Forward screening in factorial experiments

Abstract

Factorial designs have been widely utilized in many fields.

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

1 Introduction

1.1 Motivation and our contribution

(ignore the intro for now...)

Factorial designs have been widely utilized and studied in many fields, witnessing success in agricultural, industrial, and biomedical applications(Wu and Hamada (2011); Zhao and Ding (2021b); Egami and Imai (2018)). The power of factorial designs lies in their ability to simultaneously accommodate multiple factors and provide informative assessments for the magnitude of main causal effects and interactions.

wu2011experiments, zhao2021regression, egami2018c

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

However, several challenges hinder the application of heuristic statistical methods. First, due to the well-recognized “winner’s curse” phenomenon, a naive peek at the combinations with the largest effect sizes might lead to overly optimistic estimates and a deficiency in the coverage rates lee2018winner, andrews2019inference (Lee and Shen, 2018; Andrews et al., 2019). Second, the number of treatment groups in factorial experiments are often quite large and can exponentially increase in scale as the number of factors grows, which leads to less accurate estimators or confidence intervals and prohibits the implementation of many common analytical strategies such as covariate adjustments. Moreover, beyond these methodological concerns, there are a variety of practical constraints that practitioners might wish to impose. For example, in the strategy combination example presented in the previous paragraph, it is of interest to position ourselves in a scenario where only restricted resources or budgets are available and we are confronted with a maximum limit in the total number of strategies we can apply. How to incorporate such realistic constraints in the analysis stands as another important problem. Last but not least, when entangled with a finite population discussion, the aforementioned problems are all missing pieces in the puzzle of factorial experiment theories.

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

May 2, 2022, Lei – COMMENT: Also includes brief summary on the simulation results and case study. But haven't finished these sections.

1.2 Literature review

In the realm of factorial experiments, the factor-based regression typically serves as a dominant strategy for delivering point estimators and confidence regions, due to its simplicity and flexibility in real-life applications. For example, dasgupta2015causal Dasgupta et al. (2015) extended the classical notion of factorial effects to causal counterparts by introducing potential outcome framework and contrast designs. zhao2021regression Zhao and Ding (2021b) studied the use of both saturated and unsaturated linear models for estimating the factorial causal effects. They discussed the parameter specifications of the regression models and justified the commonly used ordinary least squares (OLS) practice from a theoretical perspective. pashley2019causal Pashley and Bind (2019) highlighted the desirable property of regression schemes combined with fractional factorial designs when full designs are possible due to constraints on resources such as units or cost. zhao2021covariate Zhao and Ding (2021a) explores the possibility of incorporating covariate information and applying restricted least squares (RLS) for multiple treatment experimental designs, including factorial studies as a special instance.

A closely related thread of research in factorial designs focuses on variable screening and screening. Powerful variable selection procedures can significantly reduce the complexity of the working model and lead to aditional benefits in statistical estimation and inference. In practice pre-screening serves as an appealing scheme for optimizing allocation and utilization of resources. To this end, wang2009forward Wang (2009) introduced forward regression for main effects screening and proves its screening consistency property. hao2014interaction Hao and Zhang (2014) further included second-order interactions into the linear model and proposes a two-step procedure for ultra-high dimensional variable screening. Meanwhile, to save resources and build an interpretable model with high prediction power, variable selection or screening must be employed. haris2016convex Haris et al. (2016) considered convex modelling of the factorial effects estimation and introduces strong heredity condition to achieve adaptive selection. hao2018model Hao et al. (2018) utilized a regularization scheme to tackle the curse of high dimensionality and perform valid variable screening with quadratic regression. Other works including lim2015learning, bien2013lasso Lim and Hastie (2015); Bien et al. (2013), proposed procedures for learning interactions based on ℓ_1 regularized least squares based on a purely algorithmic perspective without statistical guarantee.

May 2, 2022, Lei – COMMENT: Also include a discussion of winner's curse - Only in the "selecting the best" section.

1.3 Notations

We adopt the following notations throughout the manuscript. For asymptotic analyses, $a_N = O(b_N)$ denotes that there exists a positive constant $C > 0$ such that $a_N \leq Cb_N$. $a_N = o(b_N)$ denotes that $a_N/b_N \to 0$ as N goes to infinity. $a_N = \Theta(b_N)$ denotes that there exists positive constants c and C such that $cb_N \leq a_N \leq Cb_N$.

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

2 Factorial experiment setup

We consider a 2^K factorial experimental design for some $K \geq 2$, which encompasses K factors with binary levels indexed by $z_k \in \{0, 1\}, k = 1, \cdots, K$. Let $\boldsymbol{z}_{\mathcal{K}} = (z_k)_{k \in \mathcal{K}}$ index the combination of factors in $\mathcal{K} \subset [K]$, with $\boldsymbol{z}_{[K]}$ abbreviated to $\boldsymbol{z}$ specially. These factors define a collection of $Q = 2^K$ treatments, which we denote as $\mathcal{T} = \{\boldsymbol{z} = (z_1 \cdots z_K) \mid z_k \in \{0, 1\}, k = 1, \cdots, K\}$. We specially introduce the subsets $\mathcal{T}_{K_0}$ of $\mathcal{T}$, which contains the combinations with at most K_0 factors set as 1. N units are enrolled in the experiment, with $N(\boldsymbol{z})$ units in the group with the treatment $\boldsymbol{z}$. For simplicity, we can also index the treatments $\boldsymbol{z}$ in $\mathcal{T}$ using $1, \cdots, Q$ and write $\mathcal{T} = [Q]$. Unit i has potential outcome $Y_i(\boldsymbol{z})$ if assigned to treatment $\boldsymbol{z}$. We aggregate the potential outcomes into vectors $\boldsymbol{Y}_i = \{Y_i(\boldsymbol{z})\}_{\boldsymbol{z} \in \mathcal{T}}$ using lexicographic order.

Let $\overline{Y} = \{\overline{Y}(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}}$ be a vector defined as follows:

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

Let Z_i encode the treatment that the i-th unit received under a random permutation. More concretely, for given $N(\boldsymbol{z})$, we have

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

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

We define factorial effects for any subset $\mathcal{K}$ of the K factors following the discussion of

dasgupta2015causal,zhao2021regression, wu2011experiments

Dasgupta et al. (2015); Zhao and Ding (2021b); Wu and Hamada (2011). We introduce a set of vectors $\{g_{\mathcal{K}} \mid g_{\mathcal{K}} \in \mathbb{R}^Q\}$ defined in the following way: for $|\mathcal{K}| = |\{k\}| = 1$,

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

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

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

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

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

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

$$\tau = (\tau_{\mathcal{K}})_{\mathcal{K}\subset[K]} = Q^{-1}G^{\top}\overline{Y},\ G = (g_{\mathcal{K}})_{\mathcal{K}\subset[K]}. \tag{2.1}$$ eqn:def-fa

absorb Q^{-1} into G?

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

Following zhao2021regression Zhao and Ding (2021b), we consider factor-based regression. For the saturated regression, the regressor t_i is a vector indexed by combination of factors (equivalently, subsets of $[K]$). More precisely, t_i is constructed from $Z_i = (z_{i,k})_{k=1}^K$ such that

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

Then *the saturated regression* simply means regressing Y_i on the full t_i. More generally, we denote a collection of combination of factors by $\mathbb{M}$, $\mathbb{M} \subset \mathbb{K} = \{\mathcal{K} \mid \mathcal{K} \subset [K]\}$. In particular, we let $\mathbb{K}_k = \{\mathcal{K} \mid |\mathcal{K}| = k\}$ be the collection of indices corresponding to all the k-way interactions. We can partition $\mathbb{K}$ as

$$\mathbb{K} = \bigcup_{k=1}^{K} \mathbb{K}_k.$$

Let $\mathbb{M}_k^\star$ be the collection of the nonzero effects in $\mathbb{K}_k$. The combinations that correspond to the nonzero effects are aggregated into $\mathbb{M}^\star$, which is composed of K levels of indices:

$$\mathbb{M}^\star = \bigcup_{k=1}^{K} \mathbb{M}_k^\star.$$

For *the unsaturated regression*, we only use a sub-vector of t_i indexed by $\mathcal{K} \in \mathbb{M}$. Denote this sub-vector by $t_{i,\mathbb{M}}$, then the unsaturated regression over $\mathbb{M}$ translates to regressing Y_i on $t_{i,\mathbb{M}}$. Although we do not assume the data generating process is based on linear models, to align with the terminology in linear regression, we call a collection of indices $\mathbb{M}$ a *working model*.

Regarding the results, for regression over $\mathbb{M}$, we use $\widehat{\tau}(\mathbb{M})$ denote the coefficients. Moreover, the population version of the saturated and unsaturated regression are also of particular interest in our study. We use a non-hat counterpart $\tau(\mathbb{M})$ and τ to denote the population effects over model $\mathbb{M}$. Also use $G(\cdot, \mathbb{M})$ to denote the columns in G indexed by $\mathbb{M}$. The orthogonality of G implies the following fact:

Lemma 1. *The population averages $\overline{Y}$ can be represented by the factorial effects:*

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

eqn:repara

Remark 1. (*extension*) *The above construction of regressors t_i can be generalized by introducing a location-shift scheme as in* zhao2021regression *Zhao and Ding (2021b). For a given centering vector $\delta = (\delta_k)_{k=1}^K$, the t_i can be redefined as*

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

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

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

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

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

We observe (Y_i, Z_i) for the unit i, where $Z_i = (z_{i,1}, z_{i,2}, z_{i,3})$. Let $z'_{i,k} = 2z_{i,k} - 1$ be the centered version of $z_{i,k}$. For the purpose of regression, we construct an t_i from Z_i:

$$t_i = \left[1,\ z'_{i,1},\ z'_{i,2},\ z'_{i,3}, z'_{i,2} z'_{i,3},\ z'_{i,1} z'_{i,3}, z'_{i,1} z'_{i,2},\ z'_{i,1} z'_{i,2} z'_{i,3}\right].$$

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

$$Y_i \sim t_i.$$

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

$$Y_i \sim t_{i,\mathbb{M}}, \ \textit{where } t_{i,\mathbb{M}} = \Big[1,\ z'_{i,1},\ z'_{i,1}z'_{i,2},\ z'_{i,1}z'_{i,3}, z'_{i,1}z'_{i,2}z'_{i,3}\Big].$$

(i) Change the notation for t later! Maybe to W. (ii) Use $t(i,\cdot)$... for unification

3 Motivation

In randomized experiments, we are usually interested in estimating linear transformations of potential outcomes:

$$\gamma = \boldsymbol{f}^\top \overline{Y} \tag{3.3}$$

eqn:FTY

where $\boldsymbol{f}$ is a vector in $\in \mathbb{R}^Q$. When the number of treatment groups are small, estimation and inference for (eqn:FTY 3.3) have been well understood li2017general, zhao2021regression, zhao2021covariate (Li and Ding, 2017; Zhao and Ding, 2021b,a). One common condition imposed across literature is:

$$Q \text{ is fixed}, \ N \to \infty, \ N(\boldsymbol{z})/N \to e(\boldsymbol{z}) \in (0,1). \tag{3.4}$$

eqn:old-co

While Condition (eqn:old-cond 3.4) is able to guarantee satisfactory asymptotic results in small Q regimes, it is restrictive in practice and in more complicated statistical setups. Such concern is especially salient in factorial experiments, where the number of treatment groups is $Q = 2^K$. Therefore, Q can be large in general when the number of factors is large. Moreover, Condition (eqn:old-cond 3.4) requires that each arm contains a large collection of units, which might be violated in reality due to the limited resources and the difficulty of recruiting subjects. Mathematically, Condition (eqn:old-cond 3.4) breaks down even if we consider a slowly growing Q in asymptotic regimes. To see this, in a uniform design with N_0 units in each arm ($N(\boldsymbol{z}) = N_0$ for all $\boldsymbol{z} \in \mathcal{T}$), we have:

$$\frac{N(\boldsymbol{z})}{N} = \frac{N_0}{Q * N_0} = \frac{1}{Q} \to 0, \ \text{for } Q \to \infty \text{ with arbitrary rates.}$$

These issues call for new ideas of analyzing factorial experiments.

One promising idea to circumvent such drawbacks is to take advantage of the special structural information in factorial experiments. As wu2011experiments Wu and Hamada (2011) points out, there are several important principles in analyzing factorial experiments:

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

These principles suggest the possibility to exploit the hierarchy and sparsity structure in the factorial effects (eqn:def-factorial-effects 2.1) to overcome the barrier of limited replications. More specifically, we can perform an *effect screening* (or variable selection from a regression perspective) step that is tailored to the structure of factorial experiments and reduce the complexity of the model, then use the selected working model to analyze the target parameters.

We elaborate the above ideas in several interesting examples.

rt-effects **Example 2** (Inference on factorial effects with many factors)**.** zhao2021regression *Zhao and Ding (2021b) has discussed the statistical property of saturated and unsaturated linear regressions in finite population factorial experiment study:*

$$Y_i \sim t_{i,\mathbb{M}},$$

for a pre-specified working model $\mathbb{M} \subset [K]$.

Several questions remain unresolved. First, how should practitioners decide $\mathbb{M}$ *in real world scientific study? Second, if the number of factors* K *is allowed to grow, how will the story change? For these problems, effect screening is a natural consideration and directly serves for dimension reduction.*

-contrasts **Example 3** (Inference on general contrasts of potential outcomes)**.** *In many settings we are interested in testing general multiple linear transformations of average of potential outcomes:*

$$H_0 : F^\top \overline{Y} = \gamma_0. \tag{3.5}$$

eqn:H0

*Here $F \in \mathbb{R}^{H\times Q}$ is a general contrast matrix of interest. For example, F can a vector in $\mathbb{R}^Q$ with 1 indicating interested arms and 0 for the rest. Testing (*eqn:H0 *3.5) has been well studied in classical settings. If we can screen the effects, it is possible to exploit the shared information between treatment arms and reduce the complexity of the problem. In other words, incorporating the following information can potentially help with the testing:*

$$G(\cdot, \mathbb{M}^{\star c})^\top \overline{Y} = 0. \tag{3.6}$$

eqn:constr

*Again, $\mathbb{M}^\star$ is unknown to people without further exploration. Utilizing the information (*eqn:constraint *3.6) is also related to the restricted least squares in a multi-arm treatment experiment *zhao2021covariate *(Zhao and Ding, 2021a).*

elect-best **Example 4** (Select the best factor combinations)**.** *In many applications, we want to identify the "best" treatment group:*

$$\boldsymbol{z}_{\max} \in \mathop{\arg\max}_{\boldsymbol{z}\in\mathcal{T}'\subset\mathcal{T}} \overline{Y}(z_1\cdots z_K). \tag{3.7}$$

eqn:hat-bz

*For example, in field experiments regarding charitable giving, people wish to pick the most effective pricing policies *karlan2007does, wei2022inference *(Karlan and List, 2007; Wei et al., 2022). In financial portfolio management, managers want to learn about the best-performing strategies among many alternatives. Classical methods target each treatment arm separately and fail to incorporate the information that is shared across arms. Screening, on the other hand, provides one approach to utilize the shared information among arms. Suppose we can select the true working model (nonzero effects) $\mathbb{M}^\star$, we will have additional information to estimate $\boldsymbol{z}_{\max}$:*

$$\begin{gathered}\boldsymbol{z}_{\max} = \mathop{\arg\max}_{\boldsymbol{z}\in\mathcal{T}'\subset\mathcal{T}} \overline{Y}(z_1\cdots z_K), \\ s.t.\ G(\cdot, \mathbb{M}^{\star c})^\top \overline{Y} = 0.\end{gathered} \tag{3.8}$$

eqn:hat-bz

The constraints reflect the between-arm information which grants the problem a lower dimensionality than the original formulation.

After we present the main procedure and theoretical results in Section sec:forward-framework 4, we will come back and revisit these examples with more detailed discussion.

4 Forward screening in factorial experiments

-framework

4.1 Procedure

In this section we introduce the forward screening framework. The factorial effects have a natural hierarchical structure, which motivates a selection procedure that proceeds in a level-by-level style. Within each level, we apply some screening methods to select nonzero effects, such as marginal t tests with Bonferroni correction, lasso, etc. Between levels, we transition from lower order effects to a pre-selected working model for higher order effects following certain logic such as heredity principles. The forward selection procedure is summarized in Algorithm alg:forward-ms 1.

We briefly elaborate the high level intuition behind the forward screening procedure. The selected $\widehat{\mathbb{M}}$ can be partitioned as follows:

$$\widehat{\mathbb{M}} = \bigcup_{d=1}^{D} \widehat{\mathbb{M}}_d, \text{ where } \widehat{\mathbb{M}}_d = \widehat{\mathbb{M}} \cap \mathbb{K}_d.$$

Algorithm alg:forward-ms 1 introduces one operator to select models within each layer and another operator to advance the selected working model to a pre-model for the next layer. We call the within-level selection a "S-step", which is captured by a data-dependent operator $\widehat{\mathtt{S}} = \widehat{\mathtt{S}}(\cdot; \{Y_i, Z_i\}_{i=1}^N)$, and the between-level progress a "H-step", which is captured by a deterministic operator $\mathtt{H} = \mathtt{H}(\cdot)$. Concretely speaking, Step alg:BC-1 5 - Step alg:BC-3 7 gives a stochastic (or data dependent) operator $\widehat{\mathtt{S}}(\cdot) = \widehat{\mathtt{S}}(\cdot; \{Y_i, t_i\}_{i=1}^N)$ on any given model $\mathbb{M}$, while Step alg:step-prune 3 gives a deterministic operator $\mathtt{H}(\cdot)$ on a given model $\mathbb{M}$. The working models are updated in the following pattern:

$$\widehat{\mathbb{M}}_1 \xrightarrow{\mathtt{H}} \cdots \xrightarrow{\widehat{\mathtt{S}}} \widehat{\mathbb{M}}_{d-1} \xrightarrow{\mathtt{H}} \widehat{\mathbb{M}}_{d,+} \xrightarrow{\widehat{\mathtt{S}}} \widehat{\mathbb{M}}_d \to \cdots \xrightarrow{\widehat{\mathtt{S}}} \widehat{\mathbb{M}}_D. \quad (4.9)$$

eqn:track-

Both $\widehat{\mathtt{S}}(\cdot)$ and $\mathtt{H}(\cdot)$ can be general. In Algorithm alg:forward-ms 1, our S-step is based on marginal t tests and H-step is to proceed with one of the following heredity principles:

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

$$\mathcal{K}' \notin \widehat{\mathbb{M}}' \text{ for all } \mathcal{K}' \subset \mathcal{K},\ |\mathcal{K}'| = |\mathcal{K}| - 1. \quad (4.10)$$

eqn:weak-h

```
Algorithm 1: Forward screening under heredity
Input: Factorial data $(Y_i, Z_i)$; predetermined integer $D$; initial model for factorial
       effects $\widehat{\mathbb{M}} = \{\varnothing\}$; significance level $\{\alpha_d\}_{d=1}^{D}$.
Output: Selected working model $\widehat{\mathbb{M}}$.
Define an intermediate working model $\widehat{\mathbb{M}}' = \widehat{\mathbb{M}}$ for convenience.
for $d = 1, \cdots, D$ do
    Update intermediate working model to include all the $d$-order terms:
g:step-add-d-way
      $\widehat{\mathbb{M}}' = \widehat{\mathbb{M}} \cup \mathbb{K}_d$. screening Prune interactions according to the heredity principle
alg:step-prune
      eqn:weak-heredity eqn:strong-heredity
      (4.10) or (4.11) and still denote the pruned working model as $\widehat{\mathbb{M}}'$.
    Run weighted least squares on the working model $\widehat{\mathbb{M}}'$:
        $Y_i \sim t_{i,\widehat{\mathbb{M}}'}$, with weights $w_i = N/N_i$.
    Obtain coefficients $\widehat{\tau}(\widehat{\mathbb{M}}')$ and robust covariance estimation $\widehat{\boldsymbol{\Sigma}}(\widehat{\mathbb{M}}')$:
        $\widehat{\boldsymbol{\Sigma}}(\widehat{\mathbb{M}}') = \frac{1}{Q^2} G(\cdot, \widehat{\mathbb{M}}')^\top \mathbf{Diag}\left\{ N(\boldsymbol{z})^{-1} \widehat{S}(\boldsymbol{z}, \boldsymbol{z}) \right\} G(\cdot, \widehat{\mathbb{M}}')$.
alg:BC-1
alg:BC-2
    Extract $\widehat{\tau}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ and $\widehat{\sigma}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ for all $\mathcal{K} \in \widehat{\mathbb{M}}'$ with $|\mathcal{K}| = d$.
    Run marginal t-test using the above $\widehat{\tau}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ and $\widehat{\sigma}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ under significance level
alg:BC-3
      $\min\{\alpha_d/(|\widehat{\mathbb{M}}'| - |\widehat{\mathbb{M}}|), 1\}$ and remove the non-significant terms from $\widehat{\mathbb{M}}' \backslash \widehat{\mathbb{M}}$.
    Set $\widehat{\mathbb{M}} = \widehat{\mathbb{M}}'$.
lg:forward-ms
return $\widehat{\mathbb{M}}$
```

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

$$\mathcal{K}' \notin \widehat{\mathbb{M}}' \text{ for some } \mathcal{K}' \subset \mathcal{K},\ |\mathcal{K}'| = |\mathcal{K}| - 1. \tag{4.11}$$

eqn:strong

In practice, one can choose $\widehat{\mathtt{S}}(\cdot)$ and $\mathtt{H}(\cdot)$ based on the data structure as well as the domain knowledge.

4.2 Consistency of forward screening based on marginal t tests

Now we analyze the statistical property of Algorithm alg:forward-ms 1, which relies on an understanding of the property of the marginal t tests. The factorial effects are a special case of parameters that can be expressed as the linear combination of all the average potential outcomes:

$$\gamma = \sum_{\boldsymbol{z}\in\mathcal{T}} \boldsymbol{f}(\boldsymbol{z})\overline{Y}(\boldsymbol{z}). \tag{4.12}$$

eqn:target

(Change of notation here!) The moment estimator for (eqn:target-gamma 4.12) can be obtained by plugging in the sample averages $\widehat{Y}(\boldsymbol{z})$'s:

$$\widehat{\gamma} = \sum_{\boldsymbol{z}\in\mathcal{T}} \boldsymbol{f}(\boldsymbol{z})\widehat{Y}(\boldsymbol{z}), \tag{4.13}$$

eqn:hgamma

which also inspires the following variance estimation if $N(z) \geq 2$:

$$\widehat{v}^2 = \sum_{\boldsymbol{z}\in\mathcal{T}} \boldsymbol{f}(\boldsymbol{z})^2 N(\boldsymbol{z})^{-1}\widehat{S}(\boldsymbol{z},\boldsymbol{z}), \text{ where } \widehat{S}(\boldsymbol{z},\boldsymbol{z}) = \frac{1}{N(\boldsymbol{z})-1}\sum_{Z_i=\boldsymbol{z}}(Y_i - \widehat{Y}(\boldsymbol{z}))^2. \tag{4.14}$$

eqn:hv2R

It is known that li2017general (Li and Ding, 2017)

$$\mathbb{E}\{\widehat{Y}\} = \overline{Y},\ V_{\widehat{Y}} = \mathrm{Var}\left\{\widehat{Y}\right\} = \mathbf{Diag}\left\{N(\boldsymbol{z})^{-1}S(\boldsymbol{z},\boldsymbol{z})\right\} - N^{-1}S, \tag{4.15}$$

eqn:exp-va

where $S \in \mathbb{R}^{Q\times Q}$ is the covariance matrix for potential outcomes. Then (eqn:exp-var-hY 4.15) further leads to the following facts:

$$\mathbb{E}\{\widehat{\gamma}\} = \sum_{\boldsymbol{z}\in\mathcal{T}} \boldsymbol{f}(\boldsymbol{z})\overline{Y}(\boldsymbol{z}) = \gamma, \tag{4.16}$$

eqn:mean-h

$$\mathrm{Var}\left\{\widehat{\gamma}\right\} = \sum_{\boldsymbol{z}\in\mathcal{T}} \boldsymbol{f}(\boldsymbol{z})^2 N(\boldsymbol{z})^{-1}S(\boldsymbol{z},\boldsymbol{z}) - N^{-1}\boldsymbol{f}^\top S\boldsymbol{f}, \tag{4.17}$$

eqn:var-hg

$$\mathbb{E}\{\widehat{v}^2\} = \sum_{\boldsymbol{z}\in\mathcal{T}} \boldsymbol{f}(\boldsymbol{z})^2 N(\boldsymbol{z})^{-1}S(\boldsymbol{z},\boldsymbol{z}). \tag{4.18}$$

eqn:mean-v

The key part in our proof is to utilize non-asymptotic Berry-Esseen bounds to analyze the statistical property of (eqn:hgamma 4.13). In general the behavior of $\widehat{\gamma}$ relies on several ingredients of the outcome model, including uniformity of the design, magnitude of the true effects, etc. We start by introducing the following condition of *nearly uniform design*:

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

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

The classical assumption in literature (3.4) is a special case of Condition 1 when one takes Q to be a fixed integer.

Besides, we also need to quantify the order of the size of the true effects $\tau_{\mathcal{K}}$'s and the tuning parameters α_d's. We allow them to change with the total number of units N in certain rates:

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

- *True parameter:* $|\tau_{\mathcal{K}}| = \Theta(N^{\delta})$ *for some* $-1/2 < \delta \leq 0$ *and all* $\mathcal{K} \in \mathbb{M}^{\star}$*.*
- *Tuning parameter:* $\alpha_d = \Theta(N^{-\delta'})$ *for all* $d \in [D]$ *with some* $\delta' > 0$*.*

Condition 2 allows the order of the true factorial effects to decrease at the rate $\Theta(N^{\delta})$, for some $\delta > -1/2$, which is the boundary of statistical idenfiability. The tuning parameter α_d converges to zero, which ensures no marginal Type I error asymptotically. Wasserman and Roeder (2009) (Theorem 4.1 and 4.2) assumed similar conditions in high dimensional model selection.

The next two conditions are a set of regularity assumptions on the potential outcomes. Condition 3 requires the correlation matrix of $\widehat{Y}$ to be well-conditioned, and Condition 4 controls the moments of the potential outcomes. These two conditions generalize the classical assumptions in multi-arm randomization inference (Li and Ding, 2017).

Condition 3 (Nondegenerate correlation matrix)**.** *Let* $V^{\star}$ *be the correlation matrix of* $\widehat{Y}$*. There exists a* $\sigma > 0$*, such that*

$$\varrho_{\min}(V^{\star})/\varrho_{\max}(V^{\star}) \geq \sigma^2. \tag{4.19}$$

Condition 4 (Bounded fourth central moments)**.** *There exists a universal constant* $\Delta > 0$ *such that*

$$\max_{\boldsymbol{z}\in[Q]} \frac{1}{N}\sum_{i=1}^{N}\{Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\}^4 \leq \Delta^4. \tag{4.20}$$

Last but not least, we assume the following structural condition on the factorial effects:

Condition 5 (Hierarchical structure in factorial effects)**.** *The nonzero factorial effects have one of the following hierarchical structure:*

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

(Clarify the difference between this condition and (4.10) (4.11).)

Now we can derive the following theorem:

marginal-t **Theorem 1** (Bonferroni corrected marginal t test)**.** *Assume Conditions 1 - 5. Then the screening procedure based on Bonferroni corrected marginal t-test achieves perfect screening for the first D levels of effects asymptotically:*

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

(Comment on the literature?)

5 Inference under perfect screening

5.1 Construction of two estimators

In the motivation section, we discussed the possible benefits that effect screening can bring to estimation and inference in factorial experiments. In this section we formalize the ideas into rigorous theoretical results. Revisiting the target parameter (4.12), we have the moment estimator, which has been applied and studied widely in classical settings:

$$\widehat{\gamma} = \boldsymbol{f}^\top \widehat{Y}, \quad \widehat{v}^2 = \boldsymbol{f}^\top \widehat{V}_{\widehat{Y}} \boldsymbol{f}. \tag{5.21}$$ eqn:WLS-1

We can verify

$$\widehat{Y} = \arg\min_{\mu\in\mathbb{R}^Q} \|\widehat{Y} - \mu\|_2^2.$$

With the selected working model $\widehat{\mathbb{M}}$, it is also natural to consider a restricted least squares estimator:

$$\widehat{Y}_r = \arg\min_{\mu\in\mathbb{R}^Q} \|\widehat{Y} - \mu\|_2^2, \tag{5.22}$$ eqn:RLS-1

$$\text{s.t. } G(\cdot, \widehat{\mathbb{M}}^c)^\top \mu = 0. \tag{5.23}$$

The restricted least squares formulation (eqn:RLS-1 5.22) has a closed-form solution due to the orthogonality of G:

lem:RLS **Lemma 2.** *$\widehat{Y}_r$ from (eqn:RLS-1 5.22) can be expressed as:*

$$\widehat{Y}_r = Q^{-1}G(\cdot, \widehat{\mathbb{M}})G(\cdot, \widehat{\mathbb{M}})^\top \widehat{Y}. \tag{5.24}$$

eqn:RLS-2

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

Provide a proof. Define $f^\star$.

Lemma lem:RLS 2 suggests that under perfect screening, $\widehat{Y}_r$ is an alternative unbiased estimator for $\overline{Y}$. Let $\boldsymbol{f}[\widehat{\mathbb{M}}] = Q^{-1}G(\cdot, \widehat{\mathbb{M}})G(\cdot, \widehat{\mathbb{M}})^\top \boldsymbol{f}$. (eqn:RLS-2 5.24) motivates a point estimator $\widehat{\gamma}_r$ and a variance estimator $\widehat{v}_r^2$:

$$\widehat{\gamma}_r = \boldsymbol{f}^\top \widehat{Y}_r = \boldsymbol{f}[\widehat{\mathbb{M}}]^\top \widehat{Y}, \quad \widehat{v}_r^2 = \boldsymbol{f}[\widehat{\mathbb{M}}]^\top \widehat{V}_{\widehat{Y}} \boldsymbol{f}[\widehat{\mathbb{M}}]. \tag{5.25}$$

eqn:RLS-3

The question is to compare the inferential properties of (eqn:WLS-1 5.21) and (eqn:RLS-3 5.25).

5.2 Asymptotic normality under perfect screening

(A little dense here. How to improve it?)

For γ given by (eqn:target-gamma 4.12), one can reparameterize the parameter using decomposition (eqn:reparametrizat 2.2). If for some $\mathbb{M}$, $\tau(\mathbb{M}^c)$ is negligible, then we could reduce the dimension of the parameter space. Specifically, define the reparametrization-based quantities:

$$\widehat{\gamma}[\mathbb{M}] = \boldsymbol{f}^\top G(\cdot, \mathbb{M})\widehat{\tau}(\mathbb{M}), \ \gamma[\mathbb{M}] = \boldsymbol{f}^\top G(\cdot, \mathbb{M})\tau(\mathbb{M}),$$
$$v^2[\mathbb{M}] = \boldsymbol{f}^\top G(\cdot, \mathbb{M})\mathrm{Var}\left\{\widehat{\tau}(\mathbb{M})\right\}G(\cdot, \mathbb{M})^\top \boldsymbol{f}.$$

For simplicity, introduce the notation

$$\boldsymbol{f}[\mathbb{M}] = Q^{-1}G(\cdot, \mathbb{M})G(\cdot, \mathbb{M})^\top \boldsymbol{f}, \quad \boldsymbol{f}^\star = \boldsymbol{f}[\mathbb{M}^\star]. \tag{5.26}$$

eqn:bsf-M

$\boldsymbol{f}^\star$ captures the inherent information of $\boldsymbol{f}$ in defining the target estimand because the following fact holds:

$$(\boldsymbol{f}^\star)^\top \overline{Y} = \boldsymbol{f}^\top \overline{Y}, \text{ for all } \boldsymbol{f} \in \mathbb{R}^Q.$$

(State this as a lemma...)

With perfect screening, we have the following result:

perfect-ms **Theorem 2** (Asymptotic normality under under perfect selection)**.** *Let $\boldsymbol{f}^\star$ be given by (*eqn:bsf-M *5.26). Assume Condition* cond:nondegenerate-corr *3. If*

$$\mathbb{P}\left\{\widehat{\mathbb{M}}=\mathbb{M}^\star\right\}\longrightarrow 1 \quad \textit{and} \quad \frac{\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}\cdot\frac{\|\boldsymbol{f}^\star\|_\infty}{\|\boldsymbol{f}^\star\|_2}\longrightarrow 0, \tag{5.27}$$ eqn:asp-co

then

$$\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma[\mathbb{M}^\star]}{v[\mathbb{M}^\star]}\rightsquigarrow\mathcal{N}(0,1).$$

We briefly comment on the condition (eqn:asp-condition 5.27). The first part of (eqn:asp-condition 5.27) assumes perfect model selection, which can be justified for many procedures under certain conditions (Theorem thm:marginal-t 1). The second part of (eqn:asp-condition 5.27) is a sufficient condition for CLT that reflects the interplay among the potential outcomes, arm size and the reparametrized contrast vector $\boldsymbol{f}^\star$. In general the ratio $\|\boldsymbol{f}^\star\|_\infty/\|\boldsymbol{f}^\star\|_2$ is bounded by

$$\frac{\|\boldsymbol{f}^\star\|_\infty}{\|\boldsymbol{f}^\star\|_2}\le\min\left\{\frac{|\mathbb{M}^\star|\|\boldsymbol{f}^\top G(\cdot,\mathbb{M}^\star)\|_\infty}{\sqrt{Q}\|\boldsymbol{f}^\top G(\cdot,\mathbb{M}^\star)\|_2},1\right\}. \tag{5.28}$$ eqn:ratio-

When Q is fixed or $|\mathbb{M}^\star|$ is large, the upper bound (eqn:ratio-f 5.28) gives constant order control. Therefore we require $N_0\to\infty$ to obtain a useful CLT. When the true working model is sparse, i.e., $|\mathbb{M}^\star|$ is much smaller than Q, the ratio (eqn:ratio-f 5.28) has a vanishing order even N_0 is upper bounded by some constant.

(Add discussions on benefits of perfect screening; add discussions on conditions for $\widehat{\gamma}$ without model selection for comparison.)

To further elaborate the benefits of perfect screening for CLT, we make some simple comparison in Example exp:sparse-bw 5 based on one concrete choice of $\boldsymbol{f}$. For simplicity we assume the potential outcomes are upper bounded and $\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})$ is lower bounded by some universal constants.

:sparse-bw **Example 5.** *Let $\boldsymbol{f}=(1,0,\dots,0)^\top$. Then we can compute*

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

Applying Theorem thm:be-perfect-ms *2, we can formulate the conditions for CLT in Table* tab:sparse-bw *1.*

When Q is large and $\mathbb{M}^\star$ is sparse, CLT with screening holds under weaker conditions and the robust variance is smaller in expectation. Intuitively speaking, although the coefficient $\boldsymbol{f}$ is only relevant to treatment arm $\boldsymbol{z}=1$, by screening we can actually incorporate information from other arms to establish CLT and variance estimator under relaxed conditions.

Table 1: Comparison of conditions for CLT and limit of variance estimate for sparse $\boldsymbol{f}$

:sparse-bw

Perfect selection	CLT	$\mathbb{E}\left\{\widehat{v}_R^2\right\}$
Yes	$\mathbb{P}\left\{\widehat{\mathbb{M}} = \mathbb{M}^\star\right\} \to 1$ and $\frac{\lvert\mathbb{M}^\star\rvert}{QN_0} \to 0$	$\sum_{\boldsymbol{z}=1}^{Q} \frac{\boldsymbol{f}^\star(\boldsymbol{z})^2 S(\boldsymbol{z},\boldsymbol{z})}{N(\boldsymbol{z})} \le \frac{\lvert\mathbb{M}^\star\rvert}{Q} \max_{\boldsymbol{z}\in[Q]}\left\{\frac{S(\boldsymbol{z},\boldsymbol{z})}{N(\boldsymbol{z})}\right\}$
No	$\frac{1}{N(1)} \to 0$	$\frac{S(1,1)}{N(1)}$

Theorem thm:be-perfect-ms 2 provides direct solutions for the motivating examples we introduced earlier. For reporting factorial effects (Example exp:report-effects 2), we can report the effects in the selected working model $\widehat{\mathbb{M}}$. If the screening step is correct with high probability, we can obtain consistent estimates and apply Theorem thm:be-perfect-ms 2 to establish CLTs. For Example exp:general-contrasts 3, to do inference on general contrasts $F^\top\overline{Y}$, we can build reparametrization-based estimators for $\overline{Y}$ using the selected working model $\widehat{\mathbb{M}}$:

$$\widehat{Y}_r = Q^{-1}G(\cdot,\widehat{\mathbb{M}})G(\cdot,\widehat{\mathbb{M}})^\top\widehat{Y}, \tag{5.29}$$

eqn:hY-rp

then deduce the asymptotic results for $F^\top\widehat{Y}_r$. For Example exp:select-best 4, the story is slightly more complex and we will add detailed discussion in Section sec:select-best 5.3.

Even if we do not have enough confidence for perfect screening over high order interactions, we can still focus on an unsaturated regression based on a working model that contains the lower order interactions, which are more important in general. Section sec:imperfect-ms 6 has more discussion on imperfect screening.

5.3 Application: select the best factor combinations

elect-best

We study Example exp:select-best 4 in this section. Our goal is to identify the treatment arms that demonstrates the "best" performance measured by level of average potential outcome and do inference on the average potential outcome for the selected arms. For ease of presentation, we focus on selecting the maximal potential outcome in this section and defer discussions on general ordered average potential outcomes to the appendix.

Denote the maximal average of potential outcome as $\overline{Y}_{(1)} = \max_{\boldsymbol{z}\in\mathcal{T}'}\overline{Y}(\boldsymbol{z})$. (Explain $\mathcal{T}'$). The arms that achieved maximum might not be unique. this is the key difference -

otherwise trivial. Besides, arms with similar values might not be statistically distinguishable with limited sample size. We introduce the following notion $\mathcal{T}_1$ to include all arms that achieve or are close to $\overline{Y}_{(1)}$:

$$\mathcal{T}_1 = \left\{ \boldsymbol{z} \in \mathcal{T}' \subset \mathcal{T} \mid |\overline{Y}(\boldsymbol{z}) - \overline{Y}_{(1)}| = \Theta(N^{-\delta_3}) \right\}, \text{ for some } \delta_3 > 0. \tag{5.30}$$

eqn:near-t

Here $\mathcal{T}'$ is a subset of $\mathcal{T}$. In practice $\mathcal{T}'$ incorporate people's decision on which arms are interesting for comparison. As a naive example, when the number of arms is not very large, one can simply take $\mathcal{T}' = \mathcal{T}$. As a less trivial example, suppose the K factors in the study represent strategies that people can take. Due to resource constraint, at most K_0 factors can be set as 1, then

$$\mathcal{T}' = \left\{ \boldsymbol{z} \in [Q] \mid \sum_{k=1}^{K} z_k \leq K_0 \right\}.$$

With the above basic setup, we introduce the following Algorithm alg:select-tie 2 for selecting the most effective arm(s). give intuition

We provide some theoretical analysis of Algorithm alg:select-tie 2 from a theoretical perspective. We define

$$d_h = \max_{\boldsymbol{z} \in \mathcal{T}_1} |\overline{Y}(\boldsymbol{z}) - \overline{Y}_{(1)}|, \; d_h^\star = \min_{\boldsymbol{z} \notin \mathcal{T}_1} |\overline{Y}(\boldsymbol{z}) - \overline{Y}_{(1)}|.$$

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

For theoretical interests, we quantify the order of the involved quantities:

d:distance **Condition 6** (Order of d_h and $d_h^\star$)**.** *Assume the following scaling of parameters:*

$$d_h^\star = \Theta(N^{\delta_1}), \eta_{L,N} = \Theta(\eta_N), \eta_{R,N} = \Theta(\eta_N)$$

with $\eta_N = \Theta(N^{\delta_2}), d_h = \Theta(N^{\delta_3})$ *with* $\delta_3 \leq -1/2 < \delta_2 < \delta_1 \leq 0.$

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

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

nfer-order **Theorem 3** (Asymptotic results on the estimated effects)**.** *Assume Condition* con[illegible] *3, 4 and 6. Assume* $|\mathbb{M}^\star| = \Theta(N^{\delta_4})$ *for some* $\delta_4 \geq 0$ *and*

$$\mathbb{P}\left\{ \widehat{\mathbb{M}} = \mathbb{M}^\star \right\} \to 1, \; N^{-(1+2\delta_2-\delta_4)} \to 0, \; |\mathcal{T}'| \cdot |\mathcal{T}_1| N^{-\frac{1-\delta_4}{2}} \to 0.$$

```
Algorithm 2: Select the most effective arms
  Input: Factorial data $(Y_i, Z_i)$; predetermined integer $D$; initial model for factorial
         effects $\widehat{\mathbb{M}} = \{\varnothing\}$; significance level $\{\alpha_d\}_{d=1}^{D}$.
  Output: Selected working model $\widehat{\mathbb{M}}$.
  Perform effects screening with Algorithm alg:forward-ms 1 and obtain working model $\widehat{\mathbb{M}}$.
  Obtain reparametrization-based estimates:
      $\widehat{Y}_{\mathrm{r}}(\cdot) = G(\cdot, \widehat{\mathbb{M}})\widehat{\tau}(\widehat{\mathbb{M}}).$
  Select the best factor level combinations.
      $\widehat{\mathcal{T}}_1 = \left\{ \boldsymbol{z} \in \mathcal{T}' \mid -\eta_{L,N} \le \widehat{Y}_{\mathrm{r}}(\boldsymbol{z}) - \max_{\boldsymbol{z}' \in \mathcal{T}'} \widehat{Y}_{\mathrm{r}}(\boldsymbol{z}') \le \eta_{R,N} \right\}.$
    Here $\eta_{L,N}, \eta_{R,N} > 0$ are some tuning parameters.
  Define
      $\boldsymbol{f}_{(1)} = (Q|\widehat{\mathcal{T}}_1|)^{-1} \sum_{\boldsymbol{z} \in \widehat{\mathcal{T}}_1} G(\cdot, \widehat{\mathbb{M}}) G(\boldsymbol{z}, \widehat{\mathbb{M}})^\top.$
    Then generate point estimates and variance estimator for the effect size over the
    selected tie:
      $\widehat{Y}_{(1)} = \frac{1}{|\widehat{\mathcal{T}}_1|} \sum_{\boldsymbol{z} \in \widehat{\mathcal{T}}_1} \widehat{Y}_{\mathrm{r}}(\boldsymbol{z}) = \boldsymbol{f}_{(1)}^\top \widehat{Y},$
      $\widehat{v}_{(1)} = \boldsymbol{f}_{(1)}^\top \widehat{V}_Y \boldsymbol{f}_{(1)}.$
lg:select-tie
  return $\widehat{\mathcal{T}}_1, \widehat{Y}_{(1)}, \widehat{v}_{(1)}$
```

Then the point estimates are asymptotically jointly normal:

$$\frac{\widehat{Y}_{(1)} - \overline{Y}_{(1)}}{(\boldsymbol{f}_{(1)}^\top V_Y \boldsymbol{f}_{(1)})^{1/2}} \rightsquigarrow \mathcal{N}(0,1),$$

Moreover, $\boldsymbol{f}_{(1)}^\top \widehat{V}_Y \boldsymbol{f}_{(1)}$ is a robust variance estimator for $\boldsymbol{f}_{(1)}^\top V_Y \boldsymbol{f}_{(1)}$ which satisfies

$$N(\boldsymbol{f}_{(1)}^\top \widehat{V}_Y \boldsymbol{f}_{(1)} - \boldsymbol{f}_{(1)}^\top \mathbb{E}\{\widehat{V}_Y\} \boldsymbol{f}_{(1)}) \xrightarrow{\mathbb{P}} 0, \ \ \boldsymbol{f}_{(1)}^\top \mathbb{E}\{\widehat{V}_Y\} \boldsymbol{f}_{(1)} \succeq \boldsymbol{f}_{(1)}^\top V_Y \boldsymbol{f}_{(1)}.$$

Theorem thm:infer-order 3 implies some benefits of reparametrization:

- Sufficient conditions for CLT. When the size of the true working model is small (say $\delta_4 = 0$) and screening is consistent, one only needs

  $$|\mathcal{T}'||\mathcal{T}_1|\left(\frac{\mathbb{M}^\star}{N}\right)^{1/2} \to 0. \text{ (because } N^{-(1+2\delta_2)} \text{ always converge to 0 as } N \to \infty\text{)}$$

  The condition relies on the scaling of N instead of a particular set of $N(\boldsymbol{z})$'s, meaning that we can incorporate information from other treatment arms.

- Constructing confidence intervals. The length of confidence intervals with reparametrization is given by $\frac{1}{|\mathcal{T}_1|}\sum_{\boldsymbol{z}\in\mathcal{T}_1} N(\boldsymbol{z})^{-1}S(\boldsymbol{z},\boldsymbol{z})$. We have upper bound

  $$\boldsymbol{f}_{(1)}^\top \mathbb{E}\{\widehat{V}_Y\}\boldsymbol{f}_{(1)} = \sum_{\boldsymbol{z}\in\mathcal{T}}[\boldsymbol{f}_{(1)}(\boldsymbol{z})]^2 N(\boldsymbol{z})^{-1}S(\boldsymbol{z},\boldsymbol{z}) \le \frac{|\mathbb{M}^\star|}{Q}\max_{\boldsymbol{z}\in\mathcal{T}}\left\{\frac{S(\boldsymbol{z},\boldsymbol{z})}{N_{\boldsymbol{z}}}\right\}.$$

  When Q (or equivalently K) is large and the working model is sparse, clearly the confidence intervals based on reparametrization has more advantage.

6 Imperfect screening

perfect-ms

Perfect screening might be an overly optimistic pursuit, which is subject to complexity of the parameter structure and the subtlety of tuning. The story of general non-perfect selection is complicated due to several reasons:

- The estimators built from a selected working model have a complex probability structure because the randomness in the selection step and the inference step is entangled.

- The validity of commonly used post-selection inference strategies, such as data splitting, simultaneous inference, selective inference, among others, has not been fully understood for complete randomization. Moreover, each of these strategies has its only conceptual or practical difficulty. For example, data splitting typically leads to loss of efficiency and high variability due to the randomness of splitting. Simultaneous inference suffers from conservativeness and computational issues. Selective inference relies heavily on the specific selection methodology used prior to inference. For a critical discussion of post-selection inference, see kuchibhotla2022post Kuchibhotla et al. (2022).

- The performance of the forward selection algorithm varies with the specific screening scheme applied in each layer.

We want to explore useful and safe strategies under reasonable conditions when perfect selection fails. Our proposals are summarized in Figure 1.

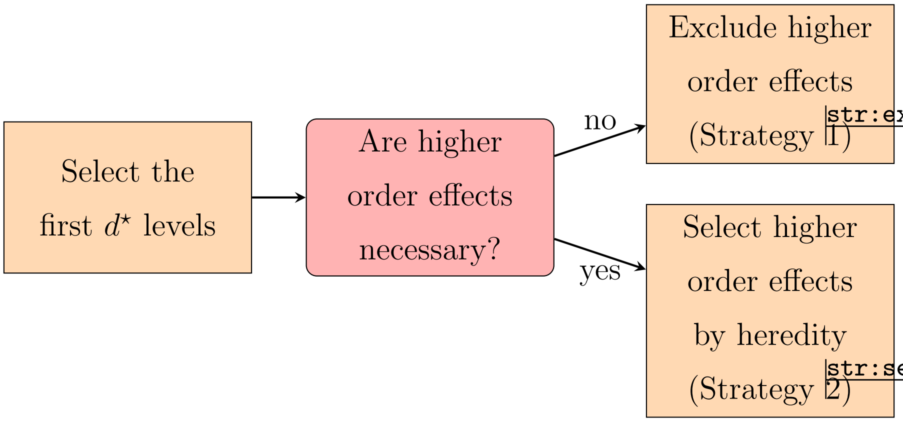
Select the first $d^\star$ levels
Are higher order effects necessary?
no
yes
Exclude higher order effects (Strategy 1)
Select higher order effects by heredity (Strategy 2)

Figure 1: General strategy for factorial screening

When using Algorithm 1, we might believe that perfect selection is more plausible in the initial levels (say the main effects plus two way interactions). However, it can be problematic in the higher order interactions if the higher order interactions are too small in magnitude or the parameter α_d at a particular layer is not well-tuned. Therefore, it is reasonable to start from the situation where we are only assured perfect selection for the first several levels:

Condition 7. *The selected working model is correct up to the level $d^\star \le D^\star$ with probability tending to 1:*

$$\mathbb{P}\left\{\widehat{\mathbb{M}}_1 = \mathbb{M}_1^\star, \ldots, \widehat{\mathbb{M}}_{d^\star} = \mathbb{M}_{d^\star}^\star\right\} \to 1.$$

Condition 7 does not impose any restriction on the selection results beyond level $d^\star$. For example, in practice we can set $d^\star = 2$ if we do not have confidence in the selection results beyond main effects and second-order interactions. This motivates the first strategy without perfect model selection:

Strategy 1 (Exclude higher order interactions)**.** *In Algorithm 1, set $\alpha_d = \infty, d \ge d^\star + 1$ so that no effects beyond level $d^\star$ will be selected and $\widehat{\mathbb{M}} = \cup_{d=1}^{d^\star} \widehat{\mathbb{M}}_d$.*

Strategy str:exclude-all 1 leads to under-selection: $\widehat{\mathbb{M}} \subset \mathbb{M}^\star$. In this case, unbiasedness usually fails for estimating general linear combination of average potential outcomes:

$$\begin{aligned}\boldsymbol{f}^\top \overline{Y} =& \boldsymbol{f}^\top G(\cdot, \mathbb{M}^\star)\tau(\mathbb{M}^\star) \\ =& \sum_{d=1}^{d^\star} \boldsymbol{f}^\top G(\cdot, \mathbb{M}^\star_d)\tau(\mathbb{M}^\star_d) + \sum_{d=d^\star+1}^{D^\star} \boldsymbol{f}^\top G(\cdot, \mathbb{M}^\star_d)\tau(\mathbb{M}^\star_d). \qquad (6.31)\end{aligned}$$

eqn:decomp

Because the selected working model might deviate from the truth beyond level $d^\star$, we do not have a consistent estimation for the second part of (eqn:decomposition 6.31). Therefore, we will need to focus on coefficient vectors $\boldsymbol{f}$ that satisfy the some orthogonality conditions as introduced in Theorem thm:strategy-I 4 below:

strategy-I

Theorem 4 (Guarantee for Strategy str:exclude-all 1). *Assume Conditions* cond:nondegenerate *3 and* cond:under-selection *7. Also assume* $\boldsymbol{f}$ *satisfies the following orthogonality condition:*

$$G(\cdot, \mathbb{M}^\star_d)^\top \boldsymbol{f} = 0, \; d^\star + 1 \leq d \leq D^\star. \qquad (6.32)$$

eqn:orthog

Let $\boldsymbol{f}[\mathbb{M}^\star_{1:d^\star}]$ *be given by (*eqn:bsf-M *5.26) with* $\mathbb{M}^\star_{1:d^\star} = \cup_{d=1}^{d^\star} \mathbb{M}^\star_d$. *If*

$$\frac{\max_{i\in[N], \boldsymbol{z}\in[Q]} |Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\sqrt{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z}, \boldsymbol{z})} \cdot \sqrt{N_0}} \cdot \frac{\|\boldsymbol{f}[\mathbb{M}^\star_{1:d^\star}]\|_\infty}{\|\boldsymbol{f}[\mathbb{M}^\star_{1:d^\star}]\|_2} \longrightarrow 0,$$

then

$$\frac{\widehat{\gamma}[\widehat{\mathbb{M}}] - \gamma}{v[\mathbb{M}^\star_{1:d^\star}]} \rightsquigarrow \mathcal{N}(0,1).$$

(eqn:orthogonality 6.32) states that the coefficient vector $\boldsymbol{f}$ should be orthogonal to the higher order contrasts, which makes sense because the under selection procedure cannot provide screening guarantee for high order interactions. One straightforward example satisfying (eqn:orthogonality 6.32) is linear combination of lower-order contrasts, given by

$$\boldsymbol{f} = G(, \cup_{d=1}^{d^\star} \mathbb{M}^\star_d)\boldsymbol{b}.$$

For any $d_0 \geq d^\star + 1$, we can calculate

$$G(\cdot, \mathbb{M}^\star_{d_0})^\top \boldsymbol{f} = G(\cdot, \mathbb{M}^\star_{d_0})^\top G(, \cup_{d=1}^{d^\star} \mathbb{M}^\star_d)\boldsymbol{b} = 0,$$

which holds due to the orthogonality between low order and higher order contrasts.

On the other hand, in certain scenarios people do care about higher order interactions, for which the screening procedure might not work perfectly. To this end, we propose to select the effects by hierarchy which takes higher order effects into consideration and guarantees high interpretability over the selected model.

y-heredity **Strategy 2** (Select higher order interactions by heredity)**.** *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*

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

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

We have the following results for Strategy 2:

trategy-II **Theorem 5** (Guarantee for Strategy 2)**.** *Assume Conditions 3, 5 and 7. Let*

$$\mathbb{M}^{\star\star} = \bigcup_{d=1}^{D} \mathbb{M}_d^{\star\star},$$

where

$$\mathbb{M}_d^{\star\star} = \begin{cases} \mathbb{M}_d^\star, & d \leq d^\star; \\ \mathtt{H}^{(d-d^\star)}(\mathbb{M}_{d^\star}^\star), & d^\star + 1 \leq d \leq D. \end{cases}$$

Then

$$\mathbb{M}^\star \subset \mathbb{M}^{\star\star}, \; \mathbb{P}\left\{\widehat{\mathbb{M}} = \mathbb{M}^{\star\star}\right\} \to 1.$$

Let $\boldsymbol{f}[\mathbb{M}^{\star\star}]$ *be given by (5.26) with* $\mathbb{M} = \mathbb{M}^{\star\star}$*. If*

$$\frac{\max_{i\in[N],\boldsymbol{z}\in[Q]} |Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\sqrt{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})} \cdot \sqrt{N_0}} \cdot \frac{\|\boldsymbol{f}[\mathbb{M}^{\star\star}]\|_\infty}{\|\boldsymbol{f}[\mathbb{M}^{\star\star}]\|_2} \longrightarrow 0,$$

then

$$\frac{\widehat{\gamma}[\widehat{\mathbb{M}}] - \gamma}{v[\mathbb{M}^{\star\star}]} \rightsquigarrow \mathcal{N}(0,1).$$

Strategy str:select-by-heredity 2 is an over-selection procedure because $\mathbb{M}^\star \subset \widehat{\mathbb{M}}$ with high probability. For one thing, Strategy str:select-by-heredity 2 can guarantee the inclusion of the true model. For another thing, Strategy str:select-by-heredity 2 preserves the heredity structure in the effects so that the selected working model has high interpretability in practice.

When analyzing Strategy str:exclude-all 1 and str:select-by-heredity 2, the selected working model recovers a fixed model with high probability. Both strategies have advantages and disadvantages. Under-selection reflects bias-variance trade-off: it can induce more bias for certain target parameters, but the constructed estimator typically enjoys smaller variance. Over-selection typically reduces bias for estimation but suffers from loss of efficiency compared with perfect screening because one might include too many redundant terms into the selected model. In general, if higher order interactions are not crucial for study, Strategy str:select-by-heredity 2 should be applied. If high order interactions are of interest and hard to select, one could pursue Strategy str:exclude-all 1 as a practically useful and interpretable solution.

7 Simulation

8 More discussions

8.1 Discussion of several screening methods in finite population factorial experiments

- Bonferroni correction:
- LASSO: soft thresholding!

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

- AIC: hard thresholding!

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

- BIC: hard thresholding!

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

8.2 Discussion: more general centering values

8.2.1 Unsaturated weighted least square: a closed form expression

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

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

$$\Big(f_{\delta'_{[K]}}(z^*_{\mathcal{K}})\Big)_{\mathcal{K} \subset [K]} = \Big(f_{\delta_{[K]}}(z^*_{\mathcal{K}})\Big)_{\mathcal{K} \subset [K]} \boldsymbol{P}_{\Delta\delta_{[K]}}. \tag{8.33}$$

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

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

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

$$\boldsymbol{P}_{\Delta\delta_{[K]}}(\mathcal{K}, \mathcal{K}') = \begin{cases} \prod_{k \in \mathcal{K}' \backslash \mathcal{K}} (\Delta\delta)_k & , \ \mathcal{K} \subset \mathcal{K}', \\ 0 & , \ \mathcal{K} \subsetneq \mathcal{K}'. \end{cases} \tag{8.34}$$

eqn:Pmatri

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

$$\boldsymbol{Q}_{\Delta\delta_{[K]}}(\mathcal{K}, \mathcal{K}') = \begin{cases} (-1)^{|\mathcal{K}'| - |\mathcal{K}|} \prod_{k \in \mathcal{K}' \backslash \mathcal{K}} (\Delta\delta)_k & , \ \mathcal{K} \subset \mathcal{K}', \\ 0 & , \ \mathcal{K} \subsetneq \mathcal{K}'. \end{cases} \tag{8.35}$$

eqn:Qmatri

We shall give an example of the above matrix in the three-factor case, which appears (incompletely) in the appendix of zhao2021regression Zhao and Ding (2021b). Let $A' = A - \delta_A$, $B' = B - \delta_B$,

$C' = C - \delta_C$.

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

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

thm:closeD **Theorem 6.** *Consider weighted least squares with centering factors $\delta_{[K]}$ and weights proportional to size of each stratum. Let $\Delta\delta_{[K]} = \delta_{[K]} - (1/2)_{k=1}^{K}$. The unsaturated regression on up to all m-level main/interactions terms has coefficient vector:*

$$(\widetilde{\tau}_{\mathcal{K}})_{\{|\mathcal{K}|\le m\}} = (\tau_{\mathcal{K}})_{\{|\mathcal{K}|\le m\}} + \boldsymbol{D}_h \cdot (\tau_{\mathcal{K}})_{\{|\mathcal{K}|>m\}}, \tag{8.36}$$ eqn:wls-re

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

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

cor:closeD **Corollary 1.** *The matrix $\boldsymbol{D}$ has a closed form expression:*

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

$$\boldsymbol{D}_h(\mathcal{K},\mathcal{K}') = 0. \tag{8.37}$$

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

$$\boldsymbol{D}_h(\mathcal{K},\mathcal{K}') = \sum_{l=0}^{m-k} (-1)^{k'-k+1-l} \binom{k'-k+1}{l} \prod_{t\in\mathcal{K}'\backslash\mathcal{K}} \left(\delta_t - \frac{1}{2}\right). \tag{8.38}$$

Proof. This result can be derived through careful calculation based on the definition of $\boldsymbol{P}$ and $\boldsymbol{Q}$ from (eqn:Pmatrix 8.34) and (eqn:Qmatrix 8.35) along with Theorem thm:closeD 6 thus omitted here. □

8.2.2 A sufficient condition for sign consistency in population WLS regression

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

This definition is mainly devoted to finer technical purposes in Theorem thm:suffcond 7.

m:suffcond **Theorem 7.** *Assume we run weighted least square under the setting depicted in Theorem thm:closeD 6. Define the maximal decaying rate $c_K = \max_{l\in[K]} c_l$. Recall the predefined maximal active interaction number s_K from Definition def:sparsity 1. If we have*

$$s_K c_K \max_{k=1,\dots,K} |\delta_k - 1/2| < \ln 2, \tag{8.39}$$ cond:suffi

then the unsaturated regression coefficients $(\widetilde{\tau}_{\mathcal{K}})_{\{|\mathcal{K}|\le m\}}$ and the corresponding saturated regression coefficients $(\tau_{\mathcal{K}})_{\{|\mathcal{K}|\le m\}}$ from (eqn:wls-res 8.36) have same signs on every term.

Condition (cond:sufficient 8.39) unifies the property of factorial effects and the information of the design pattern(the centering factors $\delta_{[K]}$). The product of s_K and c_K demonstrates a trade-off between the active interaction number and the hierarchy structure. Sparser interactions require slower decaying rate and vice versa. Besides, the product of $s_K c_K$ and $\max_{k=1,\dots,K} |\delta_k - 1/2|$ shows that if δ_k lies more close to 1/2, less restriction are needed on the effect structure. This aligns with the result in zhao2021regression Zhao and Ding (2021b): when $\delta_k = 1/2$ holds for all $k = 1,\dots,K$, $\boldsymbol{D}_h = \boldsymbol{0}$, so that forward selection always works.

9 Conclusion

References

9inference Andrews, I., Kitagawa, T., and McCloskey, A. (2019). Inference on winners. Technical report, National Bureau of Economic Research.

n2013lasso Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of statistics*, 41(3):1111.

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

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

o2018model Hao, N., Feng, Y., and Zhang, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 113(522):615–625.

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

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

an2007does Karlan, D. and List, J. A. (2007). Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–1793.

la2022post Kuchibhotla, A. K., Kolassa, J. E., and Kuffner, T. A. (2022). Post-selection inference. *Annual Review of Statistics and Its Application*, 9:505–527.

2018winner Lee, M. R. and Shen, M. (2018). Winner's curse: Bias estimation for total effects of features in online controlled experiments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 491–499.

017general Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769.

15learning Lim, M. and Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654.

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

009forward Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524.

an2009high Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37(5A):2178.

2inference Wei, W., Zhou, Y., Zheng, Z., and Wang, J. (2022). Inference on the best policies with many covariates. *arXiv preprint arXiv:2206.11868*.

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

1covariate Zhao, A. and Ding, P. (2021a). Covariate-adjusted fisher randomization tests for the average treatment effect. *Journal of Econometrics*, 225(2):278–294.

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

A General results on consistency of forward screening

We are now ready to show the guarantee for our procedure. Our presentation starts from a theorem (Theorem thm:ms-consistency 8) quantifying the performance of the forward procedure (eqn:track-model 4.9) under general assumptions on the S-step and P-step. Then we derive a corollary (Theorem thm:marginal-t 1) by specifying the S-step as Bonferroni corrected marginal t tests and the P-step as screening based on heredity. The high level of the proof proceeds through mathematical induction:

1. (Base case) Show that forward screening selects the correct main effects with probability tending to one.

2. (Induction step) Show that if we correctly screen the effects up to k-way interactions (main effects if $k = 1$), then we are able to correctly detect the non-nulls among all $(k + 1)$-way interactions.

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

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

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

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

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

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

$$\textit{Validity:}\ \limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d \cap \mathbb{M}^{\star c}_d \neq \varnothing\right\} \leq \alpha_d, \tag{1.40}$$

eqn:validi

$$\textit{Consistency:}\ \limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}^c_d \cap \mathbb{M}^{\star}_d \neq \varnothing\right\} = 0. \tag{1.41}$$

eqn:consis

This assumption can be verified for many screening procedures. In Theorem thm:marginal-t 1 we will show it holds for the layer-wise Bonferroni corrected marginal testing procedure in Algorithm alg:forward-ms 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 wasserman2009high (Wasserman and Roeder, 2009).

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

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

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

One special case of $\mathtt{H}(\cdot)$ operator satisfying Assumption asp:H-heredity 2 is naively adding all the the higher order interactions regardless of the lower-order screening results. Besides, if we have evidence that the effects have particular hierarchical structure, applying the corresponding heredity principle such as (4.10) or (4.11) can improve screening accuracy as well as interpretability of the screening results.

Theorem 8 (Screening consistency)**.** *Assume $\mathbb{M}^\star \neq \varnothing$. Assume Assumption 1 and 2. Then the forward screening procedure (4.9) has the following properties:*

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

$$\limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_d \cap \mathbb{M}_d^{\star c} \neq \varnothing \textit{ for some } d \in [D]\right) \leq \alpha = \sum_{d=1}^{D} \alpha_d. \tag{1.42}$$

(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:*

$$\limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star \textit{ for all } d \in [D]\right) = 1. \tag{1.43}$$

Theorem 8 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 8 to specific procedures, the key step is to verify Assumption 1 and justify Assumption 2, which we will do for Bonferroni corrected marginal t tests as an example in the next section.

Moreover, the scaling of α_N plays an important role in theoretical discussion. To achieve perfect selection, we hope α_N decays as fast as possible; ideally if α_N equals zero then we do not commit any type I error (or equivalently, we will never select redundant effects). However, for many data-dependent selection procedure α can only decay at certain rates, because a fast decaying α means higher possibility of rejection, thus can lead to strict under-selection. Therefore, in the tuning process, α_d should be scaled properly if one wants to pursue perfect selection. Nevertheless, even if the tuning is hard and perfect model selection can not be achieved, we still have many strategies to exploit the advantage of the forward screening procedure. We will have more discussions in later sections.

B Technical proofs

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

Consider an estimator of the form

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

with variance estimation

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

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

finite-pop **Lemma 3** (Variance concentration and Berry-Esseen bounds in finite population)**.** *Denote* $\gamma = \mathbb{E}\{\widehat{\gamma}\}$, $v^2 = \mathrm{Var}(\widehat{\gamma})$ *and* $v_R^2 = \mathbb{E}\{\widehat{v}_R^2\}$. *Suppose the following conditions hold:*

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

$$Q^{-2} \sum_{\boldsymbol{z}=1}^{Q} w(\boldsymbol{z})^2 N_{\boldsymbol{z}}^{-1} S(\boldsymbol{z}, \boldsymbol{z}) \leq \sigma_w^2 v^2. \tag{2.44}$$

eqn:nondeg

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

$$\max_{\boldsymbol{z} \in [Q]} \frac{1}{N} \sum_{i=1}^{N} \{Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})\}^4 \leq \Delta^4. \tag{2.45}$$

eqn:bounde

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

$$\mathbb{P}\left\{N|\widehat{v}_R - v_R| > t\right\} \leq \frac{C\overline{c}^3 \underline{c}^{-4} \|w\|_\infty^2 \Delta^4}{QN_0} \cdot \frac{1}{t^2}. \tag{2.46}$$

eqn:tail-v

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

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left\{ \frac{\widehat{\gamma} - \gamma}{v} \leq t \right\} - \Phi(t) \right| \leq 2C\sigma_w \frac{\underline{c}^{-1} \|w\|_\infty \max_{i \in [N], \boldsymbol{z} \in [Q]} |Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\|w\|_2 \sqrt{\overline{c}^{-1} \min_{\boldsymbol{z} \in [Q]} S(\boldsymbol{z}, \boldsymbol{z})} \cdot \sqrt{N_0}}. \tag{2.47}$$

eqn:unifor

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

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

Proof of Lemma lem:BE-finite-pop *3.* 1. BE-PCLT Corollary 2.

2. BE-PCLT Theorem 8.

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

$$\sup_{t\in\mathbb{R}} |\Phi\{(1+a)t+b\} - \Phi\{t\}| \le |a| + |b|. \tag{2.48}$$

form:Phi-b

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

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

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

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

Then we can show that

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

For the first term, we have

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

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

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

Besides, by (form:Phi-bD 2.48), when $\epsilon_N\leq 1/2$, we also have

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

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

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

On the other hand, we can show that

$$\begin{aligned}
\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{\widehat{v}_R}\leq t\right\}&\geq\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v}\leq\frac{\widehat{v}_R}{v}t,\left|\frac{\widehat{v}_R-v_R}{v}\right|\leq\epsilon_N\right\} \\
&\geq\mathbb{P}\left\{\frac{\widehat{\gamma}-\gamma}{v}\leq\left(\frac{v_R}{v}-\epsilon_N\right)t\right\}-\mathbb{P}\left\{\left|\frac{\widehat{v}_R-v_R}{v}\right|\geq\epsilon_N\right\}. \qquad (2.49)
\end{aligned}$$

eqn:upper-

By (form:Phi-bD 2.48), when $\epsilon_N\leq 1/2$, we also have

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

So we can derive a lower bound analogous to (eqn:upper-bd 2.49). Note that the results can be analogously generalized to $t\leq 0$. Putting pieces together, we can show that for any

$t \ge 0$,

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

□

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

student-be **Corollary 2.** *Assume Condition (eqn:nondegenerate-var 2.44) and (eqn:bounded-moments 2.45) hold. Let $w = g_{\mathcal{K}}$ for some $\mathcal{K} \in \mathbb{K}$. Then we have a Berry-Esseen bound with the estimated variance:*

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

Proof of Corollary cor:factorial-student-be *2.* **Lower bound for Nv^2.** Note that $\|w\|_2^2 = Q, \|w\|_\infty = 1$. Using Condition (eqn:nondegenerate-var 2.44), we have

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

Therefore, the Berry-Esseen bound becomes

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

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

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

□

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

perfect-ms **Lemma 4** (Berry Esseen bound with screening)**.** *Let*

$$\boldsymbol{f}[\mathbb{M}]=Q^{-1}G(\cdot,\mathbb{M})G(\cdot,\mathbb{M})^\top\boldsymbol{f}. \tag{2.50}$$

eqn:tw

Assume there exists $\sigma_w>0$ *such that*

$$\sum_{\boldsymbol{z}=1}^{Q}\boldsymbol{f}[\mathbb{M}](\boldsymbol{z})^2N_{\boldsymbol{z}}^{-1}S(\boldsymbol{z},\boldsymbol{z})\le\sigma_w^2v^2(\mathbb{M}). \tag{2.51}$$

eqn:nondeg

Then it holds

$$\begin{aligned}&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma[\mathbb{M}]}{v(\mathbb{M})}\le t\right\}-\Phi(t)\right|\\&\le 2\mathbb{P}\left\{\widehat{\mathbb{M}}\ne\mathbb{M}\right\}+2C\sigma_w\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}\cdot\frac{\|\boldsymbol{f}[\mathbb{M}]\|_\infty}{\|\boldsymbol{f}[\mathbb{M}]\|_2}.\end{aligned} \tag{2.52}$$

eqn:tail-p

lem:tail-perfect-ms

Proof of Lemma 4. With the selected working model we have

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

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 (Corollary 2 of BE-PCLT), we have a Berry-Esseen bound with the true variance:

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

□

B.2 Proof of Theorem thm:ms-consistency 8

Proof of Theorem thm:ms-consistency *8.* **Induction proof of a basic fact.** According to the orthogonality of designs, the signs for all terms in the studied unsaturated population regressions are consistent with those of saturated regressions, which saves the effort of differentiating true models for partial and full regression. By induction we hope to prove the following fact under the given assumptions:

For all $D_0 \le D$, we have

$$\left|\mathbb{P}\left(\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d=1,\cdots,D_0\right)-\mathbb{P}\left(\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d=1,\cdots,D_0\right)\right| \to 0. \qquad (2.53)$$ eqn:zerofn-gene

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

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

eqn:zerofn-general (2.53) is equivalent to: for all $D_0 \le D$,

$$\mathbb{P}\left(\text{for all } d\in[D_0], \widehat{\mathbb{M}}_d\subset\mathbb{M}_d^\star;\ \text{there exists } d\in[D_0], \widehat{\mathbb{M}}_d \subsetneq \mathbb{M}_d^\star\right)\to 0. \qquad (2.54)$$ eqn:zerofn

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

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

Therefore,

$$\mathbb{P}\left(\widehat{\mathbb{M}}_1 \subsetneq \mathbb{M}_1^\star\right) \to 0. \text{ (no under selection for main effects)} \tag{2.55}$$

eqn:zerofn

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

   The case for $D_0 = 1$ has been shown in the previous part. Now assume (2.53) or (2.54) for some $D_0 \geq 1$. For $D_0 + 1$, the following holds:

   eqn:zerofneqnzerofn-general-

$$\begin{aligned}
0 &\leq \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0+1\right\}\right) - \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right)\\
&= \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0+1\right\} - \left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right)\\
&\leq \mathbb{P}\left(\forall d \in [D_0+1], \widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star; \exists d \in [D_0], \widehat{\mathbb{M}}_d \subsetneq \mathbb{M}_d^\star\right)\\
&\leq \mathbb{P}\left(\forall d \in [D_0], \widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star; \exists d \in [D_0], \widehat{\mathbb{M}}_d \subsetneq \mathbb{M}_d^\star\right) \to 0. \text{ (by (2.54))}
\end{aligned}$$

   eqn:zerofn-general-1

   Hence

$$\left|\mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d \subset \mathbb{M}_d^\star, d \leq D_0+1\right\}\right) - \mathbb{P}\left(\left\{\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star, d \leq D_0; \widehat{\mathbb{M}}_{D_0+1} \subset \mathbb{M}_{D_0+1}^\star\right\}\right)\right| \to 0. \tag{2.56}$$

   eqn:zerofu

   Now $\widehat{\mathbb{M}}_{D_0+1}$ is generated based on $\widehat{\mathbb{M}}_{D_0}$ and the set of estimates over the prescreened effect set $\widehat{\mathbb{M}}_{D_0+1,+}$. Under Assumption asp:H-heredity 2, on the event $\widehat{\mathbb{M}}_d = \mathbb{M}_d^\star$ we have

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

   Hence we can compute

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

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

   The induction can be proceeded.

Proof of the first result. Now it follows

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

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

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

(using (eqn:zerofn-general 2.53) and the fact that D is a fixed integer)

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

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

$$\leq \limsup_{N\to\infty} \mathbb{P}\left(\widehat{\mathbb{M}}_1 \cap \mathbb{M}_1^{\star c} \neq \varnothing\right) + \sum_{D_0=2}^{D} \mathbb{P}\left(\widetilde{\mathbb{M}}_{D_0} \cap \mathbb{M}_{D_0}^{\star}{}^c \neq \varnothing\right) \leq \sum_{D_0=1}^{D} \alpha_{D_0} = \alpha. \qquad (2.57) \quad \boxed{\texttt{eqn:fwer-g}}$$

Therefore the target probability gets controlled under α.

Proof of the second result. Under $\alpha = \alpha_N \to 0$, (eqn:fwer-general 2.57) implies that with probability tending to one,

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

(hope this time it is easier to read?) Now apply (eqn:zerofn-general 2.53), we obtain

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

with probability tending to one, which concludes the proof.

□

B.3 Proof of Theorem thm:marginal-t 1

We state and prove a more general version of Theorem thm:marginal-t 1:

Theorem 9 (Bonferroni corrected marginal t test). *Let* $\widetilde{\mathbb{M}}_d = \widehat{\mathtt{S}}(\mathbb{M}^\star_{d,+})$ *where* $\mathbb{M}^\star_{d,+} = \mathtt{P}(\mathbb{M}^\star_{d-1})$. *Assume Conditions* [illegible] *1, 2, 3, 4 and 5. Then we have the following results for the screening procedure based on Bonferroni corrected marginal t-test:*

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

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

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

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

(iv) (Perfect selection) When δ' *is strictly positive (not exactly Condition cond:order 2! $\delta' = 0$ is excluded here), 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 thm:marginal-t 1 justified Assumption asp:validity 1 and asp:consistency 2 respectively, which build up the basis for applying Theorem thm:ms-consistency 8. 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 thm:marginal-t 1. 1. First, we show validity:

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

2. Second, we show consistency.

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

For simplicity, let

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

Then

$$\begin{aligned}
&\limsup_{N\to\infty} \mathbb{P}\left\{\widetilde{\mathbb{M}}_d^c \cap \mathbb{M}_k^\star \neq \varnothing\right\} \\
&\leq \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^\star} \left(\mathbb{P}\left\{-Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}} \leq \frac{\widehat{\tau}_{\mathcal{K}}}{\sigma_{\mathcal{K}}} - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}} \leq Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right\} + \frac{\widetilde{C}}{(QN_0)^{1/3}}\right) \\
&= \limsup_{N\to\infty} \sum_{\mathcal{K}\in\mathbb{M}_d^\star} \Phi\left\{r_{\mathcal{K}}^{-1}\left(Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\} - \Phi\left\{r_{\mathcal{K}}^{-1}\left(-Z_d^\star - \frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}\right)\right\}. \qquad (2.58)
\end{aligned}$$

eqn:typeII

With Condition cond:order 2, we have

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

Because $\delta > -1/2$ and $\delta' \geq 0$, we have $|\frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}| \to \infty$ and $Z_d^\star/(|\frac{\tau_{\mathcal{K}}}{\sigma_{\mathcal{K}}}|) \to 0$. Hence the above limit (eqn:typeII-limit 2.58) converges to zero. This concludes the proof.

3. Based on the above two parts and Theorem thm:ms-consistency 8, it suffices to conclude the Type I error rate control. A more delicate analysis in this particular setup can actually lead to

sharper bound. Based on (2.57), we directly compute

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

4. The perfect selection result follows from Part 1,2 and Theorem 8.

□

B.4 Proof of Theorem 10

Theorem 10 (Consistency of the selected tie sets)**.** *Assume Conditions 3, 4 and 6. There exists universal constants $C, C' > 0$, such that when $N > n(\delta_1, \delta_2, \delta_3)$,*

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

Theorem 10 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.

thm:wt-converge

Proof of Theorem 10. 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

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

lem:tail-perfect-ms

We apply Lemma 4 to establish a Berry-Esseen bound for each $\widehat{Y}_{\mathrm{r}}^\star(\boldsymbol{z})$. Note that

$$\widehat{Y}_{\mathrm{r}}^\star(\boldsymbol{z}) = \boldsymbol{f}_{\boldsymbol{z}}^\top \widehat{Y}, \ \boldsymbol{f}_{\boldsymbol{z}}^\top = Q^{-1}G(\boldsymbol{z}, \mathbb{M}^\star)G(\cdot, \mathbb{M}^\star)^\top. \tag{2.59}$$

eqn:hY-r-s

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

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left\{ \frac{\widehat{Y}_{\mathrm{r}}^\star(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})}{v_N} \le t \right\} - \Phi(t) \right| \le 2C\sigma \frac{\underline{c}^{-1} \max_{i\in[N], \boldsymbol{z}\in[Q]} |Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z}, \boldsymbol{z})\}^{1/2}} \sqrt{\frac{|\mathbb{M}^\star|}{QN_0}}. \tag{2.60}$$

eqn:BE-mar

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

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

We can show that

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

Hence we know that

$$\begin{aligned}
&\mathbb{P}\left\{\sqrt{N}\left|\widehat{Y}_{\mathrm{r}}^{\star}(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})\right| \geq \sqrt{N} d_h^{\star}\right\} \\
\leq & \frac{v_N}{\sqrt{2\pi} d_h^{\star}} \cdot \exp\left(-\frac{d_h^{\star 2}}{2 v_N^2}\right)+2 C \sigma \frac{\underline{c}^{-1} \max_{i \in[N], \boldsymbol{z} \in[Q]}\left|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})\right|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z} \in[Q]} S(\boldsymbol{z}, \boldsymbol{z})\}^{1/2}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}} .
\end{aligned} \tag{2.61}$$

eqn:dev-bo

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

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

Using (eqn:dev-bound 2.61) we have

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

Now a union bound gives

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

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

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

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

$$\begin{aligned}
&\mathbb{P}\left\{\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}') - \widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) < 0\right\} \\
\geq & 1 - C|\mathcal{T}_1||\mathcal{T}'|\left\{\sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^\star|}{N^{1+2\delta_1}}}\exp\left\{-\frac{C'N^{1+2\delta_1}}{\bar{c}\bar{s}|\mathbb{M}^\star|}\right\} + \sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^\star|}{N_0Q}}\right\}.
\end{aligned} \tag{2.62}$$

eqn:order

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

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}})|/\eta_N \geq 2\bar{\epsilon}\right\} \\
&\geq \mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}')|/\eta_N \geq 2\bar{\epsilon}, \widetilde{m} \in \mathcal{T}_1\right\} \\
&\geq \mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}')|/\eta_N \geq 2\bar{\epsilon}\right\} + \mathbb{P}\{\widetilde{m} \in \mathcal{T}_1\} - 1 \\
&\geq \mathbb{P}\{\widetilde{m} \in \mathcal{T}_1\} - \sum_{\boldsymbol{z}\notin\mathcal{T}_1,\boldsymbol{z}'\in\mathcal{T}_1}\mathbb{P}\left\{|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}}')|/\eta_N \leq 2\bar{\epsilon}\right\}.
\end{aligned} \tag{2.63}$$

eqn:separa

To proceed we have the following tail bound:

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

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

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

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

$$\sum_{\boldsymbol{z}\notin\mathcal{T}_1, \boldsymbol{z}'\in\mathcal{T}_1} \mathbb{P}\left\{|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}')|/\eta_N \le 2\bar{\epsilon}\right\}$$

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

Combined with (eqn:separation-1 2.63), we have:

$$\begin{aligned}
&\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}})|/\eta_N \ge 2\bar{\epsilon}\right\}\\
\ge&\mathbb{P}\left\{\widetilde{m}\in\mathcal{T}_1\right\} - \underbrace{4|\mathcal{T}_1||\mathcal{T}'|\frac{\sqrt{2\bar{c}\bar{s}|\mathbb{M}^\star|}}{\sqrt{\pi N_0 Q}d_h^\star}\cdot\exp\left(-\frac{N_0 Q d_h^{\star 2}}{8\bar{c}\bar{s}|\mathbb{M}^\star|}\right)}_{\text{Term I}}\\
&-\underbrace{4|\mathcal{T}_1||\mathcal{T}'|2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^\star|}{N_0 Q}}}_{\text{Term II}}.
\end{aligned}\tag{2.64}$$

eqn:large

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

$$\mathbb{P}\left\{\min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}})|/\eta_N \ge 2\bar{\epsilon}\right\}$$

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

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

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

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

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

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,

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

Hence we have:

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

Again, by the conditions, we can show

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

Specifying the random indices. Introduce the following random indices:

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

Applying (eqn:order 2.62) we know that

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

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

$$\begin{aligned}&\mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}})| \le \underline{\epsilon}\eta_N, \min_{\boldsymbol{z}\notin\mathcal{T}_1}|\widehat{Y}^\star_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^\star_{\mathrm{r}}(\widetilde{\boldsymbol{z}})| \ge \bar{\epsilon}\eta_N\right\}\\ \ge& 1 - C|\mathcal{T}'||\mathcal{T}_1|\left\{\sqrt{\frac{\bar{c}\bar{s}|\mathbb{M}^\star|}{N^{1+2\delta_2}}}\exp\left(-\frac{C'N^{1+2\delta_2}}{\bar{c}\bar{s}|\mathbb{M}^\star|}\right) + \sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\bar{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}}\cdot\sqrt{\frac{|\mathbb{M}^\star|}{N_0 Q}}\right\}.\end{aligned} \tag{2.65}$$

eqn:key-bo

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;\\
&\qquad |\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;\\
&\qquad |\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}\}\\
&\quad \text{(where we introduce random index } \widetilde{\boldsymbol{z}}_h \text{ to record the position that achieves maximum)}\\
\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;\\
&\qquad |\widehat{Y}^{\star}_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| > \overline{\epsilon}\eta_N, \text{ for } \boldsymbol{z} \notin \mathcal{T}_1\Bigg\} - \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^{\star}\}\\
&\quad \text{(simply using the fact that } \overline{\epsilon} > \underline{\epsilon}\text{)}\\
=&\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)| \le \underline{\epsilon}\eta_N; \min_{\boldsymbol{z}\notin\mathcal{T}_1} |\widehat{Y}^{\star}_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| > \overline{\epsilon}\eta_N\Bigg\}\\
&\quad - \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^{\star}\}\\
\ge&1 - \sum_{h=1}^{H_0}\left(1 - \mathbb{P}\left\{\max_{\boldsymbol{z}\in\mathcal{T}_1} |\widehat{Y}^{\star}_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| \le \underline{\epsilon}\eta_N; \min_{\boldsymbol{z}\notin\mathcal{T}_1} |\widehat{Y}^{\star}_{\mathrm{r}}(\boldsymbol{z}) - \widehat{Y}^{\star}_{\mathrm{r}}(\widetilde{\boldsymbol{z}}_h)| > \overline{\epsilon}\eta_N\right\}\right)\\
&\quad - \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^{\star}\}\\
\ge&1 - \mathbb{P}\{\widehat{\mathbb{M}} \neq \mathbb{M}^{\star}\}\\
&-C|\mathcal{T}'||\mathcal{T}_1|\left\{\sqrt{\frac{\overline{c}\overline{s}|\mathbb{M}^{\star}|}{N^{1+2\delta_2}}}\exp\left(-\frac{C'N^{1+2\delta_2}}{\overline{c}\overline{s}|\mathbb{M}^{\star}|}\right) + \sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z}) - \overline{Y}(\boldsymbol{z})|}{\overline{c}^{-1/2}\{\min_{\boldsymbol{z}\in[Q]} S(\boldsymbol{z},\boldsymbol{z})\}^{1/2}} \cdot \sqrt{\frac{|\mathbb{M}^{\star}|}{N_0 Q}}\right\}.
\end{aligned}
$$

□

B.5 Proof of Theorem thm:be-perfect-ms 2

We prove a Berry Esseen bound, which naturally translates into the desired CLT result:

Theorem 11 (Berry-Esseen bound under perfect screening)**.** *Let* $\boldsymbol{f}[\mathbb{M}]$ *be given by* (eqn:tw 2.50)*. Assume* (eqn:nondegenerate-var-tw 2.51)*. Then it holds*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}(\widehat{\mathbb{M}})-\gamma}{v(\mathbb{M}^\star)}\leq t\right\}-\Phi(t)\right|$$
$$\leq 2\mathbb{P}\left\{\widehat{\mathbb{M}}\neq\mathbb{M}^\star\right\}+2C\sigma_w\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}\cdot\frac{\|\boldsymbol{f}[\mathbb{M}^\star]\|_\infty}{\|\boldsymbol{f}[\mathbb{M}^\star]\|_2}. \quad (2.66)$$

eqn:be-per

Proof of Theorem thm:be-perfect-ms *2.* This theorem is a direct application of Lemma lem:tail-perfect-ms 4. First we check that

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

From the definition of γ (eqn:mean-hgamma 4.16), we have

$$\begin{aligned}\gamma &= \boldsymbol{f}^\top\overline{Y}\\ &= \boldsymbol{f}^\top G\tau = \boldsymbol{f}^\top G(\cdot,\mathbb{M}^\star)\tau(\mathbb{M}^\star) \text{ (due to (eqn:reparametrization 2.2))}\\ &= Q^{-1}\boldsymbol{f}^\top G(\cdot,\mathbb{M}^\star)G(\cdot,\mathbb{M}^\star)^\top\overline{Y} = \gamma(\mathbb{M}^\star).\end{aligned}$$

Now apply Lemma lem:tail-perfect-ms 4 with $\mathbb{M}=\mathbb{M}^\star$ gives the result. □

B.6 Proof of Theorem thm:strategy-I 4

Theorem 12 (Guarantee for Strategy str:exclude-all 1)**.** *Assume Condition* cond:under-selection *7. Also assume* $\boldsymbol{f}$ *satisfies the following orthogonality condition:*

$$G(\cdot,\mathbb{M}_d^\star)^\top\boldsymbol{f}=0,\ d^\star+1\leq d\leq D^\star. \quad (2.67)$$

eqn:orthog

Let $\boldsymbol{f}[\mathbb{M}^\star_{1:d^\star}]$ *be given by* (eqn:bsf-M 5.26) *with* $\mathbb{M}^\star_{1:d^\star}=\cup_{d=1}^{d^\star}\mathbb{M}^\star_d$*. Assume Condition* cond:nondegenerate-corr *3. Then*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma}{v[\mathbb{M}^\star_{1:d^\star}]}\leq t\right\}-\Phi(t)\right|$$
$$\leq 2\mathbb{P}\left\{\widehat{\mathbb{M}}\neq\mathbb{M}^\star_{1:d^\star}\right\}+2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})}\cdot\sqrt{N_0}}\cdot\frac{\|\boldsymbol{f}[\mathbb{M}^\star_{1:d^\star}]\|_\infty}{\|\boldsymbol{f}[\mathbb{M}^\star_{1:d^\star}]\|_2}. \quad (2.68)$$

eqn:tail-u

Proof of Theorem thm:strategy-I *4.* According to Condition cond:under-selection 7, with Strategy str:exclude-all 1,

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

We will apply Lemma lem:tail-perfect-ms 4 with

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

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

$$\begin{aligned}\gamma &= \boldsymbol{f}^\top\overline{Y}\\ &= \boldsymbol{f}^\top G\tau = \boldsymbol{f}^\top G(\cdot,\mathbb{M})\tau(\mathbb{M}) + \boldsymbol{f}^\top G(\cdot,\mathbb{M}^c)\tau(\mathbb{M}^c) \text{ (due to (eqn:reparametrization 2.2)).}\end{aligned}$$

Now by (eqn:orthogonality-app 2.67), $\boldsymbol{f}^\top G(\cdot,\mathbb{M}^c) = 0$. Hence

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

□

B.7 Proof of Theorem thm:strategy-II 5

Theorem 13 (Guarantee for Strategy str:select-by-heredity 2)**.** *Assume Conditions* cond:nondegeneracy *3,* cond:heredity *5 and* cond:under-selection *7. Let*

$$\mathbb{M}^{\star\star} = \bigcup_{d=1}^{D}\mathbb{M}_d^{\star\star},$$

where

$$\mathbb{M}_d^{\star\star} = \begin{cases} \mathbb{M}_d^\star, & d \le d^\star;\\ \mathtt{H}^{(d-d^\star)}(\mathbb{M}_{d^\star}^\star), & d^\star+1 \le d \le D.\end{cases}$$

Then

$$\mathbb{M}^\star \subset \mathbb{M}^{\star\star},\ \mathbb{P}\left\{\widehat{\mathbb{M}} = \mathbb{M}^{\star\star}\right\} \to 1.$$

Let $\boldsymbol{f}[\mathbb{M}^{\star\star}]$ *be given by (*eqn:bsf-M *5.26) with* $\mathbb{M} = \mathbb{M}^{\star\star}$*. Then we have*

$$\begin{aligned}&\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{\widehat{\gamma}[\widehat{\mathbb{M}}]-\gamma}{v[\mathbb{M}^{\star\star}]} \le t\right\} - \Phi(t)\right|\\ &\le 2\mathbb{P}\left\{\widehat{\mathbb{M}} \neq \mathbb{M}^{\star\star}\right\} + 2C\sigma\frac{\underline{c}^{-1}\max_{i\in[N],\boldsymbol{z}\in[Q]}|Y_i(\boldsymbol{z})-\overline{Y}(\boldsymbol{z})|}{\sqrt{\overline{c}^{-1}\min_{\boldsymbol{z}\in[Q]}S(\boldsymbol{z},\boldsymbol{z})\cdot\sqrt{N_0}}}\cdot\frac{\|\boldsymbol{f}[\mathbb{M}^{\star\star}]\|_\infty}{\|\boldsymbol{f}[\mathbb{M}^{\star\star}]\|_2}. \qquad (2.69)\end{aligned}$$

eqn:tail-o

Proof of Theorem thm:strategy-II *5.* This proof can be finished by applying Lemma lem:tail-perfect-ms 4 with $\mathbb{M} = \mathbb{M}^{\star\star}$ and checking $\gamma(\mathbb{M}^{\star\star}) = \gamma$, which is omitted here. □

C General theory on selecting the best tie sets