Forward factorial screening with statistical inference

Ph.D. Qualification Exam

Lei Shi
Division of Biostatistics, UC Berkeley
November 30th, 2022

Outline

2^K Factorial Experiments

- K factors: $z_k \in \{0, 1\}, k = 1, \ldots, K$.
- $Q = 2^K$ treatment combinations:

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

- N: sample size
- $N(\boldsymbol{z})$: sample size under treatment $\boldsymbol{z} \in \mathcal{T}$.
- $Y_i(\boldsymbol{z})$: potential outcome if i-th unit was assigned to the treatment $\boldsymbol{z}$.
- Z_i: the observed treatment for unit i.
- $Y_i = Y_i(Z_i)$: the observed outcome for unit i.
- **Complete randomization**: randomly permuting the treatment labels among the N units:

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

A 2^3 Factorial Experiment: Hiring Discrimination

- Pedulla et al. (2022) conducted a field experiment examining the hiring discrimination by **race**, **gender**, and **parental status** in the United States.
- Fake CVs were then generated for randomization.
- Collect feedback from the employers.

Use our notation, this example translates into a 2^3 factorial design:

- Three binary factors z_1 (race), z_2 (gender) and z_3 (parental status);
- 8 treatment combinations:

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

Factorial Effect: Main Effect

- Defined by a special set of **contrast vectors** (Wu and Hamada, 2011; Dasgupta et al., 2015).
- Denote the average of potential outcomes as $\bar{Y} = \{\bar{Y}(\boldsymbol{z})\}_{\boldsymbol{z}\in\mathcal{T}}$:

$$\bar{Y}(\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^{N} Y_i(\boldsymbol{z}).$$

- Contrast vector $g_{\{k\}}$:

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

  (aggregating centered treatment indicators into a vector)
- The **main effect** for factor z_k:

$$\tau_{\{k\}} = Q^{-1} \cdot g_{\{k\}}^{\top}\bar{Y}, \quad k \in [K].$$

Factorial Effect: Interaction Effect

- Contrast vector $g_{\mathcal{K}}$: for multiple factors $\mathcal{K} = \{k_1, \ldots, k_d\}$ with $d \geq 2$, do **element-wise product** of $g_{\{k_1\}}, \cdots, g_{\{k_d\}}$:

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

- The **interaction effect** among factors in $\mathcal{K}$:

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

- $\tau_{\mathcal{K}}$ is a **parent effect** of $\tau_{\mathcal{K}'}$ if $\mathcal{K} \subset \mathcal{K}'$ and $|\mathcal{K}| = |\mathcal{K}'| - 1$.
- Stack the contrast vectors and factorial effects:

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

A 2^3 Factorial Experiment: Hiring Discrimination

- The vector of factorial effects:

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

  where G is the contrast matrix

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

Estimation with Factor-based Regression

Factor-based regression is a model-assisted strategy for estimating factorial effects.

- **Saturated regression**: $Y_i \sim t_i$.
  - t_i is the row vector in G indexed by $Z_i = (z_{i,1}, \ldots, z_{i,K})^\top$.
- **Unsaturated regression**: $Y_i \sim$ a subvector of t_i.
  - Denote the subvector of t_i as $t_{i,\mathbb{M}}$ for a collection of factor combinations $\mathbb{M}$.
  - We refer to $\mathbb{M}$ as a **working model**.
- $\hat{\tau}(\mathbb{M})$: estimated coefficients with working model $\mathbb{M}$.
- $\tau(\mathbb{M})$: the collection of true factorial effects in $\mathbb{M}$.
- With weighted least squares: $\hat{\tau}(\mathbb{M})$ are unbiased for $\tau(\mathbb{M})$ (Zhao and Ding, 2021b).

A 2^3 Factorial Experiment: Hiring Discrimination

- The regressor t_i can be constructed from Z_i:

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

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

$$Y_i \sim t_i.$$

- For the unsaturated regression, if we use the working model

$$\mathbb{M} = \{\varnothing, \{1\}, \{1,2\}, \{1,3\}, \{1,2,3\}\},$$

it means the following regression:

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

Outline

Motivation

Challenge	Goal
Too many causal parameters in large K settings: $2^K - 1$ factorial effects ($2^8 - 1 = 255$, $2^{10} - 1 = 1023$...)	Identify a few significant factorial effects and screen out negligible effects
No principled methods for effect screening in design-based framework (what if we select $\{\{1\}, \{2, 3\}\}$...)	Develop a procedure that fully respects the structure of factorial experiments
No discussion on how effect screening affects estimation and inference	Exploring post-screening inference for general causal parameters

Motivation

Can we take advantage of the **unique structure** of factorial experiments?

- Three principles summarized by Wu and Hamada (2011):
  - **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, all of its parent effects should be significant (*strong heredity*) or at least one of its parent effects should be significant (*weak heredity*).
- Motivates a natural screening procedure that proceeds in a forward style!

Literature Review

- Linear regressions in factorial experiments (and general randomized experiments):
  - Zhao and Ding (2021b): saturated regression and unsaturated regression with given $\mathbb{M}$;
  - Zhao and Ding (2021a): regression adjustment and restricted least squares;
    (no data-driven guidance on how to decide a specification; theoretical analysis is performed under fixed Q settings)
  - Bloniarz et al. (2016): regression adjustment with Lasso;
    (binary treatment, i.e., 2^1 factorial experiment; not targeting effect screening)
- Forward selection:
  - Wang (2009); Wieczorek and Lei (2022): sequentially adding terms with smallest RSS;
    (requires strong modelling assumptions such as normality and homoscedasticity)
- Interaction screening with heredity principles:
  - Yuan et al. (2007); Lim and Hastie (2015); Bien et al. (2013); Haris et al. (2016): Lasso variants with heredity; (purely algorithmic, no statistical guarantees)
  - Hao and Zhang (2014); Hao et al. (2018): extension to regression with interactions;
    (relies heavily on normal distributions; cannot handle three-way interactions)

Our Goal

Develop a screening procedure that ...

- respects the natural hierarchical structure of factorial experiments;
- can be easily implemented with low computational cost;
- demonstrates good finite sample performance as well as strong large sample theoretical guarantees;
- can facilitate estimation and inference for general causal parameters in factorial experiments.

Outline

Forward Factorial Screening: A Running Example

2^3 factorial experiment, forward screening with Bonferroni corrected marginal t-tests

- Input: a sequence of significance levels $\alpha_1, \alpha_2, \alpha_3$;
- Heredity: strong heredity.

Step	Working Model
1. Start with an intercept term	$\hat{\mathbb{M}} = \{\varnothing\}$
2. Add all the main effects	$\hat{\mathbb{M}} = \{\varnothing, \{1\}, \{2\}, \{3\}\}$
3. Regress $Y_i \sim t_{i,\hat{\mathbb{M}}}$	$\hat{\mathbb{M}} = \{\varnothing, \{1\}, \{2\}, \{3\}\}$
4. Do t-tests on main effects with level $\alpha_1/\lvert\hat{\mathbb{M}}\backslash\{\varnothing\}\rvert$ and drop the non-rejections	$\hat{\mathbb{M}} = \{\varnothing, \{1\}, \{2\}\}$
5. Add two way interactions under strong heredity	$\hat{\mathbb{M}} = \{\varnothing, \{1\}, \{2\}, \{1,2\}\}$
6. Regress $Y_i \sim t_{i,\hat{\mathbb{M}}}$	$\hat{\mathbb{M}} = \{\varnothing, \{1\}, \{2\}, \{1,2\}\}$
7. Do t-tests on two-way interactions with level $\alpha_2/\lvert\hat{\mathbb{M}}\backslash\{\varnothing, \{1\}, \{2\}\}\rvert$ and drop the non-rejections	$\hat{\mathbb{M}} = \{\varnothing, \{1\}, \{2\}\}$
8. No two-way effects identified; return $\hat{\mathbb{M}}$	$\hat{\mathbb{M}} = \{\varnothing, \{1\}, \{2\}\}$

Forward Factorial Screening: Formal Algorithm

A forward screening procedure based on Bonferroni corrected marginal t-test:

- Initialization: factorial data $\{(Y_i, Z_i)\}_{i=1}^N$; target level D; initial working model $\hat{\mathbb{M}} = \{\varnothing\}$; significance level $\{\alpha_d\}_{d=1}^D$.
- for $d = 1, \cdots, D$:
  1. Update the intermediate working model to include all the d-order (interaction) terms: $\hat{\mathbb{M}}' = \hat{\mathbb{M}} \cup \{\mathcal{K} \mid |\mathcal{K}| = d\} \triangleq \hat{\mathbb{M}} \cup \mathbb{K}_d$.
  2. Screen out indices in $\hat{\mathbb{M}}'$ according to either the weak or strong heredity principles;
  3. Run the unsaturated factor-based regression on the working model $\hat{\mathbb{M}}'$:

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

  4. Obtain coefficients $\hat{\tau}(\hat{\mathbb{M}}')$ and robust covariance estimation $\hat{\mathbf{\Sigma}}(\hat{\mathbb{M}}')$:
  5. Extract $\hat{\tau}_{\mathcal{K}}(\hat{\mathbb{M}}')$ and $\hat{\sigma}_{\mathcal{K}}(\hat{\mathbb{M}}')$ for all $\mathcal{K} \in \hat{\mathbb{M}}'$ with $|\mathcal{K}| = d$.
  6. Run marginal t-test using the above $\hat{\tau}_{\mathcal{K}}(\hat{\mathbb{M}}')$ and $\hat{\sigma}_{\mathcal{K}}(\hat{\mathbb{M}}')$ under significance level $\min\{\alpha_d/(|\hat{\mathbb{M}}'| - |\hat{\mathbb{M}}|), 1\}$ and remove the non-significant terms from $\hat{\mathbb{M}}' \backslash \hat{\mathbb{M}}$.
  7. Set $\hat{\mathbb{M}} = \hat{\mathbb{M}}'$.

Forward Factorial Screening: Dissection

- The screening procedure is iterated in a forward style:

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

  - Respects the "Effect Hierarchy Principle": forward algorithm
  - Respects the "Effect Sparsity Principle": S-step
  - Respects the "Effect Heredity Principle": H-step
- Generates a highly interpretable working model
- Compatible with many screening methods: Marginal t-test (Wasserman and Roeder, 2009), Lasso (Zhao and Yu, 2006), SIS (Fan and Lv, 2008)...

Forward Factorial Screening: Screening Consistency

Regularity conditions:

- **Nearly uniform design**: $N(\boldsymbol{z}) = c(\boldsymbol{z})N_0 \geq 2,$ where $\underline{c} \leq c(\boldsymbol{z}) \leq \bar{c}$.
- **Regularity conditions on potential outcomes**: bounded moments, nondegenerate correlation, etc. (See the manuscript for details)
- **Effect heredity**: true effects follow the heredity structure.
- **Scale of parameters**: parameters can change with N.
  1. Nonzero effects: $|\tau_{\mathcal{K}}| = \Theta(N^{\delta})$ for some $-1/2 < \delta \leq 0$ and all $\mathcal{K} \in \cup_{d=1}^{D} \mathbb{M}_d^{\star}$.
  2. Significance level: $\sum_{d \in [D]} \alpha_d = \Theta(N^{-\delta'})$ with some $\delta' > 0$.
  3. Size of the target working model: $\sum_{d=1}^{D} |\mathbb{M}_d^{\star}| = \Theta(N^{\delta''/3})$ for some $0 \leq \delta'' < 1$.

Quick discussion on scale of parameters:

- Nonzero effects: standard Beta-min condition in literature (e.g. Zhao and Yu, 2006).
- Significance level: asymptotically vanishing false rejections (Wasserman and Roeder, 2009);
- Size of the target working model: $o(N^{1/3})$ might not be optimal; conjecture $o(N^{1/2})$.
  - Wasserman and Roeder (2009): $O(1)$;
  - Wieczorek and Lei (2022): $o(\sqrt{N})$;
  - Zhao and Yu (2006): $O(N^{1-c})$;

Forward Factorial Screening: Screening Consistency

Theorem (Perfect screening property)

Under the regularity conditions in the last slide, the working model selected by the forward factorial screening procedure converges to the true model with probability one as the sample size goes to infinity:

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

- The asymptotic regime is on N, not N_0. (allow large Q and small N_0).
- D can be any integer in $[K]$. (allow perfect screening for only lower order effects)

Outline

Inference under Perfect Screening: General Effects

- A general causal parameter of interest is

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

  where $\boldsymbol{f} = \{\boldsymbol{f}(\boldsymbol{z})\}_{\boldsymbol{z} \in \mathcal{T}}$ is a pre-specified weighting vector.

- Concrete examples:
  - $\boldsymbol{f}$ can be taken as the dense contrast vectors $g_{\mathcal{K}}$'s;
  - $\boldsymbol{f}$ can be taken as sparse contrast vectors; for example, $\boldsymbol{f} = (-1, 1, 0, \ldots, 0)^\top$, which measures causal difference of two specific arms.

Inference under Perfect Screening: Two Estimators

- Estimator **without** effect screening (simple moment estimator):

$$\hat{\gamma} = \sum_{z \in \mathcal{T}} f(z)\hat{Y}(z) \triangleq f^\top \hat{Y}, \quad \hat{v}^2 = \sum_{z \in \mathcal{T}} f(z)^2 N(z)^{-1} \hat{S}(z,z) \triangleq f^\top \hat{V}_{\hat{Y}} f$$

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

- Estimator **after** effect screening (RLS based estimator):

$$\hat{\gamma}_{\mathsf{r}} = f^\top \hat{Y}_{\mathsf{r}} \triangleq f[\hat{\mathbb{M}}]^\top \hat{Y}, \quad \text{and} \quad \hat{v}^2_{\mathsf{r}} = f[\hat{\mathbb{M}}]^\top \hat{V}_{\hat{Y}} f[\hat{\mathbb{M}}],$$

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

  – Motivated by restricted least squares:

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

Inference under Perfect Screening: Theory

Theorem (Statistical properties of $\hat{\gamma}_r$ and $\hat{v}_r^2$)

Let N goes to infinity. Under the previous conditions and

$$N_0^{-1/2} \cdot \|\boldsymbol{f}^\star\|_\infty / \|\boldsymbol{f}^\star\|_2 \longrightarrow 0, \quad \boldsymbol{f}^\star = \boldsymbol{f}[\mathbb{M}^\star] = Q^{-1} G(\cdot, \mathbb{M}^*) G(\cdot, \mathbb{M}^*)^\top \boldsymbol{f}, \tag{1}$$

we have

$$\frac{\hat{\gamma}_r - \gamma}{v_r} \rightsquigarrow \mathcal{N}(0,1)$$

where $v_r^2 = \boldsymbol{f}^{\star\top} V_{\hat{\gamma}} \boldsymbol{f}^\star, \quad V_{\hat{\gamma}} = \mathrm{Diag}\left\{N(\boldsymbol{z})^{-1} S(q,q)\right\} - N^{-1} S$. *Furthermore, assume* $\|\boldsymbol{f}^\star\|_\infty = O(Q^{-1})$, $\hat{v}_r^2$ *is consistent and robust:*

$$N(\hat{v}_r^2 - v_{r,\lim}^2) \xrightarrow{\mathbb{P}} 0, \quad v_{r,\lim}^2 \geq v_r^2,$$

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

Inference under Perfert Screening: Comparison

- Conditions for asymptotic normality:
  - $\hat{\gamma}$: $N_0^{-1/2} \cdot \|\boldsymbol{f}\|_\infty / \|\boldsymbol{f}\|_2 \longrightarrow 0$.
    Requires $\boldsymbol{f}$ to be dense or N_0 to be large.
  - $\hat{\gamma}_{\mathrm{r}}$: $N_0^{-1/2} \cdot \|\boldsymbol{f}^\star\|_\infty / \|\boldsymbol{f}^\star\|_2 \longrightarrow 0$.
    Holds even for sparse $\boldsymbol{f}$ and small N_0 if $\mathbb{M}^\star$ sparse, due to the following bound:

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

- Asymptotic relative efficiency: let $\kappa(V_{\hat{\gamma}})$ be the condition number of $V_{\hat{\gamma}}$.
  - In general, the relative efficiency depends on choice of $\boldsymbol{f}^\star$. (A well known drawback for sandwich estimators under heteroskedasticity; see Zhao and Ding (2021b))
  - If $\kappa(V_{\hat{\gamma}}) = 1$ (i.e., $V_{\hat{\gamma}} = I_Q$), then $v_{\mathrm{r}}^2 / v^2 \leq 1$.
  - If $\boldsymbol{f}^\star$ has $s^\star$ nonzeros, then

$$\frac{v_{\mathrm{r}}^2}{v^2} \leq \kappa(V_{\hat{\gamma}}) \cdot \frac{s^\star |\mathbb{M}^\star|}{Q}.$$

Outline

Inference under Imperfect Screening: Difficulties

- Perfect screening property can be too much to hope for...
- Usually fails when the higher order factorial effects are **marginal**.
- For example, if the minimal signal condition

$$|\tau_{\mathcal{K}}| = \Theta(N^{\delta}) \text{ for some } -1/2 < \delta \leq 0$$

fails, we lose the perfect screening property.

Inference under Imperfect Screening: Difficulties

- Inference without perfect screening is hard...
  - the selected model can be anything;
  - difficult for estimating and inferring parameters independent of the selected model.

- Classical post-selection inference methods have their own drawback in randomized factorial experiments:
  - Sample splitting: permutational distributions; expensive data collection.
  - Selective inference: data-dependent CI; relies heavily on the specific selection methodology (Kuchibhotla et al., 2022).
  - Simultaneous inference: no concept of ”true model” in design-based framework.

- Want to focus on schemes that work best for the purpose of factorial experiments

General Strategies under Imperfect Screening

If only the first $d^\star$ levels have large effect size...

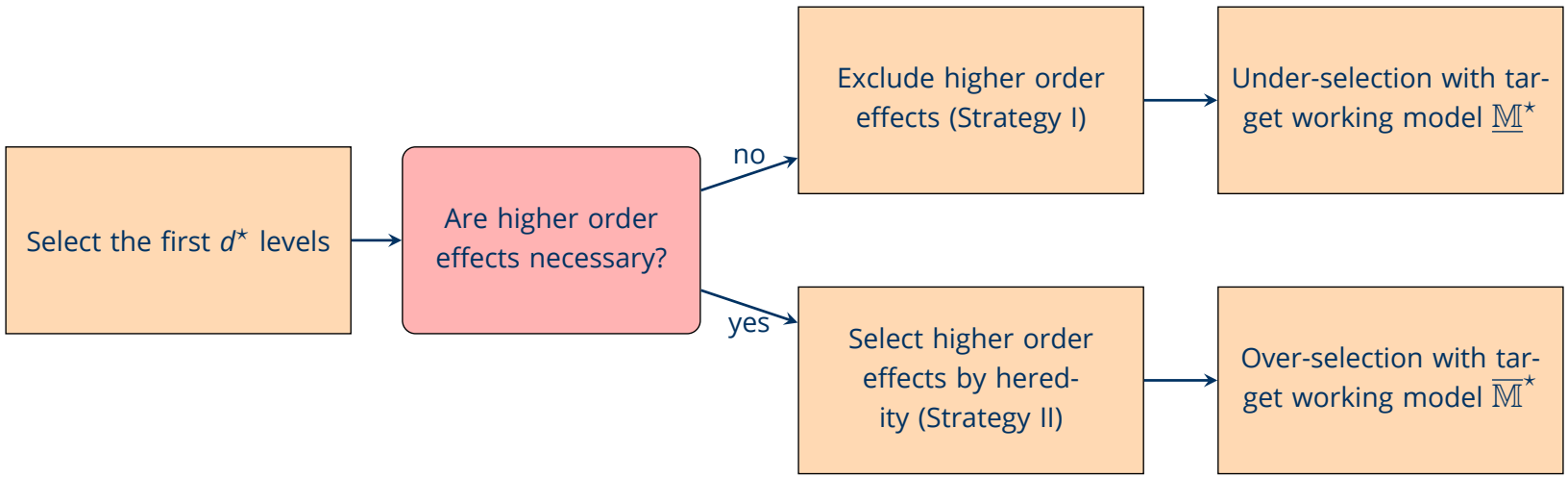

Figure 1: General strategies for factorial screening

Under-selection: Excluding High Order Interactions

Scenario 1: high order interactions are not necessary for study.

- Early stop the screening procedure at level $d = d^*$.
- Selected working model: $\underline{\hat{\mathbb{M}}} = \cup_{d=1}^{d^\star} \hat{\mathbb{M}}_d$.
- Target working model: $\underline{\mathbb{M}}^\star = \cup_{d=1}^{d^\star} \mathbb{M}_d^\star \subseteq \mathbb{M}^\star$. (may lead to under-selection)
- Construct an estimator of $\gamma = \boldsymbol{f}^\top \bar{Y}$ based on the restricted least squares:

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

Over-selection: Screening by Effect Heredity

Scenario 2: high order interactions are necessary for study.

- Apply the screening procedure until level $d = d^*$; select higher level interactions by a heredity principle (either weak or strong).
- Selected working model: $\hat{\overline{\mathbb{M}}} = \cup_{d=1}^{D} \hat{\mathbb{M}}_d$; $\hat{\mathbb{M}}_d = \mathbf{H}^{(d-d^\star)}(\hat{\mathbb{M}}_{d^\star}), d \geq d^\star + 1$.
- Target working model:

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

(may lead to under-selection)

- Construct an estimator of $\gamma = \boldsymbol{f}^\top \bar{Y}$ based on the restricted least squares:

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

Theoretical Guarantee for Under-selection

Theorem

Define $\underline{\boldsymbol{f}}^\star = Q^{-1} G(\cdot, \underline{\mathbb{M}}^\star) G(\cdot, \underline{\mathbb{M}}^\star)^\top \boldsymbol{f}$. *Assume some regularity conditions. Also assume* $\boldsymbol{f}$ *satisfies the following orthogonality condition:*

$$G(\cdot, \mathbb{M}_d^\star)^\top \boldsymbol{f} = 0, \; d^\star + 1 \leq d \leq D^\star. \tag{2}$$

If as N tends to infinity, $N_0{}^{-1/2} \cdot \|\underline{\boldsymbol{f}}^\star\|_\infty / \left\|\underline{\boldsymbol{f}}^\star\right\|_2 \longrightarrow 0$, *then*

$$\frac{\hat{\gamma}_{\text{ru}} - \gamma}{v_{\text{ru}}} \rightsquigarrow \mathcal{N}(0, 1).$$

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

$$N(\hat{v}_{\text{ru}}^2 - v_{\text{ru,lim}}^2) \xrightarrow{\mathbb{P}} 0, \quad v_{\text{ru,lim}}^2 \geq v_{\text{ru}}^2,$$

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

Theoretical Guarantee for Over-selection

Theorem

Assume some regularity conditions. Define the weighting vector $\bar{\boldsymbol{f}}^\star = Q^{-1} G(\cdot, \overline{\mathbb{M}}^\star) G(\cdot, \overline{\mathbb{M}}^\star)^\top \boldsymbol{f}$. *If as N tends to infinity,*

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

then

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

where $v_{\text{ro}}^2 = \bar{\boldsymbol{f}}^{\star\top} V_{\hat{\mathbf{y}}} \bar{\boldsymbol{f}}^\star$. *Furthermore, without loss of generality, assume* $\left\| \bar{\boldsymbol{f}}^\star \right\|_\infty = O(Q^{-1})$, *the variance estimator* $\hat{v}_{\text{ro}}^2$ *is consistent and robust:*

$$N(\hat{v}_{\text{ro}}^2 - v_{\text{ro,lim}}^2) \xrightarrow{\mathbb{P}} 0, \quad v_{\text{ro,lim}}^2 \geq v_{\text{ro}}^2,$$

where $v_{\text{ro,lim}}^2 = \bar{\boldsymbol{f}}^{\star\top} \operatorname{Diag}\left\{ N(\boldsymbol{z})^{-1} S(\boldsymbol{z}, \boldsymbol{z}) \right\} \bar{\boldsymbol{f}}^\star$ *is the robust asymptotic variance of* $\hat{\gamma}_{\text{ro}}$.

Trade-off between the two strategies

There is a trade-off between the two strategies:

Under-selection:

- more bias for general causal parameter;
- give a parsimonious working model, but possibly too small;
- smaller asymptotic variance in some interesting regimes.

Over-selection:

- less bias for general causal parameter;
- give a complete working model, but possibly too large;
- larger asymptotic variance in some interesting regimes.

Outline

Simulation study

- Set up a 2^8 uniform factorial experiment ($K = 8$).
- N_0 units in each treatment arm.
- Potential outcomes are generated independently from an exponential distribution:

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

- $\mu(\boldsymbol{z})$ are super population means which gives factorial effects with strong heredity:
  - Main effects: the first five factors, $\tau_{\{1\}}, \ldots, \tau_{\{5\}}$, are nonzero; the rest, $\tau_{\{6\}}, \tau_{\{7\}}, \tau_{\{8\}}$, are zero.
  - Two-way interactions: the two-way interactions associated with the first five factors are nonzero, i.e.,

    $$\tau_{\{kl\}} \neq 0 \text{ for } k \neq l, k, l \in [5].$$

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

Simulation Study

- Four methods for comparison:
  - **Forward Bonferroni**: forward screening based on Bonferroni corrected margin t tests;
  - **Forward Lasso**: forward screening based on Lasso;
  - **Naive Bonferroni**: full screening based on Bonferroni corrected margin t tests;
  - **Naive Lasso**: full screening with based on Lasso.
- Three criteria:
  - **Perfect selection probability**: $\mathbb{P}\{\hat{\mathbb{M}} = \mathbb{M}^\star\}$.
  - **Power of $\hat{\gamma}_{\mathbf{r}}$**: testing a causal effect γ_{target} specified by a sparse vector:

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

  - **Coverage probability of the RLS based confidence interval**.

Varying Sample Size

In the first study, we fix the magnitude of the nonzero effects (0.20 for main effects and 0.10 for interactions) and vary the number of replications N_0.

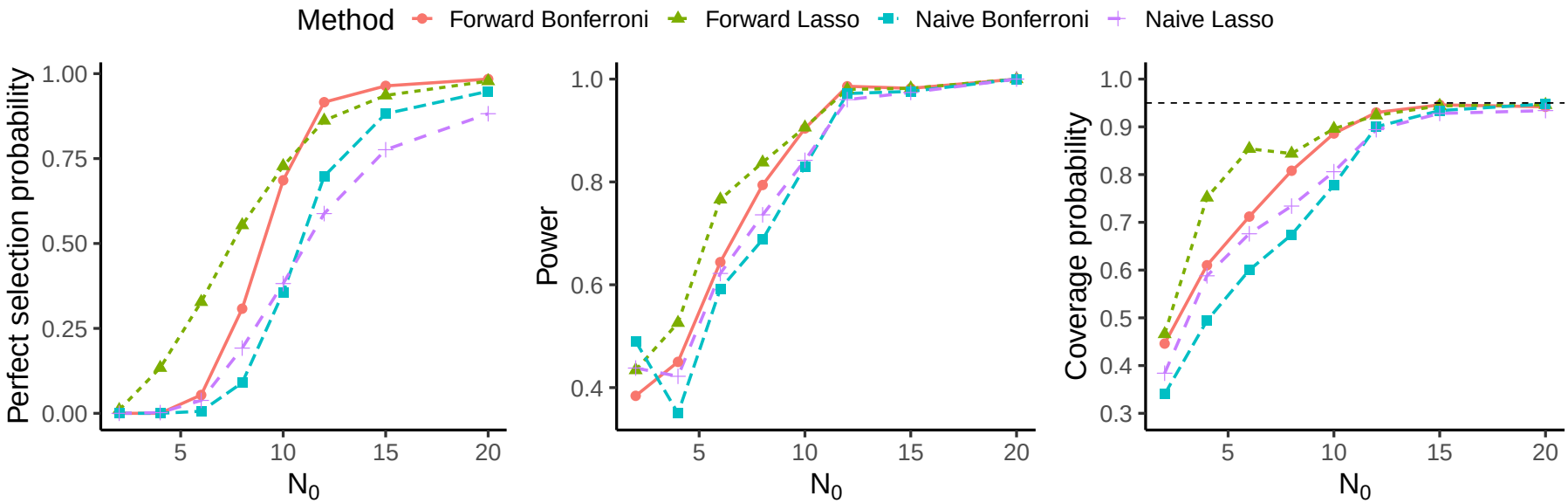

Figure 2: Simulation study with varying sample size

Varying Effect Size

In the second study, we fix the number of replications within each arm ($N_0 = 2$) and vary the effect size.

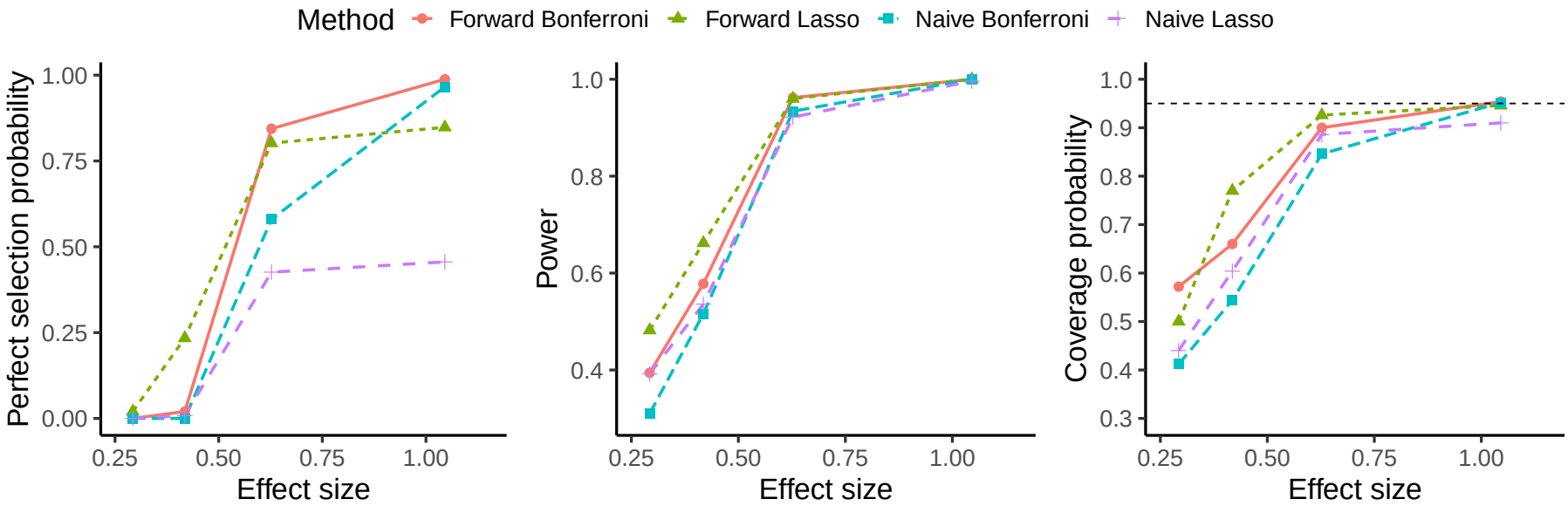

Figure 3: Simulation study with varying effects size

Outline

Generalization: Compare Multiple Causal Effects

- We study "inference on best causal effects" problem as an application.
- A classical problem with many applications (Andrews et al., 2019; Wei et al., 2022).
- Suppose we have a set of causal effects defined by pre-specified weighting vectors:

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

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

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

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

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

Generalization: Compare Multiple Causal Effects

The key part is to identify the tie set.

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

1. Perform forward effect screening: obtain working model $\hat{\mathbb{M}}$.
2. Obtain RLS based estimates: compute $\hat{\gamma}_l = \boldsymbol{f}_l^\top \hat{Y}_r = G(\cdot, \hat{\mathbb{M}})\hat{\tau}(\hat{\mathbb{M}}), \quad l \in [L]$.
3. **Select the best factor level combinations**:

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

4. Define

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

Then generate estimators

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

Generalization: Compare Multiple Causal Effects

Theorem (Asymptotic results on the estimated effects with screening)

Assume some regularity conditions. Assume $|\mathbb{M}^\star| = \Theta(N^{\delta_4})$ *for some* $\delta_4 \geq 0$ *and* $N^{-(1+2\delta_2-\delta_4)} \to 0, \quad L \cdot |\mathcal{L}_1| \cdot N^{-\frac{1-\delta_4}{2}} \to 0.$

Then the point estimates are asymptotically jointly normal:

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

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

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

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

Outline

Discussion

- Covariate adjustment?
  - Lin (2013) proposed regressions with t, X and interactions $t \otimes X$.
  - Can be singular if either number of treatment levels or dimension of X is large.
- Other definitions of factorial effects?
  - More general centering regimes (Zhao and Ding, 2021b).
  - Partial results in our Appendix: under certain conditions, unsaturated regressions will change the estimates but keep the signs.
- Unreplicated designs?
  - In some extreme cases, some treatment arms only contain 1 unit;
  - The commonly used variance estimator ill-defined;
  - General discussion on inference in another manuscript: Shi and Ding (2022).
- Data-driven tuning?
  - Wei et al. (2022) proposed a double bootstrap procedure for comparing multiple effects;
  - Unknown if similar extension works for design-based framework, both computationally and theoretically.

Discussion

- Multiple-level factorial experiments?
  - Should be a straightforward extension.
- Sample splitting and inference for general screening procedure?
  - Studied frequently for iid data; some for weak correlated data under α-mixing conditions.
  - Never studied for **permutational distribution**, but seems doable.
- Selective inference?
  - An empty research area in design-based framework.
- Fractional factorial design?
  - The story should be different due to aliasing.

Thank you!

I would like to thank Jingshen and Peng for the wonderful advising on the whole project. Also thanks a ton to Lexin and Will for kindly serving on my QE committee and offering all the helpful comments.

Reference I

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

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

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390.

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.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

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

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

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

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

Reference II

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

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

Pedulla, D. S., Muñoz, J., Wullert, K. E., and Dias, F. A. (2022). Field experiments and job posting sources: The consequences of job database selection for estimates of racial discrimination. *Sociology of Race and Ethnicity*, 8(1):26–42.

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

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

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

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

Wieczorek, J. and Lei, J. (2022). Model selection properties of forward selection and sequential cross-validation for high-dimensional regression. *Canadian Journal of Statistics*, 50(2):454–470.

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

Reference III

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

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

Zhao, A. and Ding, P. (2021b). Regression-based causal inference with factorial experiments: estimands, model specifications and design-based properties. *Biometrika*, 109(3):799–815.

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