathbb{M}) G(\cdot,\mathbb{M})^\top \boldsymbol{f} \tag{4.8}$$

to simplify $\widehat{\gamma}_{\mathrm{R}}$ and its variance estimator as

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

Construct a Wald-type level-$(1-\alpha)$ confidence interval for γ:

$$\Big[\widehat{\gamma}_{\mathrm{R}} \pm z_{1-\alpha/2} \times \widehat{v}_{\mathrm{R}} \Big], \tag{4.9}$$

where $z_{1-\alpha/2}$ is $(1-\alpha/2)$th quantile of a standard normal distribution. We can also obtain point estimates and confidence intervals handily from WLS regression of Y_i on $g_{i,\widehat{\mathbb{M}}}$ with weights $1/N_i$. See Section A.1 in the supplementary material for more details.

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 $\boldsymbol{f}^\star = \boldsymbol{f}[\mathbb{M}^\star]$. Given $\mathbb{M}^\star$ is the true working model, we have $(\boldsymbol{f}^\star)^\top \overline{Y} = \boldsymbol{f}^\top \overline{Y}$, for all $\boldsymbol{f} \in \mathbb{R}^Q$. This identify holds for the true working model, not a general model, suggested by the following algebraic facts:

$$\begin{aligned}
\boldsymbol{f}^\top \overline{Y} &= \boldsymbol{f}^\top \{Q^{-1} G(\cdot, \mathbb{M}^\star) G(\cdot, \mathbb{M}^\star)^\top + Q^{-1} G(\cdot, \mathbb{M}^{\star c}) G(\cdot, \mathbb{M}^{\star c})^\top\} \overline{Y} \text{ (orthogonality of } G) \\
&= (\boldsymbol{f}^\star)^\top \overline{Y} + G(\cdot, \mathbb{M}^{\star c}) \tau(\mathbb{M}^{\star c}) \text{ (definition of } \boldsymbol{f}^\star \text{ based on (4.8))} \\
&= (\boldsymbol{f}^\star)^\top \overline{Y}. \text{ (using } \tau(\mathbb{M}^{\star c}) = 0)
\end{aligned}$$

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 $\widehat{v}_{\mathrm{R}}^2$)**.** *Let $N \to \infty$. Assume Conditions 1-4. We have*

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

where $v_{\mathrm{R}}^2 = \boldsymbol{f}^{\star\top} V_{\widehat{Y}} \boldsymbol{f}^\star$. Further assume $\|\boldsymbol{f}^\star\|_\infty = O(Q^{-1})$. The variance estimator $\widehat{v}_{\mathrm{R}}^2$ is conservative in the sense that:

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

where $v_{\mathrm{R,lim}}^2 = \boldsymbol{f}^{\star\top} D_{\widehat{Y}} \boldsymbol{f}^\star$ is the limiting value of $\widehat{v}_{\mathrm{R}}^2$.

Theorem 2 above guarantees that the proposed confidence interval in (4.9) for γ attains the nominal coverage probability asymptotically. 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)**.** *Without factor screening, the simple plug-in estimator $\widehat{\gamma}$ in* (4.6) *satisfies a central limit theorem if*

$$N_0^{-1/2} \cdot \frac{\|\boldsymbol{f}\|_\infty}{\|\boldsymbol{f}\|_2} \to 0 \tag{4.10}$$

recalling the definition of N_0 in Condition 1 (Shi and Ding, 2022, Theorem 1). Condition (4.10) *fails when N_0 is small and $\boldsymbol{f}$ is sparse. Besides, it does not incorporate the sparsity information in the structure of factorial effects. With factor screening, however, we can borrow the benefit of a sparse working model and overcome the above drawbacks. Therefore, factor screening broadens the applicability of our proposed estimator $\widehat{\gamma}_{\mathrm{R}}$ by weakening the assumptions for the Wald-type inference.*

To elaborate the benefits of conducting forward factorial screening in terms of asymptotic efficiency, we make a simple comparison of the asymptotic variances of $\widehat{\gamma}$ and $\widehat{\gamma}_{\mathrm{R}}$ in Proposition 1 below. In the most general setup, there is no ordering relationship between v_{R}^2 and v^2. That is, the RLS based estimator may have higher variance than the unrestricted OLS estimator. This is a known fact due to heteroskedasticity and the use of sandwich variance estimators (Meng and Xie, 2014; Zhao and Ding, 2021). Nevertheless, in many interesting scenarios, we can prove an improvement of efficiency by factor screening. Two conditions are summarized in Proposition 1:

Proposition 1 (Asymptotic relative efficiency comparison between $\widehat{\gamma}$ and $\widehat{\gamma}_{\mathrm{R}}$)**.** *Assume that both $\widehat{\gamma}$ and $\widehat{\gamma}_{\mathrm{R}}$ converge to a normal distribution as $N \to \infty$.*

(i) *If the condition number of $V_{\widehat{Y}}$ satisfies $\kappa(V_{\widehat{Y}}) = 1$, then*

$$\frac{v_{\mathrm{R}}^2}{v^2} \leq 1.$$

(ii) *Let $s^\star$ denote the number of nonzero elements in $\boldsymbol{f}$. Then the asymptotic relative efficiency between $\widehat{\gamma}$ and $\widehat{\gamma}_{\mathrm{R}}$ is upper bounded by*

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

Now we add some interpretation for Proposition 1. The condition $\kappa(V_{\widehat{Y}}) = 1$ in Part (i) holds when the variance of $\widehat{Y}(\boldsymbol{z})$ does not change with its treatment group membership $\boldsymbol{z}$. One concrete problem of interest is testing the *sharp null hypothesis* of constant effects in uniform factorial designs (with N_0 replications in each arm), i.e.,

$$\mathrm{H}_{0\mathrm{F}} : Y_i(\boldsymbol{z}) = Y_i \text{ for all } i \in [N] \text{ and } \boldsymbol{z} \in \mathcal{T}.$$

Under $\mathrm{H}_{0\mathrm{F}}$, we have

$$V_{\widehat{Y}} = N_0^{-1}\sigma \cdot I_Q, \text{ where } \sigma = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2 \text{ and } \overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

The proposed RLS-based estimator $\widehat{\gamma}_{\mathrm{R}}$ is in general more efficient than the plug-in estimator $\widehat{\gamma}$. Part (ii) studies a general heteroskedastic setting with sparse weighting vector $\boldsymbol{f}$ and working model size $|\mathbb{M}^\star|$. 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 $\mathcal{T}$. When the variability of such changes is limited in the sense that $\kappa(V_{\widehat{Y}}) < Q/(s^\star|\mathbb{M}^\star|)$, the RLS-based estimator is more efficient than $\widehat{\gamma}$. Moreover, the above result can be extend to compare the length of the confidence intervals as well. The conclusion is similar. See Proposition S1 in the supplementary material for the details.

5 Post-screening inference under imperfect screening

Similar to many other consistency results for variable selection, the perfect screening property can be too much to hope for in practical data analysis in factorial designs. This is because the consistency of forward screening is a theoretical property under regularity conditions. Motivated by the hierarchy principle for factorial effects, the main factorial effects and lower-order factorial effects are more likely to be non-negligible than the higher-order factorial effects. With smaller effect size for the higher-order interactions, the perfect screening property more less likely to hold even asymptotically. Technically, when Condition 2(i) is violated, Algorithm 1 may no longer enjoy the perfect screening property.

Statistical inference without perfect screening is a non-trivial problem in factorial designs. If we do not put any restriction on the imperfect selection, the selected model can be anything, even without a stable limit. Classical strategies for post-selection inference (Kuchibhotla et al., 2022) will encounter various drawbacks in our current setup. For example, data splitting (Wasserman and Roeder, 2009) is a widely used strategy to validate inference after variable selection due to its simplicity. However, it highly relies on the independent sampling assumptions, which is violated under complete randomization. On the other hand, selective inference (Fithian et al., 2014) is another widely studied strategy. As another example, selective inference delivers valid inference for data-dependent parameters. However, it cannot be directly applied to analyze data collected in factorial designs. This is because selective inference strategy tends to rely on specific selection methods and parametric modelling assumptions on the outcome variables.

Rather than directly generalizing classical post selection inference methods to factorial experiments, in this section, we shall discuss two alternative strategies leveraging the special data structures in factorial experiments, along with with their statistical inference results (summarized in Figure 1).

5.1 Two alternative strategies for imperfect screening and statistical inference

The two proposed strategies are built on a belief that perfect screening is more plausible for selecting the main factorial effects and lower-order factorial effects up to level $d^\star$ than the high order effects. We will add more discussion on $d^\star$ after presenting the two strategies.

For Strategy 1, when the higher-order factorial effects are not necessary in the study, we may stop our forward screening procedure in Algorithm 1 at $d = d^*$ (instead of $d = D$). Such a strategy

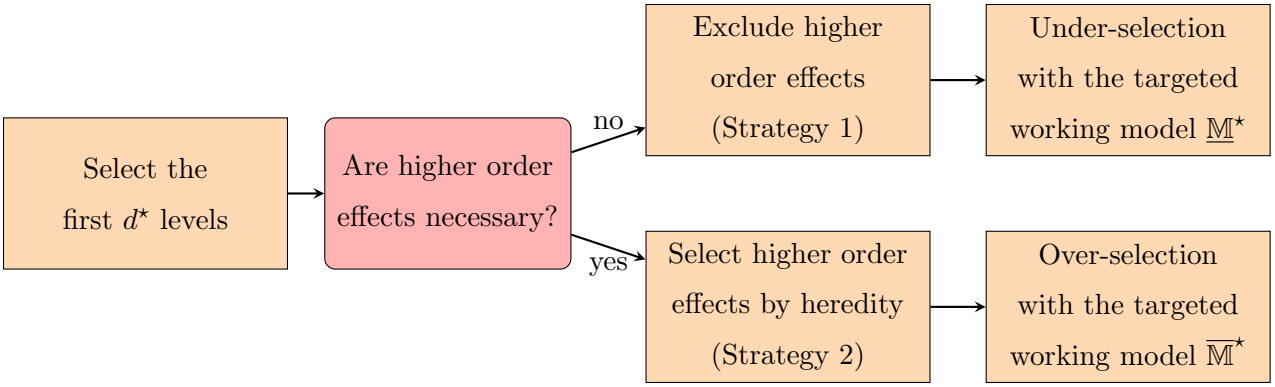

Figure 1: Two strategies for factorial screening: Strategy 1 under selects whereas Strategy 2 over selects.

focus on recovering a targeted working model $\underline{\mathbb{M}}^\star$ up to level $d^\star$, that is,

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

which leads to an under-selected parsimonious working model. We summarize this strategy below.

Strategy 1 (Under selection by excluding high-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 again construct an estimator of $\gamma = \boldsymbol{f}^\top \overline{Y}$ (defined in Section 4.1) based on RLS:

$$\widehat{\gamma}_{\text{RU}} = \boldsymbol{f}[\widehat{\underline{\mathbb{M}}}]^\top \widehat{Y}, \quad \text{and} \quad \widehat{v}_{\text{RU}}^2 = \boldsymbol{f}[\widehat{\underline{\mathbb{M}}}]^\top \widehat{V}_{\widehat{Y}} \boldsymbol{f}[\widehat{\underline{\mathbb{M}}}]. \tag{5.11}$$

For Strategy 2, rather than excluding all higher-order interactions with negligible effects, we may further leverage the 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 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 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*

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

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 RLS:

$$\widehat{\gamma}_{\mathrm{RO}} = \boldsymbol{f}[\widehat{\overline{\mathbb{M}}}]^\top \widehat{Y}, \quad \text{and} \quad \widehat{v}^2_{\mathrm{RO}} = \boldsymbol{f}[\widehat{\overline{\mathbb{M}}}]^\top \widehat{V}_{\widehat{Y}} \boldsymbol{f}[\widehat{\overline{\mathbb{M}}}]. \tag{5.12}$$

In terms of implementation, one can use WLS to conveniently obtain the point estimators in (5.11) and (5.12) and slightly more conservative variance estimators. Due to the orthogonality of the contrast matrix G, perfect screening is not required for computation. See Section A.1 in the supplementary material for more detailed discussions.

In real-world factorial experiments, how should practitioners decide which strategy to work with? This relies on the domain knowledge and the research question of interest. Strategy 1 is more suitable when there are domain-specific messages indicating that higher order interactions are small, or the research question only involves lower order factorial effects. Moreover, Strategy 1 is helpful when the number of active lower order interaction is large and Strategy 2 cannot be applied. Meanwhile, Strategy 2 works better when domain knowledge suggests non-negligible higher order interactions or the research question targets a more general parameter beyond factorial effects themselves. It also works well when the number of active lower order interaction is small, and we can include a small set of high order terms according to the heredity principle.

In the following subsection, we study the statistical properties of $\widehat{\gamma}_{\mathrm{RO}}$ and $\widehat{\gamma}_{\mathrm{RU}}$ and demonstrate the trade-offs between the two strategies for statistical inference from a theoretical perspective.

5.2 Theoretical properties under imperfect screening

Throughout this subsection, we discuss the scenario where perfect screening is hard to achieve. We work under a relaxed condition of Condition 2 defined as follows:

Condition 5 (Order of parameters up to level $d^\star$)**.** *Condition 2 holds with* $D = d^\star$*.*

Condition 5 no longer imposes any restriction on the order of the parameters beyond level $d^\star$. By Theorem 1, Condition 5 guarantees that Algorithm 1 perfectly screens the first $d^\star$ levels of factorial effects in the sense that

$$\mathbb{P}\left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star \text{ for } d = 1, \ldots, d^\star\right\} \to 1.$$

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 consistent 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)**.** *Recall the equation* (4.8) *and define* $\underline{\boldsymbol{f}}^\star = \boldsymbol{f}[\underline{\mathbb{M}}^\star] = Q^{-1}G(\cdot, \underline{\mathbb{M}}^\star)G(\cdot, \underline{\mathbb{M}}^\star)^\top \boldsymbol{f}$*. Assume Conditions 1, 3, 4, 5, and* $\boldsymbol{f}$ *satisfies the following orthogonality condition:*

$$G(\cdot, \mathbb{M}_d^\star)^\top \boldsymbol{f} = 0 \text{ for } d^\star + 1 \le d \le K. \tag{5.13}$$

Let $N \to \infty$*. We have*

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

where $v_{\text{RU}}^2 = \underline{\boldsymbol{f}}^{\star\top} V_{\widehat{Y}} \underline{\boldsymbol{f}}^\star$*. Further assume* $\|\underline{\boldsymbol{f}}^\star\|_\infty = O(Q^{-1})$*. The variance estimator* $\widehat{v}_{\text{RU}}^2$ *is conservative in the sense that:*

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

where $v_{\text{RU,lim}}^2 = \underline{\boldsymbol{f}}^{\star\top} D_{\widehat{Y}} \underline{\boldsymbol{f}}^\star$ *is the limiting value of* $\widehat{v}_{\text{RU}}^2$*.*

Now we add some discussion on Theorem 3. The orthogonality condition presented in (5.13) restricts the weighting vector $\boldsymbol{f}$ to be orthogonal to the higher order contrasts. Intuitively, because the higher-order interactions are excluded from the model, making inference on a weighted combination of those excluded interactions is infeasible. One set of weighting vectors satisfying (5.13) is the contrast vectors of nonzero canonical lower-order interactions, given by $\boldsymbol{f} = G(, \cup_{d=1}^{d^\star} \mathbb{M}_d^\star)$. In large K settings, the lower order interactions can also grow polynomially fast in K and add difficulty for interpretation. As an example, when $K = 10$, for the first two levels of factorial effects without screening, there are a total of more than 50 estimates. It can still greatly benefit the analysis and interpretation to filter out the insignificant ones and obtain a parsimonious, structured working model.

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

Theorem 4 (Guarantee for Strategy 2)**.** *Recall the equation* (4.8) *and define* $\overline{\boldsymbol{f}}^\star = f[\overline{\mathbb{M}}^\star] = Q^{-1}G(\cdot,\overline{\mathbb{M}}^\star)G(\cdot,\overline{\mathbb{M}}^\star)^\top \boldsymbol{f}$*. Assume Conditions 1, 3, 4 and 5. Let* $N \to \infty$*. If* $|\overline{\mathbb{M}}^\star|/N \to 0$*, then*

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

where $v_{\mathrm{RO}}^2 = \overline{\boldsymbol{f}}^{\star\top} V_{\widehat{Y}} \overline{\boldsymbol{f}}^\star$*. Further assume* $\|\overline{\boldsymbol{f}}^\star\|_\infty = O(Q^{-1})$*. The variance estimator* $\widehat{v}_{\mathrm{RO}}^2$ *is conservative in the sense that:*

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

where $v_{\mathrm{RO,lim}}^2 = \overline{\boldsymbol{f}}^{\star\top} D_{\widehat{Y}} \overline{\boldsymbol{f}}^\star$ *is the limiting value of* $\widehat{v}_{\mathrm{RO}}^2$*.*

We comment that there is an additional technical requirement in Theorem 4 for over-selection: we assume $|\overline{\mathbb{M}}^\star|/N \to 0$. This equation mainly serves as a sufficient condition for CLT. The reason is that we need to control the size of the target model $\overline{\mathbb{M}}^\star$ compared to the sample size N in order to infer a general causal parameter.

When analyzing Strategies 1 and 2, Algorithm 1 recovers a targeted 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 not be feasible if there are too many lower order terms which can result in 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 equal-sized designs with* $N(\boldsymbol{z}) = N_0$ *and homoskedasticity with* $S(\boldsymbol{z},\boldsymbol{z}) = S_0$ *for all* $\boldsymbol{z} \in \mathcal{T}$*, 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 show that*

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

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 Application to inference on the best arm in factorial experiments

In the previous sections, we consider the problem of making inference on a single factorial causal effect $\gamma = \boldsymbol{f}^\top \overline{Y}$. As an application of the proposed framework, we study the problem of inference on best causal effects. Section 6.1 introduces the basic problem and an inferential procedure. Section 6.2 presents the theoretical guarantees.

6.1 Best arms, tie set and statistical inference

Suppose we have a set of causal effects Γ defined by pre-specified weighting vectors $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_L$ (L is potentially large), that is

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

We aim to perform statistical inference on their ordered values

$$\gamma_{(1)} \geq \cdots \geq \gamma_{(l_0)}$$

with $l_0 < L$ being a fixed positive integer. As a concrete example, if we choose $\{\boldsymbol{f}_l\}_{l\in[L]} = \{\boldsymbol{e}(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}}$ to be the set of the canonical bases $\{\boldsymbol{e}(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}}$, then our inferential target is the maximal potential outcome means:

$$\overline{Y}_{(1)} = \max_{z\in\mathcal{T}} \overline{Y}(\boldsymbol{z}). \tag{6.14}$$

A more practically appealing consideration in factorial experiments is to incorporate structural constraints into the choices of $\{\boldsymbol{f}_l\}_{l\in[L]}$. For example, when inferring (6.14) in a real-world factorial experiment, it might be infeasible to consider all treatment levels $\mathcal{T}$ due to budget concerns or resource limitation. Instead, we may only be interested in factor combinations $\boldsymbol{z} = (z_1, \ldots, z_K)$ with at most $K_0 (\leq K)$ 1's. In other words, we replace $\mathcal{T}$ with the following $\mathcal{T}'$ in (6.14) and obtain:

$$\mathcal{T}' = \left\{ \boldsymbol{z} = (z_1, \ldots, z_K) \mid \sum_{k=1}^{K} z_k \leq K_0 \right\}, \quad \overline{Y}_{(1)} = \max_{z\in\mathcal{T}'} \overline{Y}(\boldsymbol{z}). \tag{6.15}$$

Such constraints are unique and important to factorial experiments, especially when K is large. By focusing on $\{\boldsymbol{f}\}_{l\in[L]}$ that are most pertinent to factorial designs, the problem allows us to use the available data to decide if the best causal parameter among those practically interesting ones has a non-zero causal effect.

Two challenges exist in delivering valid statistical inference on $\gamma_{(1)}, \ldots, \gamma_{(l_0)}$ in factorial experiments. On the one hand, sample analogs of the order parameters, $(\widehat{\gamma}_{(1)}, \ldots, \widehat{\gamma}_{(l_0)})$, are often biased

estimates of $(\gamma_{(1)}, \ldots, \gamma_{(l_0)})$ due to the well-known winner's curse phenomenon (Andrews et al., 2019; Guo et al., 2021; Wei et al., 2022). On the other hand, although one might argue that existing approaches can be applied to remove the winner's curse bias in $\widehat{\gamma}_{(l)}$, these approaches do not account for the special structural constraint in factorial experiments. Moreover, rigorous statistical guarantees are lacking in the context due to the unique presence of both large L and large Q in factorial designs.

To simultaneously address the above challenges, we propose a procedure that tailors the tie-set identification approach proposed in Claggett et al. (2014) and Wei et al. (2022) to our current problem setup. We focus on making inference on the first ordered value $\gamma_{(1)}$ to simplify discussion, and our approach extends naturally to other ordered values. The proposed procedure is provided in Algorithm 2.

Algorithm 2 consists of three major components. First, we need to 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 based on our previous discussion. Second, we construct $\widehat{\mathcal{L}}_1$ to include the estimates that are close to $\widehat{\gamma}_{(1)}$ (Step 3). Intuitively, these collected estimates are different due to random error. We will show that with proper tuning, this procedure will include all the l for which γ_l are statistically indistinguishable from $\gamma_{(1)}$ with high probability. Third, we construct estimators by averaging over $\widehat{\mathcal{L}}_1$ (Step 4). By averaging the estimates over the selected $\widehat{\mathcal{L}}_1$ we alleviate the impact of randomness and obtain accurate estimates for the maximal effect.

6.2 Theoretical guarantees

In the following, we present theoretical guarantees for Algorithm 2. We introduce the following notation $\mathcal{L}_1$ to include all effects that stay in a local neighbourhood of $\gamma_{(1)}$:

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

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

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

as within- and between-group distances, respectively. We work under the following condition:

Condition 6 (Order of d_h, $d_h^\star$ and η_N)**.** *Assume the within and between group distances satisfy:*

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

Algorithm 2: Inference on best causal effect(s)

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

2 Obtain RLS based estimates: use Equation (4.8) and definition of $\widehat{Y}_{\mathrm{R}}$ (4.7) to compute

$$\boldsymbol{f}_l[\widehat{\mathbb{M}}] = Q^{-1}G(\cdot,\widehat{\mathbb{M}})G(\cdot,\widehat{\mathbb{M}})^\top \boldsymbol{f}_l, \quad \widehat{\gamma}_l = \boldsymbol{f}_l^\top \widehat{Y}_{\mathrm{R}} = \boldsymbol{f}_l[\widehat{\mathbb{M}}]^\top \widehat{Y}, \quad l \in [L].$$

3 Record the set of effects close to $\widehat{\gamma}_{(1)}$:

$$\widehat{\mathcal{L}}_1 = \left\{ l \in [L] \mid |\widehat{\gamma}_l - \widehat{\gamma}_{(1)}| \leq \eta_N \right\}.$$

Here, η_N is a tuning parameter which can be selected using the algorithm provided in Wei et al. (2022), Appendix C.1.

4 Define

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

Generate point estimates and variance estimator for $\gamma_{(1)}$:

$$\widehat{Y}_{(1)} = \frac{1}{|\widehat{\mathcal{L}}_1|} \sum_{l\in\widehat{\mathcal{L}}_1} \widehat{\gamma}_l = \boldsymbol{f}_{(1)}^\top \widehat{Y}, \quad \widehat{v}_{(1)}^2 = \boldsymbol{f}_{(1)}^\top \widehat{V}_Y \boldsymbol{f}_{(1)}.$$

5 **return** $\widehat{\mathcal{L}}_1, \widehat{Y}_{(1)}, \widehat{v}_{(1)}^2$

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

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

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

We establish the following result for the procedure provided in Algorithm 2. Recall δ_2 from Condition 6 and δ'' from Condition 2(iii), which characterizes the magnitude of the within/between group distances and the size of the true working model, respectively.

Theorem 5 (Asymptotic results on the estimated effects using Algorithm 2)**.** *Assume Condition*

1–4 and 6. Let $N \to \infty$. If

$$N^{-(1+2\delta_2-\delta'')} \to 0, \tag{6.16}$$

$$L \cdot |\mathcal{L}_1| \cdot N^{-\frac{1-\delta''}{2}} \to 0, \tag{6.17}$$

then

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

where $v_{(1)}^2 = \boldsymbol{f}_{(1)}^{\star\top} V_Y \boldsymbol{f}_{(1)}^{\star}$. *Moreover,* $\widehat{v}_{(1)}^2$ *is conservative in the sense that*

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

where $v_{(1),\lim}^2 = \boldsymbol{f}_{(1)}^{\star\top} D_{\widehat{Y}} \boldsymbol{f}_{(1)}^{\star}$ *is the limiting value of* $v_{(1)}^2$.

The conditions in Theorem 5 are mild and reveals an interesting trade-off between some mathematical quantities. For the first asymptotic condition (6.16), when the size of the targeted working model is small compared to N, say $\delta'' = 0$ (meaning $|\mathbb{M}^\star|$ does not grow with N), this condition always holds. More generally, (6.16) is easier to satisfy with larger gap size (larger δ_2) and smaller true working model size (smaller δ''). The second condition (6.17) reflects the tradeoff among the total number of interested parameters (L, which is also $|\mathcal{T}'|$), the size of the neighborhood of $\gamma_{(1)}$ ($|\mathcal{L}_1|$), and the size of the true working model (δ''). The smaller these quantities are compared to N, the easier inference will be. Moreover, (6.17) is easily justifiable. Going back to the previous example (6.15), (6.17) translates into

$$\sum_{K=0}^{K_0} \binom{K}{k} \cdot |\mathcal{L}_1| \cdot \left(\frac{|\mathbb{M}^\star|}{N}\right)^{1/2} \to 0. \tag{6.18}$$

(6.18) accommodates a variety of interesting regimes with different specifications of K_0, $|\mathcal{L}_1|$ and $|\mathbb{M}^\star|$. We omit the discussion here.

Theorem 5 also suggests the benefits of factor screening compared to procedures where no screening is involved. Recall our discussion in Remark 1. Without screening, one requires Q to be small compared to N or $\{\boldsymbol{f}_l\}_{l\in[L]}$ are dense, which is violated in large Q setups and many practical scenarios such as (6.14). On the contrary, factor screening can capture the shared information across treatment arms and still deliver valid inference.

As a final comment, the result of our Theorem 5 relies on the perfect screening property (Theorem 1), which are ensured by Conditions 1 - 4. Without perfect screening, there might be additional sources of bias due to the uncertainty induced by the screening step and possible under-selection results. Nevertheless, one can consider applying the over-selection strategy (Strategy 2 in Section 5.1) to facilitate inference on the best factorial effects.

7 Simulation

In this section, we use simulation studies to demonstrate the finite-sample performance of the proposed forward screening framework and the inferential properties of the RLS-based estimator. More concretely, our simulation results verify the following properties of the proposed procedure and estimators:

(G1) The RLS-based estimator $\widehat{\gamma}_{\mathrm{R}}$ demonstrates efficiency gain (in terms of improved power and shortened confidence interval) compared to the simple moment estimator $\widehat{\gamma}$ for general causal parameters defined by sparse weighting vectors.

(G2) The factorial forward screening procedure provided in Algorithm 1 can improve the performance of effect screening compared to naive procedure (i.e., screening without leveraging the heredity principle).

(G1) echoes our discussion on the comparison of CLT conditions and asymptotic variance in Remark 1 and Proposition 1. **(G2)** verifies the results in Theorem 1 and 2 and checks the finite sample behaviors of the proposed procedures. For both goals, we will vary the sample size and effect size to provide a comprehensive understanding of their performance.

7.1 Simulation setup

We set up a 2^8 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 independent potential outcomes 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 $\boldsymbol{z}$. We choose $\mu(\boldsymbol{z})$ 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 of the two-way interactions are zero.
- Higher order interactions: all the higher-order interactions $\tau_{\mathcal{K}}$ are zero if $|\mathcal{K}| \geq 3$.

The above setup of factorial effects guarantees that they are sparse and follow the strong heredity principle. In the provided simulation results, we will vary the number of units in each treatment arm and the size of the nonzero factorial effects. More details can be found in the `R` code attached to the support materials.

7.2 Simulation results supporting (G1)

In this subsection, we evaluate the performance of the RLS-based estimators $(\widehat{\gamma}_{\mathrm{R}}, \widehat{v}_{\mathrm{R}})$ compared to $(\widehat{\gamma}, \widehat{v})$ for testing a causal effect $\gamma_{\text{target}} = \boldsymbol{f}^{\top}\overline{Y}$ specified by a sparse vector: $\boldsymbol{f} = (0, \ldots, 0, 1)^{\top} \in \mathbb{R}^{Q}$. Intuitively, γ_{target} measures the average of potential outcomes in the last level. For each estimator, we report: (i) power for testing $\mathrm{H}_0 : \gamma_{\text{target}} = 0$. (ii) coverage probability of the confidence intervals for γ_{target} at level 0.95. Figure 2 summarizes the results.

Figure 2 shows that the RLS-based estimator $\widehat{\gamma}_{\mathrm{R}}$ has power one and coverage probability 0.95 asymptotically when the effect size or the number of replications is large enough. The simple moment estimator $\widehat{\gamma}$, on the other hand, suffers from deficiency in power in finite sample. While the coverage probability appears high for small N_0 and effect size, it is largely due to an overly conservative confidence interval given the low power. This echoes our conclusion that effect screening can incorporate information across treatment levels and improve the power for testing general causal effects.

7.3 Simulation results for (G2)

In this subsection, we compare the performance of forward and naive screening procedures. We use four candidate effect screening methods:

- *Forward Bonferroni.* Forward screening based on Bonferroni corrected marginal t tests;
- *Forward Lasso.* Forward screening based on Lasso;
- *Naive Bonferroni.* Screening with the full working model based on Bonferroni corrected margin t tests;
- *Naive Lasso.* Screening with the full working model based on Lasso.

For each of the screening methods, we evaluate their performance from several perspectives: (i) perfect screening probability $\mathbb{P}\{\widehat{\mathbb{M}} = \mathbb{M}^{\star}\}$. (ii) power of $\widehat{\gamma}_{\mathrm{R}}$ for testing $\mathrm{H}_0 : \gamma_{\text{target}} = 0$ for the

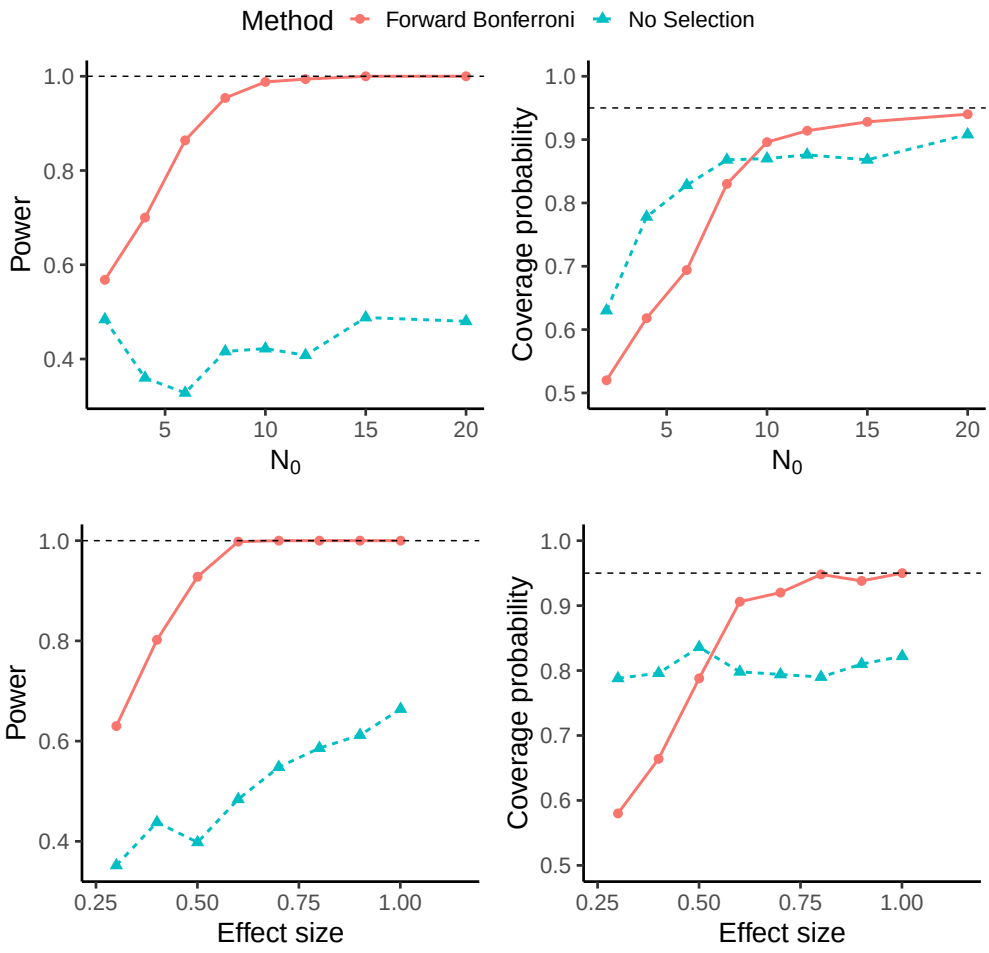

Figure 2: Simulation results on **(G1)**. (i) Top left panel: power curve with varying N_0; (ii) Top right panel: coverage probability with varying N_0; (iii) Bottom left panel: power curve with varying effect size; (iv) Bottom right panel: coverage probability with varying effect size.

same γ_{target} defined in the previous section. (iii) coverage probability of the RLS based confidence interval for γ_{target} at level 0.95. The results are summarized in Figure 3.

From Figure 3, all four effect screening methods lead to perfect selection with high probability as the number of replications N_0 or effect size increases. Nevertheless, with the forward screening procedure, the probability of perfect screening are higher than naive screening procedure. Besides, forward screening keeps the heredity structure and demonstrates higher interpretability than the naive screening methods. In terms of the power of $\widehat{\gamma}_{\text{R}}$ and $\widehat{v}_{\text{R}}$ for testing $\text{H}_0 : \gamma_{\text{target}} = 0$, all four methods have asymptotic power 1, while forward screening possesses higher power in finite samples. Analogously, we can see an improvement in coverage probability of the RLS based confidence

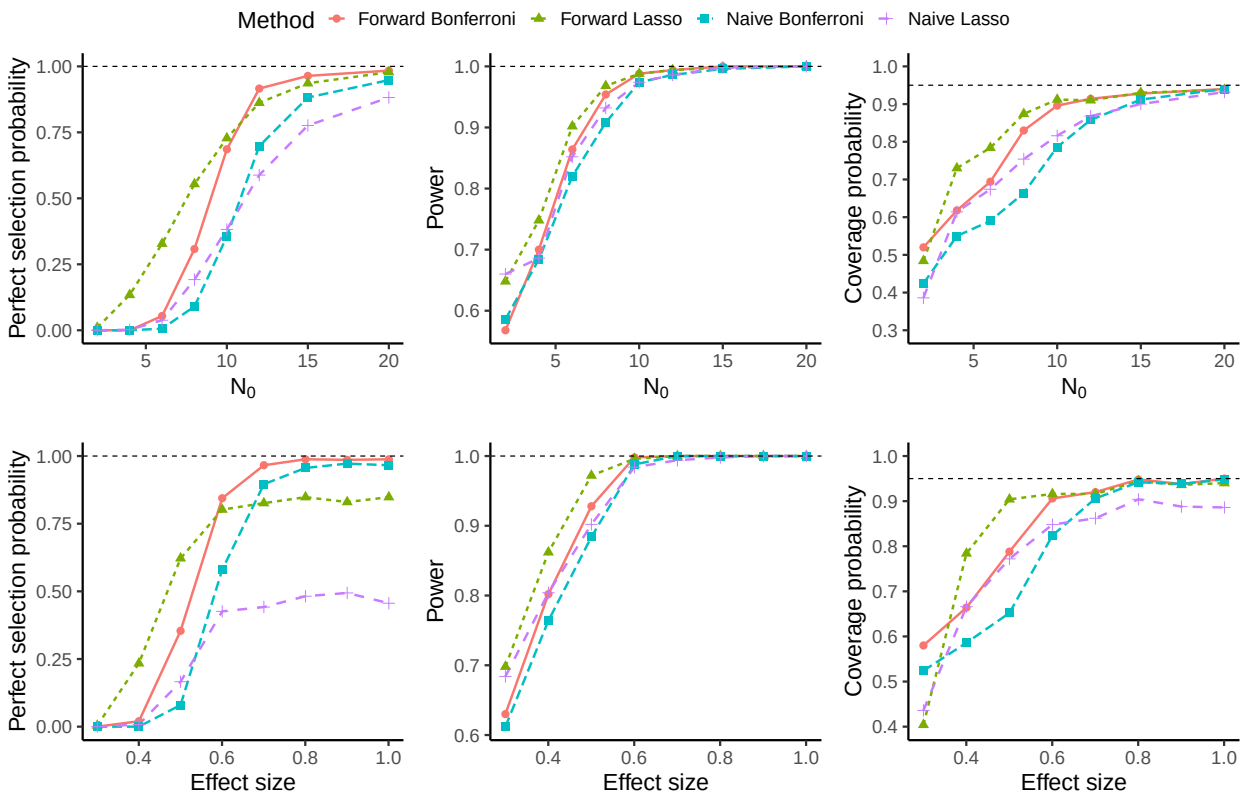

Figure 3: Simulation results on **(G2)**. (i) Top left panel: perfect screening probability with a small fixed effect size 0.20 and varying N_0; (ii) Top middle panel: power curve with a small fixed effect size 0.20 and varying N_0; (iii) Top right panel: coverage probability with a small fixed effect size 0.20 and varying N_0; (iv) Bottom left panel: perfect screening probability with a small fixed replication $N_0 = 2$ and varying effect size; (v) Bottom middle panel: power curve with a small fixed replication $N_0 = 2$ and varying effect size; (vi) Bottom right panel: coverage probability with a small fixed replication $N_0 = 2$ and varying effect size.

intervals with the forward screening procedure.

8 Discussion

We have discussed the formal theory for forward screening and post-screening inference in 2^K factorial designs. It is conceptually straightforward to extend the theory to general factorial designs with multi-valued factors. We omit the technical details. Another important direction is covariate adjustment in factorial experiments. Lin (2013), Lu (2016a) and Liu et al. (2022) demonstrated

the efficiency gain of covariate adjustment with small K. Zhao and Ding (2023) discussed covariate adjustment in factorial experiment with factors and covariates selected independent of data. We leave it to future research to establish the theory for factor screening and covariate selection in factorial designs.

References

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

Angrist, J. D. and Pischke, J.-S. (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press.

Bai, Z., Choi, K. P., Fujikoshi, Y., and Hu, J. (2022), "Asymptotics of AIC, BIC and Cp model selection rules in high-dimensional regression," *Bernoulli*, 28, 2375–2403.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2010), "Hierarchical selection of variables in sparse high-dimensional regression," *IMS Collections*, 6, 28.

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

Blackwell, M. and Pashley, N. E. (2021), "Noncompliance and Instrumental Variables for 2^K Factorial Experiments," *Journal of the American Statistical Association*, in press.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., and Yu, B. (2016), "Lasso adjustments of treatment effect estimates in randomized experiments," *Proceedings of the National Academy of Sciences*, 113, 7383–7390.

Box, G., Hunter, J., and Hunter, W. (2005), *Statistics for Experimenters: Design, Innovation, and Discovery*, Hoboken, NJ: Wiley.

Branson, Z., Dasgupta, T., and Rubin, D. B. (2016), "Improving covariate balance in 2^K factorial designs via rerandomization with an application to a New York City Department of Education High School Study," *Annals of Applied Statistics*, 10, 1958–1976.

Claggett, B., Xie, M., and Tian, L. (2014), "Meta-analysis with fixed, unknown, study-specific parameters," *Journal of the American Statistical Association*, 109, 1660–1671.

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)*, 77, 727–753.

Egami, N. and Imai, K. (2019), "Causal interaction in factorial experiments: application to conjoint analysis," *Journal of the American Statistical Association*, 114, 529–540.

Espinosa, V., Dasgupta, T., and Rubin, D. B. (2016), "A Bayesian perspective on the analysis of unreplicated factorial experiments using potential outcomes," *Technometrics*, 58, 62–73.

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

Fithian, W., Sun, D., and Taylor, J. (2014), "Optimal inference after model selection," *arXiv preprint arXiv:1410.2597*.

Freedman, D. A. (2008), "On regression adjustments to experimental data," *Advances in Applied Mathematics*, 40, 180–193.

Gerber, A. S. and Green, D. P. (2012), *Field Experiments: Design, Analysis, and Interpretation*, New York, NY: Norton.

Guo, X., Wei, L., Wu, C., and Wang, J. (2021), "Sharp Inference on Selected Subgroups in Observational Studies," *arXiv preprint arXiv:2102.11338*.

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, 615–625.

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

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

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2, New York: Springer.

Kempthorne, O. (1952), *The Design and Analysis of Experiments*, New York: Wiley.

Kuchibhotla, A. K., Kolassa, J. E., and Kuffner, T. A. (2022), "Post-selection inference," *Annual Review of Statistics and Its Application*, 9, 505–527.

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, 1759–1769.

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

Lin, W. (2013), "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *Annals of Applied Statistics*, 7, 295–318.

Liu, H., Ren, J., and Yang, Y. (2022), "Randomization-based Joint Central Limit Theorem and Efficient Covariate Adjustment in Randomized Block 2^K Factorial Experiments," *Journal of the American Statistical Association*, in press.

Lu, J. (2016a), "Covariate adjustment in randomization-based causal inference for 2^K factorial designs," *Statistics and Probability Letters*, 119, 11–20.

— (2016b), "On randomization-based and regression-based inferences for 2^K factorial designs," *Statistics and Probability Letters*, 112, 72–78.

Meng, X.-L. and Xie, X. (2014), "I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb?" *Econometric Reviews*, 33, 218–250.

Neyman, J. (1923/1990), "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." *Statistical Science*, 465–472.

Pashley, N. E. and Bind, M.-A. C. (2023), "Causal Inference for Multiple Non-Randomized Treatments using Fractional Factorial Designs," *Canadian Journal of Statistics*, in press.

Rillig, M. C., Ryo, M., Lehmann, A., Aguilar-Trigueros, C. A., Buchert, S., Wulf, A., Iwasaki, A., Roy, J., and Yang, G. (2019), "The role of multiple global change factors in driving soil functions and microbial biodiversity," *Science*, 366, 886–890.

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*.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Wainwright, M. J. (2019), *High-dimensional Statistics: A Non-asymptotic Viewpoint*, vol. 48, Cambridge: Cambridge University Press.

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

Wasserman, L. and Roeder, K. (2009), "High dimensional variable selection," *Annals of Statistics*, 37, 2178.

Wei, W., Zhou, Y., Zheng, Z., and Wang, J. (2022), "Inference on the Best Policies with Many Covariates," *arXiv preprint arXiv:2206.11868*.

Wieczorek, J. and Lei, J. (2022), "Model selection properties of forward selection and sequential cross-validation for high-dimensional regression," *Canadian Journal of Statistics*, 50, 454–470.

Wu, C. J. and Hamada, M. S. (2011), *Experiments: Planning, Analysis, and Optimization*, vol. 552, Hoboken, NJ: John Wiley & Sons.

Wu, Y., Zheng, Z., Zhang, G., Zhang, Z., and Wang, C. (2022), "Non-stationary a/b tests: Optimal variance reduction, bias correction, and valid inference," *Bias Correction, and Valid Inference (May 20, 2022)*.

Yates, F. (1937), "The design and analysis of factorial experiments," Tech. Rep. Technical Communication 35, Imperial Bureau of Soil Science, London, U. K.

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

Zhang, C. (2022), "Social construction of hate crimes in the U.S.: A factorial survey experiment," *Theses and Dissertations–Sociology*, 49.

Zhao, A. and Ding, P. (2021), "Regression-based causal inference with factorial experiments: estimands, model specifications and design-based properties," *Biometrika*, 109, 799–815.

— (2023), "Covariate adjustment in multi-armed, possibly factorial experiments," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, in press.

Zhao, P., Rocha, G., and Yu, B. (2009), "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics*, 37, 3468–3497.

Zhao, P. and Yu, B. (2006), "On model selection consistency of Lasso," *The Journal of Machine Learning Research*, 7, 2541–2563.

Zhao, S., Witten, D., and Shojaie, A. (2021), "In defense of the indefensible: A very naive approach to high-dimensional inference," *Statistical Science*, 36, 562–577.

Supplementary material

Section A provides more discussions/extensions to the results introduced in the main paper. More concretely, Section A.1 presents detailed discussion of the use of weight least squares in factorial experiments. Section A.2 extends the inference results in Section 4 to a vector of causal effects.

Section B presents general results on consistency of forward factor screening. Theorem 1 is a corollary of the results in Section B.

Section C gives the technical proofs of the results in the main paper and the Appendix.

A Additional results

This section provides more extensions to the results in the main paper. Section A.1 discusses the use of WLS in analyzing factorial experiments. Section A.2 extends the inference results under perfect screening (Section 4) to a vector of causal effects.

A.1 Weighted least squares for estimating factorial effects

In this subsection, we briefly state and prove some useful facts about weighted least squares in estimating factorial effects. More discussions can be found in Zhao and Ding (2021). Denote the design matrix as $X = (g_{1,\mathbb{M}}, \ldots, g_{N,\mathbb{M}})^\top$. Let $W = \mathrm{Diag}\{w_i\}$. The problem (2.3) has closed-form solution:

$$\begin{aligned}
\widehat{\tau} &= (X^\top W X)^{-1}(X^\top W Y) \text{ (closed form solution of WLS)} \\
&= \{G(\cdot,\mathbb{M})^\top G(\cdot,\mathbb{M})\}^{-1}\{G(\cdot,\mathbb{M})^\top \widehat{Y}\} \\
&\quad \text{(units under the same treatment arm share the same regressor)} \\
&= Q^{-1} G(\cdot,\mathbb{M})^\top \widehat{Y}. \qquad \text{(S1)}
\end{aligned}$$

The closed form (S1) motivates the variance estimation:

$$\widehat{V}_{\widehat{\tau}} = Q^{-2} G(\cdot,\mathbb{M})^\top \widehat{V}_{\widehat{Y}} G(\cdot,\mathbb{M}). \qquad \text{(S2)}$$

Alternatively, one can use the Eicker–Huber–White (EHW) variance estimation with the HC2 correction (Angrist and Pischke, 2009):

$$\widehat{V}_{\mathrm{EHW}} = (X^\top W X)^{-1} X^\top W \mathrm{Diag}\left\{\frac{\widehat{\epsilon}_i^2}{1 - N_i^{-1}}\right\} W X (X^\top W X)^{-1}, \quad \widehat{\epsilon}_i = Y_i - g_{i,\mathbb{M}}^\top \widehat{\tau}. \qquad \text{(S3)}$$

Again, because units under the same treatment arm share the same regressor, $\widehat{V}_{\text{EHW}}$ simplifies to

$$\widehat{V}_{\text{EHW}} = Q^{-2} G(\cdot, \mathbb{M})^\top \widehat{V}'_{\widehat{Y}} G(\cdot, \mathbb{M}), \tag{S4}$$

where

$$\widehat{V}'_{\widehat{Y}} = \text{Diag}\left\{ N(\boldsymbol{z})^{-1} \widehat{S}'(\boldsymbol{z}, \boldsymbol{z}) \right\}_{\boldsymbol{z} \in \mathcal{T}} \text{ with } \widehat{S}'(\boldsymbol{z}, \boldsymbol{z}) = \frac{1}{N(\boldsymbol{z}) - 1} \sum_{Z_i = \boldsymbol{z}} (Y_i - g_{i,\mathbb{M}}^\top \widehat{\tau})^2.$$

Following some algebra, we can show

$$\begin{aligned} \widehat{S}'(\boldsymbol{z}, \boldsymbol{z}) &= \frac{1}{N(\boldsymbol{z}) - 1} \sum_{Z_i = \boldsymbol{z}} (Y_i - \widehat{Y}(\boldsymbol{z}))^2 + \frac{N(\boldsymbol{z})}{N(\boldsymbol{z}) - 1} \{\widehat{Y}(\boldsymbol{z}) - G(\boldsymbol{z}, \mathbb{M}) \widehat{\tau}\}^2 \\ &= \widehat{S}(\boldsymbol{z}, \boldsymbol{z}) + \frac{N(\boldsymbol{z})}{N(\boldsymbol{z}) - 1} \{\widehat{Y}(\boldsymbol{z}) - G(\boldsymbol{z}, \mathbb{M}) \widehat{\tau}\}^2. \end{aligned}$$

Hence $\widehat{S}'(\boldsymbol{z}, \boldsymbol{z}) \geq \widehat{S}(\boldsymbol{z}, \boldsymbol{z})$. In general $\widehat{Y}(\boldsymbol{z}) \neq G(\boldsymbol{z}, \mathbb{M}) \widehat{\tau}$, so the difference is not negligible. The following Lemma S1 formally summarizes the statistical property of $\widehat{\tau}$ and its two variance estimators, $\widehat{V}_{\widehat{\tau}}$ and $\widehat{V}_{\text{EHW}}$. The proof can be done by utilizing the moment facts from Section C.2 and C.3 of Shi and Ding (2022), which we omit here.

Lemma S1. *Assume Conditions 1 and 3. For the WLS in* (2.3)*, we have*

1. $\widehat{\tau} = Q^{-1} G(\cdot, \mathbb{M})^\top \widehat{Y}$ *is unbiased for the true factorial effects* $\tau(\mathbb{M})$*; i.e.,* $\mathbb{E}\{\widehat{\tau}\} = \tau(\mathbb{M})$*.*

2. *Both variance estimators are consistent and robust:* $N(\widehat{V}_{\widehat{\tau}} - V_{\widehat{\tau},\text{lim}}) = o_{\mathbb{P}}(1)$*,* $N(\widehat{V}_{\text{EHW}} - V_{\text{EHW,lim}}) = o_{\mathbb{P}}(1)$*, with* $V_{\widehat{\tau},\text{lim}} \succcurlyeq V_{\widehat{\tau}}$ *and* $V_{\text{EHW}} \succcurlyeq V_{\widehat{\tau}}$*, where*

$$V_{\widehat{\tau},\text{lim}} = Q^{-2} G(\cdot, \mathbb{M})^\top D_{\widehat{Y}} G(\cdot, \mathbb{M}),$$

and

$$V_{\text{EHW,lim}} = Q^{-2} G(\cdot, \mathbb{M})^\top \text{Diag}\left\{ \frac{1 - N^{-1}}{N(\boldsymbol{z}) - 1} S(\boldsymbol{z}, \boldsymbol{z}) + \frac{1}{N(\boldsymbol{z}) - 1} \{\overline{Y}(\boldsymbol{z}) - G(\boldsymbol{z}, \mathbb{M}) \tau(\mathbb{M})\}^2 \right\} G(\cdot, \mathbb{M}).$$

3. *EHW variance estimator is more conservative than the direct variance estimator:* $\widehat{V}_{\text{EHW}} \succcurlyeq \widehat{V}_{\widehat{\tau}}$*.*

It is worthy of mentioning that in the fixed Q setting, if we assume that the factorial effects that are not included in $\mathbb{M}$ are all zero, Lemma S1 implies EHW variance estimator (S3) or (S4) has the same asymptotic statistical property as the direct variance estimator (S2), which agrees with the conclusion of Zhao and Ding (2021).

A.2 Extension of post-screening inference to vector parameters

In this subsection we present an extension of Theorem 2 to a vector of causal parameters:

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

For convenience we can stack $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_L$ into a weighting matrix $F = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_L)$ and write

$$\Gamma = F^\top \overline{Y}.$$

We will focus on linear projections of Γ, defined as $\gamma_b = b^\top \Gamma$ for a given $b \in \mathbb{R}^L$. Naturally, we can apply forward screening and construct RLS-based estimators for Γ:

$$\widehat{\Gamma}_{\mathrm{R}} = (\widehat{\gamma}_{1,\mathrm{R}}, \ldots, \widehat{\gamma}_{L,\mathrm{R}})^\top, \quad \widehat{V}_{\widehat{\Gamma},\mathrm{R}} = F[\widehat{\mathbb{M}}]^\top \widehat{V}_{\widehat{Y}} F[\widehat{\mathbb{M}}], \tag{S5}$$

where

$$F[\widehat{\mathbb{M}}] = Q^{-1} G(\cdot, \widehat{\mathbb{M}}) G(\cdot, \widehat{\mathbb{M}})^\top F.$$

For γ_b, an estimator based on (S5) is

$$\widehat{\gamma}_{b,\mathrm{R}} = b^\top \widehat{\Gamma}_{\mathrm{R}}, \quad \widehat{v}^2_{b,\mathrm{R}} = b^\top \widehat{V}_{\widehat{\Gamma},\mathrm{R}} b.$$

For standard factorial effects, we can use WLS to obtain the robust covariance matrix (Section A.1). For one single b, we can actually apply Theorem 2 with

$$\boldsymbol{f}_b = Fb = \sum_{l=1}^{L} b_l \boldsymbol{f}_l.$$

Define $\boldsymbol{f}_b^\star = F[\mathbb{M}^\star]b$. We then get the following theorem:

Theorem S1 (Statistical properties linear projections of Γ)**.** *Assume Conditions 1-4. Let $N \to \infty$. Then*

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

where $v^2_{b,\mathrm{R}} = \boldsymbol{f}_b^{\star\top} V_{\widehat{Y}} \boldsymbol{f}_b^\star$. Further assume $\|\boldsymbol{f}_b^\star\|_\infty = O(Q^{-1})$. The variance estimator $\widehat{v}^2_{b,\mathrm{R}}$ is conservative in the sense

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

where $v^2_{b,\mathrm{R,lim}} = \boldsymbol{f}_b^{\star\top} D_{\widehat{Y}} \boldsymbol{f}_b^\star$ is the limiting value of $\widehat{v}^2_{b,\mathrm{R}}$.

The proof of Theorem S1 is similar to that of Theorem 2, which is mainly based on Lemma S5 and thus omitted here. Moreover, for a fixed integer L, Theorem S1 implies joint normality of $\widehat{\Gamma}_{\mathrm{R}}$, a result due to the Cramér-Wold theorem. We summarize the result as the following corollary and omit the proof:

Corollary S1. *Assume a fixed L. Assume Conditions 1-4. We have*

$$V_{\widehat{\Gamma},\mathrm{R}}^{-1/2}(\widehat{\Gamma}_{\mathrm{R}} - \Gamma) \rightsquigarrow \mathcal{N}(0, I_L),$$

where $V_{\widehat{\Gamma},\mathrm{R}} = F[\mathbb{M}^\star]^\top V_{\widehat{Y}} F[\mathbb{M}^\star]$. Further assume $\max_{\|b\|_2=1} \|\boldsymbol{f}_b^\star\|_\infty = O(Q^{-1})$. The variance estimator $\widehat{v}_{b,\mathrm{R}}^2$ is conservative in the sense that

$$N(\widehat{V}_{\widehat{\Gamma},\mathrm{R}} - V_{\widehat{\Gamma},\mathrm{R},\mathrm{lim}}) \xrightarrow{\mathbb{P}} 0, \quad V_{\widehat{\Gamma},\mathrm{R},\mathrm{lim}} \succcurlyeq V_{\widehat{\Gamma},\mathrm{R}},$$

where $V_{\widehat{\Gamma},\mathrm{R},\mathrm{lim}} = F[\mathbb{M}^\star]^\top D_{\widehat{Y}} F[\mathbb{M}^\star]^\top$ is the limiting value of $\widehat{V}_{\widehat{\Gamma},\mathrm{R}}$.

B General results on consistency of forward screening

In this section we provide some theoretical insights into the forward factor screening algorithm (Algorithm 1). The discussion in this section starts from a more broad discussion where we allow the S-step to be general procedures that satisfy certain conditions. We will show Bonferroni corrected marginal t-test is a special case of these procedures.

We start with some regularization conditions to characterize a "good" layer-wise S-step, and ensure the P-step is compatible with the structure of the true factorial effects. In light of this, we use $\mathbb{M}_{d,+}^\star$ to denote the pruned set of effects on the d-th layer based on the true model $\mathbb{M}_{d-1}^\star$ on the previous layer; that is,

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

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*

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

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

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

$$\textit{Consistency:}\ \limsup_{N\to\infty} D \sum_{d=1}^{D} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\} = 0.$$

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 heredity principles can improve screening accuracy as well as interpretability of the screening results.

Theorem S2 (Screening consistency)**.** *Assume Assumption 1 and 2. Then the forward screening procedure* (3.5) *has the following properties:*

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

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

(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:*

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

Theorem S2 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 S2 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.

Lastly, as we have commented earlier, in practice people have many alternative methods for the S-step. They are attractive in factorial experiments because many lead to simple form solutions due to the orthogonality of factorial designs. For example, Lasso is a commonly adopted strategy for variable selection in linear models (Zhao and Yu, 2006). It solves the following penalized WLS problem in factorial settings:

$$\widehat{\mathbb{M}}_{\mathrm{L}} = \{\mathcal{K} : \widehat{\tau}_{\mathrm{L},\mathcal{K}} \neq 0\}, \quad \widehat{\tau}_{\mathrm{L},\mathcal{K}} = \min_{\tau' \in \mathbb{R}^H} \frac{1}{2} \sum_{\boldsymbol{z} \in \mathcal{T}} w_i (Y_i - g_i^\top \tau')^2 + \lambda_{\mathrm{L}} \|\tau'\|_1.$$

Due to the orthogonality of G, the resulting $\widehat{\mathbb{M}}$ has a closed-form solution (Hastie et al., 2009):

$$\widehat{\mathbb{M}}_{\mathrm{L}} = \{\mathcal{K} : |\widehat{\tau}_{\mathcal{K}}| \geq \lambda_{\mathrm{L}}\}.$$

Other methods, such as AIC/BIC (Bai et al., 2022), sure independence screening (Fan and Lv, 2008), etc., are also applicable. With more delicate assumptions and tuning parameter choices, these methods can also be justified theoretically for screening consistency and post-screening inference. We omit the details.

C Technical proofs

In this section we present the technical proofs for the results across the whole paper. Section C.1 presents some preliminary probabilistic results that are useful in randomized experiments which are mainly attributed to Shi and Ding (2022). The main proof starts from Section C.2.

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

In this subsection we present some preliminary probability results that are crucial for our theoretical discussion. 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 estimator

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

Li and Ding (2017) showed that

$$\mathbb{E}\{\widehat{Y}\} = \overline{Y},\ V_{\widehat{Y}} = \mathrm{Var}\left\{\widehat{Y}\right\} = D_{\widehat{Y}} - N^{-1} S. \tag{S6}$$

Then (S6) further leads to the following facts:

$$\begin{aligned} \mathbb{E}\{\widehat{\gamma}\} &= \sum_{\boldsymbol{z} \in \mathcal{T}} \boldsymbol{f}(\boldsymbol{z}) \overline{Y}(\boldsymbol{z}) = \gamma, \\ \mathrm{Var}\{\widehat{\gamma}\} &= \sum_{\boldsymbol{z} \in \mathcal{T}} \boldsymbol{f}(\boldsymbol{z})^2 N(\boldsymbol{z})^{-1} S(\boldsymbol{z}, \boldsymbol{z}) - N^{-1} \boldsymbol{f}^\top S \boldsymbol{f}, \\ \mathbb{E}\{\widehat{v}^2\} &= \sum_{\boldsymbol{z} \in \mathcal{T}} \boldsymbol{f}(\boldsymbol{z})^2 N(\boldsymbol{z})^{-1} S(\boldsymbol{z}, \boldsymbol{z}). \end{aligned} \tag{S7}$$

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

Lemma S2 (Variance concentration and Berry–Esseen bounds in finite population)**.** *Define* $\gamma = \mathbb{E}\{\widehat{\gamma}\}$, $v^2 = \mathrm{Var}(\widehat{\gamma})$ *and* $v_{\lim}^2 = \mathbb{E}\{\widehat{v}^2\}$. *Suppose the following conditions hold:*

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

$$Q^{-2} \sum_{\boldsymbol{z}=1}^{Q} w(\boldsymbol{z})^2 N_{\boldsymbol{z}}^{-1} S(\boldsymbol{z}, \boldsymbol{z}) \leq \sigma_w^2 v^2. \tag{S8}$$

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

$$\max_{\boldsymbol{z} \in [Q]} \frac{1}{N} \sum_{i=1}^{N} \{Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\}^4 \leq \Delta^4. \tag{S9}$$

Then we have the following conclusions:

1. *The variance estimator is conservative for the true variance:* $v_{\lim}^2 \geq v^2$. *Besides, the following tail bound holds:*

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

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

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

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

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

Proof of Lemma S2. 1. See Lemma S13 of Shi and Ding (2022).

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

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

$$\sup_{t\in\mathbb{R}}|\Phi\{(1+a)t+b\} - \Phi\{t\}| \leq |a| + |b|. \tag{S10}$$

(S10) is particularly useful for small choices of a and b. Intuitively, it evaluates the change of Φ under a small affine perturbation of t.

The proof of (S10) is based on a simple step of the mean value theorem: for any $t \in \mathbb{R}$,

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

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

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

Then we can show that

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

For the first term, we have

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

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

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

Besides, by (S10), when $\epsilon_N \le 1/2$, we also have

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

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

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

On the other hand, we can show that

$$\begin{aligned}\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{\widehat{v}}\le t\right\} &\ge \mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v}\le\frac{\widehat{v}}{v}t, \left|\frac{\widehat{v}-v_{\lim}}{v}\right|\le\epsilon_N\right\} \\ &\ge \mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v}\le\left(\frac{v_{\lim}}{v}-\epsilon_N\right)t\right\} - \mathbb{P}\left\{\left|\frac{\widehat{v}-v_{\lim}}{v}\right|\ge\epsilon_N\right\}. \qquad \text{(S11)}\end{aligned}$$

By (S10), when $\epsilon_N \le 1/2$, we also have

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

So we can derive a lower bound analogous to (S11). Note that the results can be analogously generalized to $t \le 0$. Putting pieces together, we can show that for any $t \ge 0$,

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

□

The following corollary shows a Berry–Esseen bound for the studentized statistic 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 S2. *Assume Condition* (S8) *and* (S9) *hold. Let $w = g_{\mathcal{K}}$ for some $\mathcal{K} \in \mathbb{K}$. Then we have a Berry–Esseen bound with the estimated variance:*

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

Proof of Corollary S2. **Lower bound for Nv^2.** Note that $\|w\|_2^2 = Q$ and $\|w\|_\infty = 1$. Using Condition (S8), we have

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

Therefore, the Berry–Esseen bound becomes

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

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

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

□

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

Lemma S3 (Berry Esseen bound with screening)**.** *Assume there exists $\sigma_w>0$ such that*

$$\sum_{\boldsymbol{z}=1}^{Q}\boldsymbol{f}[\mathbb{M}](\boldsymbol{z})^2N_{\boldsymbol{z}}^{-1}S(\boldsymbol{z},\boldsymbol{z})\le\sigma_w^2v^2(\mathbb{M}). \tag{S12}$$

Then

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

Proof of Lemma S3. 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:

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

□

A crucial quantity that appeared in Lemma S3 is the ratio of norms:

$$\frac{\|\boldsymbol{f}[\mathbb{M}]\|_\infty}{\|\boldsymbol{f}[\mathbb{M}]\|_2}. \tag{S13}$$

The following Lemma S4 provides an explicit bound on (S13) which reveals how the ratio is controlled with respect to the size of the working model.

Lemma S4. *For $\boldsymbol{f}[\mathbb{M}]\neq 0$, we have*

$$\frac{\|\boldsymbol{f}[\mathbb{M}]\|_\infty}{\|\boldsymbol{f}[\mathbb{M}]\|_2}\le\left(\frac{|\mathbb{M}|}{Q}\right)^{1/2}. \tag{S14}$$

Proof of Lemma S4. Because the LHS of (S14) is a ratio, based on the definition of $\boldsymbol{f}^\star$ (4.8) we can assume $\|\boldsymbol{f}\|_2=1$ without loss of generality. Due to the orthogonality of G, we can use the columns of G as bases and express $\boldsymbol{f}$ as

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

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

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

Hence

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

□

C.2 Proof of Theorem S2

Proof of Theorem S2. 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. We introduce several key events that will play a crucial role in the proof: for $D_0\in[D]$, define

$$\textit{Under-selection: } \mathcal{E}_{\mathrm{U}}(D_0)=\{\widehat{\mathbb{M}}_d\subset\mathbb{M}_d^\star, d\in[D_0]\},$$

$$\textit{Strict under-selection: } \mathcal{E}_{\mathrm{SU}}(D_0)=\{\widehat{\mathbb{M}}_d\subset\mathbb{M}_d^\star, d\in[D_0];\ \text{there exists } d\in[D_0],\widehat{\mathbb{M}}_d\subsetneq\mathbb{M}_d^\star\}.$$

High level idea of the proof. To prove screening consistency, we will prove two facts:

$$\mathbb{P}\left\{\mathcal{E}_{\mathrm{U}}(D) \text{ holds}\right\} \to 1, \quad \mathbb{P}\left\{\mathcal{E}_{\mathrm{SU}}(D) \text{ holds}\right\} \to 0.$$

Combining these two results together, we can conclude asymptotic screening consistency.

We start from the strict under-selection probability.

Step 1: Prove that asymptotically, there is no strict under-selection.

By definition,

$$\mathbb{P}\left\{\mathcal{E}_{\mathrm{SU}}(1) \text{ holds}\right\} = \mathbb{P}\left\{\widetilde{\mathbb{M}}_1 \subsetneq \mathbb{M}_1^\star\right\} \leq \mathbb{P}\left\{\widetilde{\mathbb{M}}_1^c \cap \mathbb{M}_1^\star \neq \varnothing\right\}.$$

We now derive a recursive bound for $\mathbb{P}\left\{\mathcal{E}_{\mathrm{SU}}(D_0+1) \text{ holds}\right\}$ where $1 \leq D_0 \leq D-1$. We have decomposition

$$\begin{aligned}\mathcal{E}_{\mathrm{SU}}(D_0+1) &= \left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0+1\right\} - \left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0+1\right\} \\ &= \mathcal{E}_{\mathrm{SU},1}(D_0+1) \cup \mathcal{E}_{\mathrm{SU},2}(D_0+1),\end{aligned}$$

where

$$\begin{aligned}\mathcal{E}_{\mathrm{SU},1}(D_0+1) &= \left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0+1\right\} - \left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}, \\ \mathcal{E}_{\mathrm{SU},2}(D_0+1) &= \left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\} - \left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0+1\right\}.\end{aligned}$$

For $\mathcal{E}_{\mathrm{SU},1}(D_0+1)$, we have

$$\begin{aligned}\mathbb{P}\left\{\mathcal{E}_{\mathrm{SU},1}(D_0+1) \text{ holds}\right\} &= \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0+1\right\} - \left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right) \\ &\leq \mathbb{P}\left(\forall d \in [D_0+1], \widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star; \exists d \in [D_0], \widehat{\mathbb{M}}_d \subsetneq \mathbb{M}_d^\star\right) \\ &\leq \mathbb{P}\left(\forall d \in [D_0], \widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star; \exists d \in [D_0], \widehat{\mathbb{M}}_d \subsetneq \mathbb{M}_d^\star\right) \\ &= \mathbb{P}\left\{\mathcal{E}_{\mathrm{SU}}(D_0) \text{ holds}\right\}. \end{aligned} \tag{S15}$$

For $\mathcal{E}_{\mathrm{SU},2}(D_0+1)$, we notice that $\widehat{\mathbb{M}}_{D_0+1}$ 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

$$\begin{aligned}\mathbb{P}\left\{\mathcal{E}_{\mathrm{SU},2}(D_0+1) \text{ holds}\right\} =& \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subsetneq \mathbb{M}_{D_0+1}^\star\right) \\ =& \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widetilde{\mathbb{M}}_{D_0+1} \subsetneq \mathbb{M}_{D_0+1}^\star\right) \\ \leq& \mathbb{P}\left(\widetilde{\mathbb{M}}_{D_0+1}^c \cap \mathbb{M}_{D_0+1}^\star \neq \varnothing\right). \end{aligned} \tag{S16}$$

Now (S15) and (S16) together suggest that

$$\begin{aligned}
&\mathbb{P}\left\{\mathcal{E}_{\mathrm{SU}}(D_0+1) \text{ holds}\right\} \\
\leq &\mathbb{P}\left\{\mathcal{E}_{\mathrm{SU}}(D_0) \text{ holds}\right\} + \mathbb{P}\left\{\widetilde{\mathbb{M}}_{D_0+1}^{c} \cap \mathbb{M}_{D_0+1}^{\star} \neq \varnothing\right\} \\
\leq &\cdots \leq \sum_{d=1}^{D_0+1} \mathbb{P}\left\{\widetilde{\mathbb{M}}_{D_0+1}^{c} \cap \mathbb{M}_{D_0+1}^{\star} \neq \varnothing\right\}. \qquad \text{(S17)}
\end{aligned}$$

Taking $D_0 = D-1$ in (S17) and apply Assumption 1, we conclude

$$\mathbb{P}\left\{\mathcal{E}_{\mathrm{SU}}(D) \text{ holds}\right\} \to 0.$$

Step 2: Prove the first part of Theorem S2 and give a probability bound for under-selection. We compute the probability for under-selection:

$$\begin{aligned}
&\mathbb{P}\left\{\mathcal{E}_{\mathrm{U}}(D) \text{ fails}\right\} \\
=&\mathbb{P}\left\{\mathcal{E}_{\mathrm{U}}(1) \text{ fails}\right\} + \sum_{D_0=2}^{D} \mathbb{P}\left\{\mathcal{E}_{\mathrm{U}}(D_0-1) \text{ holds}; \mathcal{E}_{\mathrm{U}}(D_0) \text{ fails}\right\} \\
=&\mathbb{P}\left\{\mathcal{E}_{\mathrm{U}}(1) \text{ fails}\right\} (\triangleq \circledast_1) + \sum_{D_0=2}^{D} \mathbb{P}\left\{\mathcal{E}_{\mathrm{P}}(D_0-1) \text{ holds}; \mathcal{E}_{\mathrm{U}}(D_0) \text{ fails}\right\} (\triangleq \circledast_2) \\
&+ \sum_{D_0=2}^{D} \mathbb{P}\left\{\mathcal{E}_{\mathrm{SU}}(D_0-1) \text{ holds}; \mathcal{E}_{\mathrm{U}}(D_0) \text{ fails}\right\} (\triangleq \circledast_3).
\end{aligned}$$

For $\circledast_1$, by definition of $\mathcal{E}_{\mathrm{U}}(1)$ we have

$$\circledast_1 = \mathbb{P}\left\{\mathcal{E}_{\mathrm{U}}(1) \text{ fails}\right\} = \mathbb{P}\left\{\widehat{\mathbb{M}}_1 \cap \mathbb{M}_1^{\star c} \neq \varnothing\right\} = \mathbb{P}\left\{\widetilde{\mathbb{M}}_1 \cap \mathbb{M}_1^{\star c} \neq \varnothing\right\}. \qquad \text{(S18)}$$

For $\circledast_2$, we have

$$\circledast_2 \leq \sum_{D_0=2}^{D} \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^{\star}, d \in [D_0-1]; \widetilde{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star c} \neq \varnothing\right) \leq \sum_{D_0=2}^{D} \mathbb{P}\left\{\widetilde{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star c} \neq \varnothing\right\}, \quad \text{(S19)}$$

which is because on the given event, $\widehat{\mathbb{M}}_{D_0,+} = \mathrm{H}(\widehat{\mathbb{M}}_{D_0-1}) = \mathrm{H}(\mathbb{M}_{D_0-1}^{\star}) = \mathbb{M}_{D_0,+}^{\star}$ and $\widehat{\mathbb{M}}_{D_0} = \widehat{\mathrm{S}}(\widehat{\mathbb{M}}_{D_0,+}) = \widetilde{\mathbb{M}}_{D_0}$.

From (S18) and (S19),

$$\limsup_{N\to\infty} (\circledast_1 + \circledast_2) = \sum_{D_0=1}^{D} \mathbb{P}\left\{\widetilde{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star c} \neq \varnothing\right\} \leq \sum_{D_0=1}^{D} \alpha_{D_0} = \alpha. \text{ (by Assumption 1)} \quad \text{(S20)}$$

For $\circledast_3$, we have

$$\begin{aligned}\circledast_3 &\leq \sum_{D_0=2}^{D} \mathbb{P}\left\{\mathcal{E}_{\text{SU}}(D_0-1) \text{ holds}\right\} \\ &\leq \sum_{D_0=2}^{D}\sum_{d=1}^{D_0-1} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\} \text{ (using (S17))} \\ &= \sum_{d=1}^{D-1}(D-d)\mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\} \to 0. \text{ (using Assumption 1)} \end{aligned} \tag{S21}$$

Therefore, by (S20) and (S21), the probability of failure of under-selection gets controlled under α asymptotically.

As a side product, we obtain finite sample bounds:

$$\mathbb{P}\left\{\mathcal{E}_{\text{U}}(D) \text{ fails}\right\} \leq \sum_{D_0=1}^{D} \mathbb{P}\left\{\widetilde{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star c} \neq \varnothing\right\} + \sum_{d=1}^{D-1}(D-d)\mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\}.$$

Step 3. Prove of the second part of Theorem S2 and conclude screening consistency. Under $\alpha = \alpha(N) \to 0$, the first part of the result implies that with probability tending to one, we have under-selection:

$$\mathbb{P}\left\{\mathcal{E}_{\text{U}}(D) \text{ holds}\right\} \to 1.$$

By (S17) and Assumption 1, strict under-selection will not happen with high probability:

$$\mathbb{P}\left\{\mathcal{E}_{\text{SU}}(D) \text{ holds}\right\} \to 0.$$

Therefore, we conclude the consistency of the screening procedure.

□

C.3 Proof of Theorem 1

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

Theorem S3 (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) $\limsup_{N\to\infty} \sum_{d=1}^{D} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} \neq \varnothing\right\} \leq \sum_{d=1}^{D} \alpha_d = \alpha.$

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

(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 screening) 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 S2. 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. (i) First, we show validity:

$$\begin{aligned}
\mathbb{P}\left\{\widetilde{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} \neq \varnothing\right\} &= \mathbb{P}\left\{\exists \mathcal{K} \in \mathbb{M}_{d,+}^\star \backslash \mathbb{M}_d^\star, \left|\frac{\widehat{\tau}_{\mathcal{K}}}{\widehat{v}_{\mathcal{K},\mathrm{R}}}\right| \geq \Phi^{-1}\left(1-\frac{\alpha_d}{2|\mathbb{M}_{d,+}^\star|}\right)\right\} \\
&\leq \sum_{\mathcal{K}\in\mathbb{M}_{d,+}^\star\backslash\mathbb{M}_d^\star} \mathbb{P}\left\{\left|\frac{\widehat{\tau}_{\mathcal{K}}}{\widehat{v}_{\mathcal{K},\mathrm{R}}}\right| \geq \Phi^{-1}\left(1-\frac{\alpha_d}{2|\mathbb{M}_{d,+}^\star|}\right)\right\} \\
&\leq \sum_{\mathcal{K}\in\mathbb{M}_{d,+}^\star\backslash\mathbb{M}_d^\star} \left(\frac{\alpha_d}{|\mathbb{M}_{d,+}^\star|} + \frac{\widetilde{C}}{(QN_0)^{1/3}}\right) \text{ (by Corollary S2)} \\
&\leq \left(\alpha_d + \frac{\widetilde{C}|\mathbb{M}_{d,+}^\star|}{N^{1/3}}\right).
\end{aligned}$$

Hence,

$$\sum_{d=1}^{D}\mathbb{P}\left\{\widetilde{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} \neq \varnothing\right\} \leq \sum_{d=1}^{D}\left(\alpha_d + \frac{\widetilde{C}|\mathbb{M}_{d,+}^\star|}{N^{1/3}}\right).$$

Due to the effect heredity condition 4, we have

$$|\mathbb{M}_{1,+}^\star| = |\mathbb{M}_1^\star|, \quad |\mathbb{M}_{d,+}^\star| \leq K|\mathbb{M}_{d-1}^\star|.$$

Hence

$$\limsup_{N\to\infty}\sum_{d=1}^{D}\mathbb{P}\left\{\widetilde{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} \neq \varnothing\right\} \leq \alpha + \limsup_{N\to\infty}\frac{K\widetilde{C}|\mathbb{M}^\star|}{N^{1/3}} = \alpha. \text{ (using Condition 2(iii))}$$

(ii) Second, we show consistency. Assume the nonzero $\tau_{\mathcal{K}}$'s are positive. If some are negative one can simply modify the direction of some of the inequalities and still validate the proof.

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

For simplicity, let

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

Then

$$
\begin{aligned}
&\mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\} \\
&\leq \sum_{\mathcal{K}\in\mathbb{M}_d^\star}\left(\mathbb{P}\left\{-Z_d^\star - \frac{\tau_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}} \leq \frac{\widehat{\tau}_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}} - \frac{\tau_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}} \leq Z_d^\star - \frac{\tau_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}}\right\} + \frac{\widetilde{C}}{(QN_0)^{1/3}}\right) \\
&= \sum_{\mathcal{K}\in\mathbb{M}_d^\star} \Phi\left\{r_{\mathcal{K}}^{-1}\left(Z_d^\star - \frac{\tau_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}}\right)\right\} - \Phi\left\{r_{\mathcal{K}}^{-1}\left(-Z_d^\star - \frac{\tau_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}}\right)\right\} (\triangleq \circledast) + \frac{\widetilde{C}|\mathbb{M}_d^\star|}{(QN_0)^{1/3}}.
\end{aligned}
$$

With Condition 2, we have

$$
\begin{aligned}
Z_d^\star = \Theta\left(\sqrt{2\ln\frac{2|\mathbb{M}_{d,+}^\star|}{\alpha_d}}\right) &= \Theta(\sqrt{(\delta' + \delta''/3)\ln N}), \\
\left|\frac{\tau_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}}\right| &= \Theta(N^{1/2+\delta}) = \Theta(N^{\delta_0}) \text{ (by defining } \delta_0 = 1/2 + \delta > 0).
\end{aligned}
$$

Because $\delta > -1/2$ and $\delta' \geq 0$, we have $|\frac{\tau_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}}| \to \infty$ and $Z_d^\star/(|\frac{\tau_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}}|) \to 0$. Therefore,

$$
\Phi\left\{r_{\mathcal{K}}^{-1}\left(Z_d^\star - \frac{\tau_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}}\right)\right\} - \Phi\left\{r_{\mathcal{K}}^{-1}\left(-Z_d^\star - \frac{\tau_{\mathcal{K}}}{v_{\mathcal{K}.\mathrm{R}}}\right)\right\} = \Theta(N^{-\delta_0}\exp\{-N^{2\delta_0}/2\}).
$$

Now applying Condition 2 again, we have

$$
D\sum_{d=1}^{D}\mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_d^\star \neq \varnothing\right\} = \Theta\left(D|\mathbb{M}^\star|N^{-\delta_0}\exp\{-N^{2\delta_0}/2\} + D|\mathbb{M}^\star|/N^{1/3}\right) = o(1).
$$

(iii) The Type I error rate control comes from Theorem S2.

(iv) The perfect selection result follows from Theorem S2.

□

C.4 Proof of Theorem 2

Theorem 2 is a direct result of Theorem 1, Lemma S2 and the following Berry–Esseen bound:

Lemma S5 (Berry–Esseen bound under perfect screening)**.** *Assume* (S12). *Then*

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

Proof of Lemma S5. This lemma is a direct application of Lemma S3. First we check that

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

From the definition of γ (S7), 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 S3 with $\mathbb{M}=\mathbb{M}^\star$ to get the result of Theorem 2.

C.5 Statement and proof of Lemma S6

The following lemma gives the closed form solution of the RLS estimator (4.7).

Lemma S6. $\widehat{Y}_{\mathrm{R}}$ *from* (4.7) *can be expressed as:*

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

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

Proof of Lemma S6. 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 (4.7), 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 (4.7).

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

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

□

C.6 Proof of Proposition 1

Proof of Proposition 1. (i) Based on the definition of v_{R}^2 and v^2, we have

$$\frac{v_{\mathrm{R}}^2}{v^2} = \frac{\boldsymbol{f}^{\star\top} V_{\widehat{Y}} \boldsymbol{f}^\star}{\boldsymbol{f}^\top V_{\widehat{Y}} \boldsymbol{f}} = \frac{\|\boldsymbol{f}^\star\|_2^2}{\|\boldsymbol{f}\|_2^2}$$

because $\kappa(V_{\widehat{Y}}) = 1$. We further compute

$$\frac{\|\boldsymbol{f}^\star\|_2^2}{\|\boldsymbol{f}\|_2^2} = \frac{\boldsymbol{f}^\top \{Q^{-1}G(\cdot, \mathbb{M}^\star)G(\cdot, \mathbb{M}^\star)^\top\} \boldsymbol{f}}{\boldsymbol{f}^\top \boldsymbol{f}} \leq 1$$

where the inequality holds because of the following dominance relationship:

$$Q^{-1}G(\cdot, \mathbb{M}^\star)G(\cdot, \mathbb{M}^\star)^\top \preccurlyeq I_Q.$$

(ii) Because the order of the nonzero elements in $\boldsymbol{f}$ is not crucial here, we assume the first $s^\star$ coordinates of $\boldsymbol{f}$ are nonzero while the rest are zero without loss of generality. We can compute

$$\frac{v_{\mathrm{R}}^2}{v^2} = \frac{\boldsymbol{f}^{\star\top} V_{\widehat{Y}} \boldsymbol{f}^\star}{\boldsymbol{f}^\top V_{\widehat{Y}} \boldsymbol{f}} \leq \kappa(V_{\widehat{Y}}) \cdot \frac{\|\boldsymbol{f}^\star\|_2^2}{\|\boldsymbol{f}\|_2^2}. \tag{S22}$$

For $\boldsymbol{f}^\star$ we have

$$\begin{aligned}\|\boldsymbol{f}^\star\|_2 &= \|Q^{-1}G(\cdot,\mathbb{M}^\star)G(\cdot,\mathbb{M}^\star)^\top \boldsymbol{f}\|_2 \\ &= \left\| Q^{-1}G(\cdot,\mathbb{M}^\star)G(\cdot,\mathbb{M}^\star)^\top \left\{\sum_{s=1}^{s^\star} \boldsymbol{f}(s)\boldsymbol{e}_s\right\}\right\|_2 \\ &\le \sum_{s=1}^{s^\star} |\boldsymbol{f}(s)| \|Q^{-1}G(\cdot,\mathbb{M}^\star)G(\cdot,\mathbb{M}^\star)^\top \boldsymbol{e}_s\|_2 \\ &= \left(\frac{|\mathbb{M}^\star|}{Q}\right)^{1/2} \sum_{s=1}^{s^\star} |\boldsymbol{f}(s)| = \left(\frac{|\mathbb{M}^\star|}{Q}\right)^{1/2} \|\boldsymbol{f}\|_1.\end{aligned}$$

Then we have

$$\frac{\|\boldsymbol{f}^\star\|_2^2}{\|\boldsymbol{f}\|_2^2} \le \frac{|\mathbb{M}^\star|}{Q}\frac{\|\boldsymbol{f}\|_1^2}{\|\boldsymbol{f}\|_2^2} \le \frac{s^\star|\mathbb{M}^\star|}{Q}. \tag{S23}$$

Combining (S22) and (S23), we conclude the result. □

Proposition S1 (Asymptotic length of confidence comparison)**.** *Assume that both $\widehat{\gamma}$ and $\widehat{\gamma}_{\mathrm{R}}$ converge to a normal distribution as the sample size tends to infinity. Assume the variance estimators are consistent: $N(\widehat{v}^2 - v_{\lim}^2) = o_{\mathbb{P}}(1)$, $N(\widehat{v}_{\mathrm{R}}^2 - v_{\mathrm{R},\lim}^2) = o_{\mathbb{P}}(1)$.*

(i) If the condition number of $D_{\widehat{Y}}$ satisfies $\kappa(D_{\widehat{Y}}) = 1$, we have

$$\frac{v_{\mathrm{R},\lim}^2}{v_{\lim}^2} \le 1.$$

(ii) Let $s^\star$ denote the number of nonzero elements in $\boldsymbol{f}$. , then we have

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

C.7 Proof of Theorem 3

Proof of Theorem 3. According to Condition 5 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 S5 with

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

We only need to verify $\gamma = \gamma[\mathbb{M}]$ under the orthogonality condition (5.13).

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

Now by (5.13), $\boldsymbol{f}^\top G(\cdot, \mathbb{M}^c) = 0$. 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.8 Proof of Theorem 4

Proof of Theorem 4. This proof can be finished by applying Lemma S3 and S4 with $\mathbb{M} = \overline{\mathbb{M}}^\star$ and checking $\gamma[\overline{\mathbb{M}}^\star] = \gamma$, which is omitted here. □

C.9 Proof of Proposition 1

Proof of Proposition 1. Because the order of the nonzero elements in $\boldsymbol{f}^\star$ is not crucial, we assume only the first $s^\star$ elements of $\boldsymbol{f}$ are nonzero. That is,

$$\boldsymbol{f} = f_1\boldsymbol{e}_1 + \cdots + f_{s^\star}\boldsymbol{e}_{s^\star}. \tag{S24}$$

We can verify that

$$\|Q^{-1}G(\cdot, \mathbb{M}^\star)G(\cdot, \mathbb{M}^\star)^\top \boldsymbol{e}_k\|_2 = \frac{|\mathbb{M}^\star|}{Q}, \quad \forall\ k \in [Q]. \tag{S25}$$

Therefore,

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

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

$$\frac{\|\boldsymbol{f}^\star\|_2^2}{\|\boldsymbol{f}\|_2^2} \le 1. \tag{S26}$$

On the other hand, using (S24) and (S25), we have

$$\frac{\|\boldsymbol{f}^\star\|_2^2}{\|\boldsymbol{f}\|_2^2} \le \frac{\|\boldsymbol{f}\|_1^2}{\|\boldsymbol{f}\|_2^2} \cdot \frac{|\mathbb{M}^\star|}{Q} \le \frac{s^\star|\mathbb{M}^\star|}{Q}. \tag{S27}$$

Combining (S26) and (S27) concludes the proof. □

C.10 Proof of Theorem 5

For simplicity, we focus on the case given by (6.14). The general proof can be completed similarly. We begin by the following lemma:

Lemma S7 (Consistency of the selected tie sets)**.** *Assume Conditions 1, 3 and 6. There exists universal constants $C, C' > 0$, such that when $N > n(\delta_1, \delta_2, \delta_3)$, we have*

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

Lemma S7 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 S7. 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 S3 to establish a Berry–Esseen bound for each $\widehat{Y}_{\mathrm{R}}^\star(\boldsymbol{z})$. Note that

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

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

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

and obtain

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

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

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

It is known that (Wainwright, 2019, Exercise 2.2)

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

Hence

$$\begin{aligned}
&\mathbb{P}\left\{ \sqrt{N} \left| \widehat{Y}_{\mathrm{R}}^\star(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z}) \right| \ge \sqrt{N} d_h^\star \right\} \\
\le& \frac{v_N}{\sqrt{2\pi} d_h^\star} \cdot \exp\left(-\frac{d_h^{\star 2}}{2v_N^2}\right) + 2C\sigma \frac{\underline{c}^{-1} \max_{i\in[N],\boldsymbol{z}\in[Q]} |Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2} \{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}} \cdot \sqrt{\frac{|\mathbb{M}^\star|}{N_0 Q}}.
\end{aligned} \tag{S28}$$

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 (S28) we have

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z}')-\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})<0\right\}\\
\leq&\frac{\sqrt{\bar{c}\Delta|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0 Q d_h^{\star}}}\cdot\exp\left(-\frac{N_0 Q d_h^{\star 2}}{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}.
\end{aligned}$$

Now a union bound gives

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

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

$$\frac{\sqrt{\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0 Q d_h^{\star}}}\cdot\exp\left(-\frac{N_0 Q d_h^{\star 2}}{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)=\Theta\left(\sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_1}}}\exp\left\{-\frac{C'N^{1+2\delta_1}}{\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right\}\right)$$

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

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z}')-\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})<0\right\}\\
\geq&1-C|\mathcal{T}_1||\mathcal{T}'|\left\{\sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_1}}}\exp\left\{-\frac{C'N^{1+2\delta_1}}{\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right\}+\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}
\end{aligned}\tag{S29}$$

Nice separation. Consider the following random index:

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

For any $\bar{\epsilon}>0$,

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{R}}^{\star}(\widetilde{\boldsymbol{z}})|/\eta_N\geq 2\bar{\epsilon}\right\}\\
&\geq\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1}|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z}')|/\eta_N\geq 2\bar{\epsilon},\widetilde{\boldsymbol{z}}\in\mathcal{T}_1\right\}\\
&\geq\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1}|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z}')|/\eta_N\geq 2\bar{\epsilon}\right\}+\mathbb{P}\left\{\widetilde{\boldsymbol{z}}\in\mathcal{T}_1\right\}-1\\
&\geq\mathbb{P}\left\{\widetilde{\boldsymbol{z}}\in\mathcal{T}_1\right\}-\sum_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1}\mathbb{P}\left\{|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{R}}^{\star}(\widetilde{\boldsymbol{z}}')|/\eta_N\leq 2\bar{\epsilon}\right\}.
\end{aligned}\tag{S30}$$

To proceed we have the following tail bound:

$$
\begin{aligned}
&\mathbb{P}\left\{|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z}')|/\eta_N \leq 2\bar{\epsilon}\right\}\\
=&\mathbb{P}\left\{|\{\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})\}-\{\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}')\}-\{\overline{Y}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}')\}| \leq 2\bar{\epsilon}\eta_N\right\}\\
\leq&\mathbb{P}\left\{|\overline{Y}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z}')|-|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|-|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}')| \leq 2\bar{\epsilon}\eta_N\right\}\\
\leq&\mathbb{P}\left\{|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|+|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}')| \geq 2d_h^{\star}-2\bar{\epsilon}\eta_N\right\}\\
\leq&\mathbb{P}\left\{|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})| \geq d_h^{\star}-\bar{\epsilon}\eta_N\right\}+\mathbb{P}\left\{|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}')| \geq d_h^{\star}-\bar{\epsilon}\eta_N\right\}\\
&\text{(because } \boldsymbol{z}\notin\mathcal{T}_1 \text{ and } \boldsymbol{z}'\in\mathcal{T}_1)\\
\leq&4\Bigg\{\frac{\sqrt{\bar{c}\Delta|\mathbb{M}^{\star}|}}{\sqrt{2\pi N_0 Q}(d_h^{\star}-\bar{\epsilon}\eta_N)}\cdot\exp\left(-\frac{N_0Q(d_h^{\star}-\bar{\epsilon}\eta_N)^2}{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)\\
&+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}\Bigg\}.\\
&\text{(This is deduced analogously to the proof in the previous part)}
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
&\sum_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1}\mathbb{P}\left\{|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z}')|/\eta_N \leq 2\bar{\epsilon}\right\}\\
\leq&4|\mathcal{T}_1||\mathcal{T}'|\left\{\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{\pi N_0Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0Qd_h^{\star 2}}{8\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}\right\}.
\end{aligned}
$$

Combined with (S30), we have:

$$
\begin{aligned}
&\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}_{\mathrm{R}}^{\star}(\boldsymbol{z})-\widehat{Y}_{\mathrm{R}}^{\star}(\widetilde{\boldsymbol{z}})|/\eta_N \geq 2\bar{\epsilon}\right\}\\
\geq&\mathbb{P}\{\widetilde{m}\in\mathcal{T}_1\}-\underbrace{4|\mathcal{T}_1||\mathcal{T}'|\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^{\star}|}}{\sqrt{\pi N_0Q}d_h^{\star}}\cdot\exp\left(-\frac{N_0Qd_h^{\star 2}}{8\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)}_{\text{Term I}}\\
&\underbrace{-4|\mathcal{T}_1||\mathcal{T}'|2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}}_{\text{Term II}}.
\end{aligned}
$$

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

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

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

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

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

$$\begin{aligned}
&\mathbb{P}\left\{|\widehat{Y}^\star_{\mathrm{R}}(\boldsymbol{z})-\widehat{Y}^\star_{\mathrm{R}}(\boldsymbol{z}')|/\eta_N>2\underline{\epsilon}\right\}\\
&\leq\mathbb{P}\left\{|\widehat{Y}^\star_{\mathrm{R}}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|\geq\underline{\epsilon}\eta_N-d_h\right\}+\mathbb{P}\left\{|\widehat{Y}^\star_{\mathrm{R}}(\boldsymbol{z}')-\overline{Y}(\boldsymbol{z}')|\geq\underline{\epsilon}\eta_N-d_h\right\}\\
&\leq 4\Bigg\{\frac{\sqrt{\bar{c}\bar{s}|\mathbb{M}^\star|}}{\sqrt{2\pi N_0Q}(\underline{\epsilon}\eta_N-d_h)}\cdot\exp\left(-\frac{N_0Q(\underline{\epsilon}\eta_N-d_h)^2}{2\bar{c}\bar{s}|\mathbb{M}^\star|}\right)\\
&+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^\star|}{N_0Q}}\Bigg\}.
\end{aligned}$$

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,

$$\begin{aligned}
&\mathbb{P}\left\{|\widehat{Y}^\star_{\mathrm{R}}(\boldsymbol{z})-\widehat{Y}^\star_{\mathrm{R}}(\boldsymbol{z}')|/\eta_N>2\underline{\epsilon}\right\}\\
\leq&4\left\{\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^\star|}}{\sqrt{\pi N_0Q}(\underline{\epsilon}\eta_N)}\cdot\exp\left(-\frac{N_0Q(\underline{\epsilon}\eta_N)^2}{8\bar{c}\bar{s}|\mathbb{M}^\star|}\right)+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^\star|}{N_0Q}}\right\}.
\end{aligned}$$

Hence we have:

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{R}}(\boldsymbol{z})-\widehat{Y}^\star_{\mathrm{R}}(\widetilde{\boldsymbol{z}})|/\eta_N\leq 2\underline{\epsilon}\right\}\\
\geq&\mathbb{P}\left\{\widetilde{\boldsymbol{z}}\in\mathcal{T}_1\right\}-\underbrace{4|\mathcal{T}_1||\mathcal{T}'|\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^\star|}}{\sqrt{\pi N_0Q}(\underline{\epsilon}\eta_N)}\cdot\exp\left(-\frac{N_0Q(\underline{\epsilon}\eta_N)^{\star 2}}{8\bar{c}\bar{s}|\mathbb{M}^\star|}\right)}_{\text{Term I}}\\
&\underbrace{-4|\mathcal{T}_1||\mathcal{T}'|2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^\star|}{N_0Q}}}_{\text{Term II}}.
\end{aligned}$$

Again, by the conditions, we can show

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

Applying (S29) we know that

$$\begin{aligned}
&\mathbb{P}\{\widetilde{\boldsymbol{z}}_h\in\mathcal{T}_1\}\\
\ge&1-C|\mathcal{T}'||\mathcal{T}_1|\left\{\sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_2}}}\exp\left(-\frac{C'N^{1+2\delta_2}}{\bar{c}\bar{s}|\mathbb{M}^{\star}|}\right)+\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^{\star}|}{N_0Q}}\right\}.
\end{aligned}$$

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

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

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)

$$\begin{aligned}
\ge&\mathbb{P}\Bigg\{|\widehat{Y}^{\star}_{\mathrm{R}}(\boldsymbol{z})-\widehat{Y}^{\star}_{\mathrm{R}}(\widetilde{\boldsymbol{z}}_h)|\le\underline{\epsilon}\eta_N,\ \text{for } \boldsymbol{z}\in\mathcal{T}_1;
\end{aligned}$$

$$|\widehat{Y}^{\star}_{\mathrm{R}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{R}}(\widetilde{\boldsymbol{z}}_h)| > \bar{\epsilon}\eta_N, \text{ for } \boldsymbol{z} \notin \mathcal{T}_1 \Bigg\} - \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^{\star}\}$$

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

$$\begin{aligned}
=&\mathbb{P}\Bigg\{ \max_{\boldsymbol{z}\in\mathcal{T}_1} |\widehat{Y}^{\star}_{\mathrm{R}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{R}}(\widetilde{\boldsymbol{z}}_h)| \leq \underline{\epsilon}\eta_N; \min_{\boldsymbol{z}\notin\mathcal{T}_1} |\widehat{Y}^{\star}_{\mathrm{R}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{R}}(\widetilde{\boldsymbol{z}}_h)| > \bar{\epsilon}\eta_N \Bigg\} \\
&- \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^{\star}\} \\
\geq& 1 - \sum_{h=1}^{H_0} \left(1 - \mathbb{P}\left\{ \max_{\boldsymbol{z}\in\mathcal{T}_1} |\widehat{Y}^{\star}_{\mathrm{R}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{R}}(\widetilde{\boldsymbol{z}}_h)| \leq \underline{\epsilon}\eta_N; \min_{\boldsymbol{z}\notin\mathcal{T}_1} |\widehat{Y}^{\star}_{\mathrm{R}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{R}}(\widetilde{\boldsymbol{z}}_h)| > \bar{\epsilon}\eta_N \right\} \right) \\
&- \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^{\star}\} \\
\geq& 1 - \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^{\star}\} \\
&- C|\mathcal{T}'||\mathcal{T}_1| \left\{ \sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_2}}} \exp\left(-\frac{C' N^{1+2\delta_2}}{\bar{c}\bar{s}|\mathbb{M}^{\star}|} \right) + \sigma \frac{\underline{c}^{-1} \max_{i\in[N],\boldsymbol{z}\in[Q]} |Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2} \{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}} \right\}.
\end{aligned}$$

□

Lemma S7 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 S5 and Lemma S7.