The Annals of Statistics
2025, Vol. 53, No. 2, 506–530
https://doi.org/10.1214/24-AOS2454

FORWARD SELECTION AND POST-SELECTION INFERENCE IN FACTORIAL DESIGNS

BY LEI SHI[1,a], JINGSHEN WANG[1,b] AND PENG DING[2,c]

[1]*Division of Biostatistics, University of California, Berkeley,* [a]*leishi@berkeley.edu*
[2]*Department of Statistics, University of California, Berkeley,* [b]*jingshenwang@berkeley.edu,* [c]*pengdingpku@berkeley.edu*

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

1. Introduction.

1.1. *Factorial experiments*: *Opportunities and challenges*. Ever since the seminal work of Fisher [13] and Yates [45], factorial designs have been widely used in many fields, including agricultural, industrial, biomedical and social sciences (e.g., [6, 16, 43]). Factorial experiments are popular because they can simultaneously accommodate multiple factors and offer opportunities to estimate not only the main effects of factors but also their interactions.

We focus on the replicated 2^K factorial design in which K binary factors are randomly assigned to N experimental units, and each treatment combination contains at least two replicates. Classical factorial experiments are usually conducted with a small K so that we can simultaneously estimate the $2^K - 1$ main effects and interactions. For example, Chapter 4 of [43] discussed many full factorial experiments, all of which involve less than or equal to four factors ($K \leq 4$). However, many modern factorial experiments are conducted on a much larger scale for exploring complex research questions. For example, in political science and marketing research, powered by the development of computers and web-based technology, conjoint survey experiments [8, 18, 31, 53], which can be viewed as a special type of factorial experiments, are popular for analyzing the effects of many factors together. In Table 1, we review several conjoint experiments in the literature and their corresponding setups. These modern factorial experiments involve a large number of treatment combinations, which motivate us to move beyond the classical small K regimes and develop methodology and theory for large K regimes.

Received March 2024; revised August 2024.
MSC2020 subject classifications. Primary 62K15, 62J05; secondary 62G05.
Key words and phrases. Causal inference, design-based inference, forward selection, post-selection inference.

TABLE 1
Some conjoint survey experiments and the corresponding setups

Experiment	Reference	K	Q	N	N_0
Immigrant admission experiment	[53]	6	$2^6 = 64$	$\sim 28{,}000$	~ 430
U.S. presidential election	[8]	12	$2^{12} = 4096$	$\sim 30{,}000$	~ 8
Aluminum packaging characteristics	[24]	7	$2^6 * 4 = 256$	$\sim 15{,}000$	~ 60

Note: K is the number of factors, Q is the number of treatment combinations, N is the number of units (hypothetical profiles), and $N_0 = N/Q$ is the average replications per treatment arm. See Section 2 for the definitions.

A large number of factors pose new challenges to the analysis of factorial experiments. First, estimation and inference of the causal effects fall into new regimes, especially when K is large and each treatment combination only contains limited replications. Second, a large K results in a large set of factorial effects, which complicate the interpretation of the results. This motivates us to conduct factor selection based on *sparsity, hierarchy and heredity* principles for factorial effects to reduce the dimensionality of the problem and facilitate the interpretation of the results. Wu and Hamada [43] summarized these three principles as below:

(a) (sparsity) The number of important factorial effects is small.
(b) (hierarchy) Lower-order effects are more important than higher-order effects, and effects of the same order are equally important.
(c) (heredity) Higher-order effects are important only if their corresponding lower-order effects are important.

The sparsity principle motivates conducting factor selection in factorial designs. The hierarchy principle motivates the forward selection strategy that starts from lower-order effects and then moves on to higher-order effects. The heredity principle motivates using structural restrictions to select higher-order effects based on the selected lower-order effects. Due to its simplicity and computational efficiency, the forward selection strategy has been widely used in data analysis [11, 43]. However, its design-based theory under the potential outcome framework has not been formally established. Moreover, it is often challenging to understand the impact of factor selection on the subsequent statistical inference. The overarching goal of this manuscript is to fill these gaps.

1.2. *Our contributions and literature review*. We summarize our contributions from the following three perspectives.

First, our study adds to the growing literature of factorial designs with the number of factors increasing with the sample size under the potential outcome framework [4, 7, 9–11, 29, 33, 44, 48]. To deal with a large number of factors, Espinosa, Dasgupta and Rubin [11] and Egami and Imai [10] informally used factor selection without studying its theoretical properties, whereas Zhao and Ding [48] discussed parsimonious model specifications that are chosen a priori and independent of data. The rigorous theory for factor selection is missing in this literature, let alone the theory for statistical inference after factor selection. At a high level, our paper fills the gaps.

Second, we formalize forward factor selection and establish its consistency under the design-based framework without imposing outcome modeling assumptions; see Section 3. Factor selection in factorial design sounds like a familiar statistical task if we formulate it as a variable selection problem in a linear model. Thus, forward selection is reminiscent of the vast literature on forward selection. Wang [39] and Wieczorek and Lei [42] proved the consistency of forward selection for the main effects in a linear model, whereas Hao and Zhang

[20] and Hao, Feng and Zhang [19] moved further to allow for second-order interactions. Other researchers proposed various penalized regressions to encode the sparsity, hierarchy and heredity principles (e.g., [2, 3, 21, 26, 47, 50]), without formally studying the statistical properties of the selected model. Our design-based framework departs from the literature without assuming a correctly-specified linear outcome model. This framework is classic in experimental design and causal inference with randomness coming solely from the design of experiments rather than the error terms in a linear model [9, 15, 22, 27, 37]. This framework invokes fewer outcome modeling assumptions but consequently imposes technical challenges for developing the theory. Bloniarz et al. [5] discussed the design-based theory for covariate selection in treatment-control experiments, but the corresponding theory for factorial designs is largely unexplored.

Third, we discuss statistical inference after forward factor selection with consistent (see Sections 4) and inconsistent selection (see Section 5). First, we prove the selection consistency of the forward selection procedure, which ensures that the selected factorial effects are the true, nonzero ones as the sample size grows. With this selection consistency property, we can then proceed as if the selected working model is the true model. This allows us to ignore the impact of forward selection on the subsequent inference, which is similar to the proposal of Zhao, Witten and Shojaie [52] for statistical inference after Lasso [38]. Moreover, we quantify the advantages of conducting forward selection based on the asymptotic efficiency gain for estimating factorial effects. As an application under selection consistency, we discuss statistical inference for the mean outcome under the best factorial combination in Section A.4 in the Supplementary Material [1, 17, 41]. Second, we acknowledge that selection consistency can be difficult to achieve in practice as it requires strong regularity conditions on factorial effects. As a remedy, we propose two strategies to deal with inconsistent selection in higher-order interactions and study their impacts on post-selection inference. A key motivation for our strategies is to ensure that the parameters of interest after forward factorial selection are not data-dependent, avoiding philosophical debates in the current literature of post-selection inference [14, 23].

1.3. *Notation*. We will use the following notation throughout. For asymptotic analyses, $a_N = O(b_N)$ denotes that there exists a positive constant $C > 0$ such that $a_N \le Cb_N$; $a_N = o(b_N)$ 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 $cb_N \le a_N \le Cb_N$.

For matrix V, define $\varrho_{\max}(V)$ and $\varrho_{\min}(V)$ as the largest and smallest eigenvalues, respectively, and define $\kappa(V) = \varrho_{\max}(V)/\varrho_{\min}(V)$ as its condition number. For two positive semidefinite matrices V_1 and V_2, we write $V_1 \preccurlyeq V_2$ or $V_2 \succcurlyeq V_1$ if $V_2 - V_1$ is positive semidefinite, which is called the Loewner order for positive semidefinite matrices.

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

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

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

2.1. *Potential outcomes and factorial effects.* We first introduce the framework of a 2^K factorial design, with $K \geq 2$ being an integer. The design has K binary factors, and factor k can take value $z_k \in \{-1, 1\}$ for $k = 1, \ldots, K$; we use the ± 1 coding of the factors because of its convenience for later parts and can modify the results under the 0-1 coding of the factors. Let $z = (z_1, \ldots, z_K)$ denote the treatment combination with factor levels $z_1, \ldots, z_K$. The K factors in total define $Q = 2^K$ treatment combinations, collected in the set below:

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

We follow the potential outcome notation of Dasgupta, Pillai and Rubin [9] for 2^K factorial designs. Unit i has potential outcome $Y_i(z)$ under treatment combination z. Corresponding to the $Q = 2^K$ treatment combinations, unit i has Q potential outcomes, vectorized as $\boldsymbol{Y}_i = \{Y_i(z)\}_{z\in\mathcal{T}}$ using the lexicographic order based on the treatment combinations. Over units $i = 1, \ldots, N$, the potential outcomes have finite-population mean vector $\overline{Y} = (\overline{Y}(z))_{z\in\mathcal{T}}$ and covariance matrix $S = (S(z, z'))_{z,z'\in\mathcal{T}}$, with elements defined as follows:

$$\overline{Y}(z) = \frac{1}{N}\sum_{i=1}^{N} Y_i(z), \qquad S(z, z') = \frac{1}{N-1}\sum_{i=1}^{N}\big(Y_i(z) - \overline{Y}(z)\big)\big(Y_i(z') - \overline{Y}(z')\big). \tag{2.1}$$

We then use the potential outcomes to define factorial effects. For a subset $\mathcal{K} \subset [K]$ of the K factors, we introduce the following "contrast vector" notation to facilitate the presentation. To start with, we define the main causal effect for factor k. For a treatment combination $z = (z_1, \ldots, z_K) \in \mathcal{T}$, we use $g_{\{k\}}(z) = z_k$ to denote the "centered" treatment indicator z_k. We then define a Q-dimensional contrast vector $g_{\{k\}}$ by concatenating these centered treatment indicators in the lexicographic order of $z \in \mathcal{T}$, that is,

$$g_{\{k\}} = \big\{g_{\{k\}}(z)\big\}_{z\in\mathcal{T}}, \quad \text{where } g_{\{k\}}(z) = z_k. \tag{2.2}$$

Next, for the interactions of multiple factors in $\mathcal{K}$ with $|\mathcal{K}| \geq 2$, we define the contrast vector $g_{\mathcal{K}} \in \mathbb{R}^Q$ as

$$g_{\mathcal{K}} = \big\{g_{\mathcal{K}}(z)\big\}_{z\in\mathcal{T}}, \quad \text{where } g_{\mathcal{K}}(z) = \prod_{k\in\mathcal{K}} g_{\{k\}}(z) = \prod_{k\in\mathcal{K}} z_k. \tag{2.3}$$

Finally, for the average of potential outcomes, we define $g_\varnothing = \mathbf{1}_Q$, which is orthogonal to all the contrast vectors. Stack the $g_{\mathcal{K}}$'s into a $Q \times Q$ matrix

$$G = (g_\varnothing, g_{\{1\}}, \ldots, g_{\{K\}}, g_{\{1,2\}}, \ldots, g_{\{K-1,K\}}, \ldots, g_{[K]}), \tag{2.4}$$

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

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

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

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

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

The above definition for factorial effects differs from [9] by a constant of 2, which does not change the problem fundamentally but has a better mathematical structure under our framework.

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

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

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

as the sample mean and variance of the observed outcomes under the treatment combination z. Recall that S is the finite population covariance matrix of the potential outcomes defined in (2.1). Let $D_{\widehat{Y}} = \mathrm{Diag}\{N(z)^{-1}S(z,z)\}_{z\in\mathcal{T}}$. Vectorize the sample means as $\widehat{Y} = (\widehat{Y}(z))_{z\in\mathcal{T}}$, which has mean $\overline{Y}$ and covariance matrix $V_{\widehat{Y}} = D_{\widehat{Y}} - N^{-1}S$ [25]. An unbiased estimator for $D_{\widehat{Y}}$ is

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

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

A dominant approach to estimating factorial effects from factorial designs is through estimating least-squares coefficients based on appropriate model specifications. Let g_i denote the row vector in the contrast matrix G corresponding to unit i's treatment combination Z_i, that is, $g_i = \{g_{\mathcal{K}}(Z_i)\}_{\mathcal{K}\subset[K]} \in \mathbb{R}^Q$ with $g_{\mathcal{K}}(z)$ defined in (2.3). We can run ordinary least squares (OLS) to obtain unbiased estimates for the factorial effects:

$$\widehat{\tau} = \arg\min_{\tau}\sum_{i=1}^{N}(Y_i - g_i^{\top}\tau)^2. \tag{2.5}$$

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

In contrast, an *unsaturated regression* with weighted least squares (WLS) involves fewer coefficients by regressing the observed outcome Y_i on $g_{i,\mathbb{M}}$, a subvector of g_i, where $\mathbb{M} \subset \mathbb{K}$ is a subset of the power set of all factors:

$$\widehat{\tau} = \arg\min_{\tau}\sum_{i=1}^{N} w_i(Y_i - g_{i,\mathbb{M}}^{\top}\tau)^2 \quad \text{with } w_i = 1/N_i. \tag{2.6}$$

The above least squares fit in (2.5) and (2.6) are based on a fact in factor-based regressions: to get unbiased estimates for a set of factorial effects, one can either run OLS/WLS with a saturated model or run WLS including that particular set of effects with an unsaturated model. Such a result is established in, for example, Section 5.4 and Section A.5 of [48].

For the convenience of description, we will call $\mathbb{M}$ a *working model*. We use a working model to generate estimates based on least squares without assuming the corresponding linear

model is correct. When $\mathbb{M} = \mathbb{K}$, (2.6) incorporates the saturated regression (2.5). Based on the unsaturated regression with working model $\mathbb{M}$, let

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

denote the vectors of estimated and true coefficients, respectively. $\widehat{\tau}(\mathbb{M})$ has the following closed-form expression:

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

Because $\widehat{\tau}(\mathbb{M})$ is a linear transformation of $\widehat{Y}$, we can use the following estimator for its covariance matrix:

$$\widehat{\mathbf{\Sigma}}(\mathbb{M}) = \frac{1}{Q^2}G(\cdot,\mathbb{M})^\top \widehat{V}_{\widehat{Y}} G(\cdot,\mathbb{M}). \tag{2.7}$$

See Lemma S1 in Section A.1 of the Supplementary Material [36] for more discussions on the above algebraic results for unsaturated regressions.

REMARK 1. The terminology "saturated regression" here means that the factor-based linear regression takes the full set of elements in g_i as regressors, while the "unsaturated regression" only takes a subset. In the experimental design literature, the word "saturated" has a different meaning in the terminology "saturated design," which is used to indicate that the number of observations in the design equals the number of effects in the model.

2.3. *An example of a* 2^3 *factorial design.* The above notation can be abstract. In this section, we provide an illustrating Example 1 below with $K = 3$ factors.

EXAMPLE 1 (2^3 factorial design). Suppose we have three binary factors z_1, z_2 and z_3. These three factors generate 8 treatment combinations, indexed by a triplet $(z_1z_2z_3)$ with $z_1, z_2, z_3 \in \{-1, 1\}$, in the set (keeping the signs but omitting "1")

$$\mathcal{T} = \big\{(---), (--+), (-+-), (-++), (+--), (+-+), (++-), (+++)\big\}.$$

Each unit i has a potential outcome vector $\boldsymbol{Y}_i = \{Y_i(z_1z_2z_3)\}^\top_{z_1,z_2,z_3=-1,1}$. The vector of factorial effects is

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

where G is the contrast matrix

$$G = \begin{array}{c} \\ (---)\\(--+)\\(-+-)\\(-++)\\(+--)\\(+-+)\\(++-)\\(+++)\end{array}\!\!\begin{array}{c}\begin{array}{cccccccc}\tau_\varnothing & \tau_{\{1\}} & \tau_{\{2\}} & \tau_{\{3\}} & \tau_{\{1,2\}} & \tau_{\{1,3\}} & \tau_{\{2,3\}} & \tau_{\{1,2,3\}}\end{array}\\ \begin{pmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1\\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1\\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1\\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1\\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1\\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1\\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1\\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1\end{pmatrix}\end{array}.$$

We observe the pair (Y_i, Z_i) for unit i, where $Z_i = (z_{i,1}, z_{i,2}, z_{i,3})$ is the observed treatment combination for unit i. Recall $g_{\{k\}}(Z_i) = z_{i,k}$. For the factor-based regression, the regressor g_i corresponding to the treatment combination Z_i equals

$$g_i = \big[1,\ g_{\{1\}}(Z_i),\ g_{\{2\}}(Z_i),\ g_{\{3\}}(Z_i),\ g_{\{2,3\}}(Z_i),\ g_{\{1,3\}}(Z_i),\ g_{\{1,2\}}(Z_i),\ g_{\{1,2,3\}}(Z_i)\big]^\top.$$

For instance, when $Z_i = (+\,-\,+)$, the regressor g_i corresponds to the row $(+\,-\,+)$ of the contrast matrix G. Then a saturated regression is to regress Y_i on g_i. For the unsaturated regression, if we only include indices $\varnothing$, $\{1\}$, $\{1, 2\}$, $\{1, 3\}$, $\{1, 2, 3\}$, we can form the working model $\mathbb{M} = \{\varnothing, \{1\}, \{1, 2\}, \{1, 3\}, \{1, 2, 3\}\}$ and perform WLS $Y_i \sim g_{i,\mathbb{M}}$, where

$$g_{i,\mathbb{M}} = \big[1,\ g_{\{1\}}(Z_i),\ g_{\{1,3\}}(Z_i),\ g_{\{1,2\}}(Z_i),\ g_{\{1,2,3\}}(Z_i)\big]^\top$$

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

3. Forward selection in factorial experiments. In factorial designs with small K, we can run the saturated regression to estimate all factorial effects simultaneously [29, 48]. However, when K is large, estimation and inference of the causal effects fall into new regimes, especially when each treatment combination only contains limited replications. Moreover, it is difficult to interpret a large number of estimates for factorial effects. As a remedy, forward selection is a popular strategy frequently adopted to analyze data collected from factorial experiments, due to its benefits in ruling out zero nuisance factorial effects in a systematic way. In this section, we formalize forward selection as a principled procedure to select an unsaturated working model $\widehat{\mathbb{M}}$. We first present a formal version of forward selection and then demonstrate its consistency property.

3.1. *A formal forward selection procedure.* In this subsection, we introduce a principled forward selection procedure that not only respects the effect hierarchy, sparsity and heredity principles but also results in an interpretable parsimonious model with statistical guarantees. More concretely, the algorithm starts by performing factor selection over lower-order effects, then moves forward to select higher-order effects following the heredity principle. Algorithm 1 summarizes the forward selection procedure. In what follows, we illustrate why the proposed procedure in Algorithm 1 respects the three fundamental principles in factorial experiments.

First, Algorithm 1 obeys the hierarchy principle as it performs factor selection in a forward style (coded in the global loop from $d = 1$ to $d = D \le K$ with prespecified D that represents the maximum number of factors in the selected working model, Step 2 in particular). More concretely, we begin with an empty working model. We then select main effects (Steps 4 and 8) and add them into the working model. Once the working model is updated, we continue to select higher-order interaction effects in a forward style. Such a forward selection procedure is again motivated by the hierarchy principle that lower-order effects are more important than higher-order ones.

Second, Algorithm 1 operates under the sparsity principle as it removes potentially unimportant effects using marginal t-tests with the Bonferroni correction (see Step 8). This step induces a sparse working model and helps us to identify important factorial effects. The sparsity-inducing step can incorporate many popular selection frameworks, such as marginal t-tests, Lasso [38], sure independence screening [12], etc. For simplicity, we present Algorithm 1 with marginal t-tests and relegate general discussions to Section A.3 of the Supplementary Material.

Third, Algorithm 1 incorporates the heredity principle as it rules out the interaction effects [20, 26, 43] when either none of their parent effects is included (weak heredity) or some of their parent effects are excluded (strong heredity) in the previous working model (see Step 4).

Algorithm 1: Forward factorial selection

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

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

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

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

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

4 Drop indices in $\widehat{\mathbb{M}}'$ according to either the weak or strong heredity principles, and renew the selected working model as $\widehat{\mathbb{M}}'$.

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

$$Y_i \sim g_{i,\widehat{\mathbb{M}}'}, \text{ with weights } w_i = 1/N_i.$$

6 Obtain coefficients $\widehat{\tau}(\widehat{\mathbb{M}}')$ and robust covariance estimation $\widehat{\mathbf{\Sigma}}(\widehat{\mathbb{M}}')$ defined in (2.7).

7 Obtain $\widehat{\tau}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ and $\widehat{\sigma}^2_{\mathcal{K}}(\widehat{\mathbb{M}}')$ for all $\mathcal{K} \in \widehat{\mathbb{M}}'$ with $|\mathcal{K}| = d$, where $\widehat{\sigma}^2_{\mathcal{K}}(\widehat{\mathbb{M}}')$ is the variance estimator in the diagonal values of $\widehat{\mathbf{\Sigma}}(\widehat{\mathbb{M}}')$ corresponding to the factor combination $\mathcal{K}$.

8 Run marginal t-tests using the above $\widehat{\tau}_{\mathcal{K}}(\widehat{\mathbb{M}}')$ and $\widehat{\sigma}^2_{\mathcal{K}}(\widehat{\mathbb{M}}')$ under the significance level $\min\{\alpha_d/(|\widehat{\mathbb{M}}'| - |\widehat{\mathbb{M}}|), 1\}$ and remove the nonsignificant terms from $\widehat{\mathbb{M}}' \backslash \widehat{\mathbb{M}}$.

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

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

Lastly, Algorithm 1 enhances the interpretability of the selected working model by iterating between the "Sparsity-selection" step (Step 8, called the S-step in the rest of the manuscript), captured by a data-dependent operator $\widehat{\mathrm{S}} = \widehat{\mathrm{S}}(\cdot; \{Y_i, Z_i\}_{i=1}^N)$, and the "Heredity-selection" step (Step 4, called the H-step in the rest of the manuscript), captured by a deterministic operator $\mathrm{H} = \mathrm{H}(\cdot)$. Because the working model is updated in an iterative fashion,

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

the final working model includes effects that fully respect the heredity principle.

REMARK 2. While forward selection has been set up as a standard tool in the literature (e.g., [20, 39]), we provided Algorithm 1 as it is customized for the factorial designs. More specifically, Algorithm 1 differs from existing proposals [20, 39] from several perspectives. (i) The tools for effect selection are different. In Algorithm 1, we use factor-based regression and robust variance estimation, which do not assume the true outcome model is linear. In contrast, the forward regression procedure in [20, 39] selects the variables by iteratively minimizing the residual sum of squares. The validity of their procedure relies on the linear model assumption and the homoskedasticity of the noise. (ii) The theoretical justification is different. We study the property of forward selection from the design-based perspective, where the randomness originates from the treatment assignment. On the contrary, [20, 39] focus on linear models as the underlying data-generating process, where the randomness comes from the outcomes. The forward selection procedure requires novel theoretical justification under the design-based framework.

3.2. *Consistency of forward selection.* We are now ready to analyze the selection consistency property of Algorithm 1. We shall show that Algorithm 1 selects the targeted working model up to level D with probability tending to one as the sample size goes to infinity. Here, the targeted working model at level $k \in [K]$, denoted as $\mathbb{M}_k^\star$, is the collection of $\mathcal{K}$'s where $|\mathcal{K}| = k$ and $\tau_{\mathcal{K}} \neq 0$. Define the full targeted working model up to level D as

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

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

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

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

$$N(z) = c(z)N_0 \geq 2, \quad \text{where } \underline{c} \leq c(z) \leq \overline{c},$$

where $\underline{c}$ and $\overline{c}$ are universal constants that do not depend on other quantities.

Condition 1 is a finite sample characterization of the design. It allows either Q or N_0 (thus $N(z)$ across all treatment combinations) to diverge [35]. It generalizes the classical assumption where Q is fixed, and each treatment combination contains a sufficiently large number of replications [25]. The quantities Q, N_0, $c(z)$ can vary for different design settings, which leads to different asymptotic regimes. In general, N has the order of $O(QN_0)$ under Condition 1.

Next, we quantify the order of the true factorial effect sizes $\tau_{\mathcal{K}}$'s and the tuning parameters α_d's adopted in the Bonferroni correction in Algorithm 1. We allow these parameters to change with the sample size N.

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

(i) *True nonzero factorial effects*: $|\tau_{\mathcal{K}}| = \Theta(N^{\delta})$ *for some* $-1/2 < \delta \leq 0$ *and all* $\mathcal{K} \in \mathbb{M}_{1:D}^\star$.
(ii) *Tuning parameters in Bonferroni correction*: $\alpha_d = \Theta(N^{-\delta'})$ *for all* $d \in [D]$ *for some* $\delta' > 0$.
(iii) *Size of the targeted working model*: $\sum_{d=1}^{D} |\mathbb{M}_d^\star| = \Theta(N^{\delta''})$ *for some* $0 \leq \delta'' < 1/3$.

Condition 2(i) specifies the allowable order of the true factorial effects. If Condition 2(i) fails with a $\delta \leq -1/2$, the effect size is of the same or smaller order as the statistical error, and thus is too small to be detected by marginal t-tests. As a special case, when the number of nonzero factorial effects has a finite upper limit as $N \to \infty$, then Condition 2(i) is satisfied with $\delta = 0$. Similar conditions are also adopted in the variable selection literature under the linear model, including [51] and [42]. Condition 2(ii) requires the tuning parameter α_d to converge to zero, which ensures that there is no Type I error asymptotically in our procedure as N goes to infinity, which is crucial for the selection consistency. Wasserman and Roeder [40], Theorems 4.1 and 4.2, assumed similar conditions in the variable selection literature under the linear model. Condition 2(iii) restricts the size of the targeted working model. The rate is due to our technical analysis. As a special case, when the number of nonzero effects is a constant that does not grow with N, it suffices to set $\delta'' = 0$. Similar conditions also appeared in [42, 51] and [40].

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

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

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

$$\max_{i\in[N],q\in[Q]} \left|Y_i(q) - \overline{Y}(q)\right| < \nu, \qquad \min_{q\in[Q]} S(q,q) > \underline{S}.$$

Condition 3(i) requires the correlation matrix of $\widehat{Y}$ to be well behaved. Condition 3(ii) imposes a universal bound on potential outcomes and their variances, which is a sufficient condition by [35] to prove the Berry–Esseen bound based on Stein's method.

Lastly, we impose the following structural conditions on the factorial effects.

CONDITION 4 (Hierarchical structure in factorial effects). *The nonzero true factorial effects obey the effect heredity principle*:

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

Finally, we present the selection consistency property of Algorithm 1.

THEOREM 1 (Consistent selection property). *Under Conditions 1–4, we have*

$$\lim_{N\to\infty} \mathrm{P}\{\widehat{\mathbb{M}} = \mathbb{M}^\star_{1:D}\} = 1.$$

Theorem 1 guarantees that the working model selected by Algorithm 1 converges to the targeted working model with probability one as the sample size goes to infinity. Here, we used the terminology "consistent selection" that is coherent with [34]. A closely related property is "screening consistency," which is a terminology by [12] and refers to the fact that the selected model covers the true model with high probability and allows for overselection.

4. Inference under selection consistency. Statistical inference is relatively straightforward under the selection consistency of the factorial effects as ensured by Theorem 1. If forward selection correctly identifies the true, nonzero factorial effects with probability approaching one, we can proceed as if the selected working model is not data-dependent. In Section 4.1, we present the point estimators and confidence intervals for general causal parameters. In Section 4.2, we study the advantages of forward selection in terms of asymptotic efficiency in estimating general causal parameters, compared with the corresponding estimators without forward selection. We relegate the extensions to vector parameters to Section A.2 of the Supplementary Material since it is conceptually straightforward.

4.1. *Post-selection inference for general causal parameters.* Define a general causal parameter of interest as a weighted combination of average potential outcomes,

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

where $f = \{f(z)\}_{z\in\mathcal{T}}$ is a prespecified weighting vector. For example, if one is interested in estimating the main factorial effects, then f can be taken as the contrast vectors $g_{\{k\}}$ given in (2.2). If one wants to estimate interaction effects, then f can be taken as the $g_{\mathcal{K}}$ given in (2.3). However, we allow f to be different from the contrast vectors $g_{\mathcal{K}}$. For instance, if we

focus on the first two arms in factorial experiments and estimate the average treatment effect, we shall choose $f = (1, -1, 0, \ldots, 0)^\top$. In general, researchers may tailor the choice of f to the specific research questions of interest.

Without factor selection, the plug-in estimator of γ is to replace $\overline{Y}$ with its sample analogue [25, 35, 48]:

$$\widehat{\gamma} = f^\top \widehat{Y} = \sum_{z \in \mathcal{T}} f(z)\widehat{Y}(z). \tag{4.1}$$

Under regularity conditions in [35], the plug-in estimator $\widehat{\gamma}$ satisfies a central limit theorem $(\widehat{\gamma} - \gamma)/v \rightsquigarrow \mathcal{N}(0, 1)$ with the variance $v^2 = f^\top V_{\widehat{Y}} f$. When $N(z) \geq 2$, its variance can be estimated by

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

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

$$\widehat{Y}_{\mathrm{R}} = \arg \min_{\mu \in \mathbb{R}^Q} \{\|\widehat{Y} - \mu\|_2^2 : \ G(\cdot, \widehat{\mathbb{M}}^c)^\top \mu = 0\}, \tag{4.2}$$

which leverages the information that the nuisance effects $G(\cdot, \widehat{\mathbb{M}}^c)^\top \overline{Y}$ are all zero. The $\widehat{Y}_{\mathrm{R}}$ in (4.2) has a closed-form solution:

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

To show this, due to the orthogonality of G, we have the following decomposition:

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

By the constraint in (4.2), we have

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

which is minimized at

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

$\widehat{\mu}$ satisfies the constraint in (4.2), and thus is the solution for (4.2). Using this expression, we conclude that if $\widehat{\mathbb{M}} = \mathbb{M}^\star$ almost surely, then $\mathbb{E}\{\widehat{Y}_{\mathrm{R}}\} = \overline{Y}$.

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

$$f[\mathbb{M}] = Q^{-1} G(\cdot, \mathbb{M}) G(\cdot, \mathbb{M})^\top f \tag{4.3}$$

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

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

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

$$[\widehat{\gamma}_{\mathrm{R}} \pm z_{1-\alpha/2} \cdot \widehat{v}_{\mathrm{R}}], \tag{4.4}$$

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

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

4.2. *Theoretical properties under selection consistency.* In this subsection, we first present the asymptotic normality result for $\widehat{\gamma}_{\mathrm{R}}$. To simplify the discussion, we denote $f^\star = f[\mathbb{M}^\star]$. Given $\mathbb{M}^\star$ is the true working model, we have $(f^\star)^\top \overline{Y} = f^\top \overline{Y}$, for all $f \in \mathbb{R}^Q$, because

$$\begin{aligned} f^\top \overline{Y} &= f^\top \{Q^{-1} G(\cdot, \mathbb{M}^\star) G(\cdot, \mathbb{M}^\star)^\top + Q^{-1} G(\cdot, \mathbb{M}^{\star c}) G(\cdot, \mathbb{M}^{\star c})^\top\} \overline{Y} \\ &\quad \text{(due to the orthogonality of } G\text{)} \\ &= (f^\star)^\top \overline{Y} + G(\cdot, \mathbb{M}^{\star c}) \tau(\mathbb{M}^{\star c}) \text{ (definition of } f^\star \text{ based on (4.3))} \\ &= (f^\star)^\top \overline{Y}. \text{ (using } \tau(\mathbb{M}^{\star c}) = 0\text{).} \end{aligned}$$

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

THEOREM 2 (Statistical properties of $\widehat{\gamma}_{\mathrm{R}}$ and $\widehat{v}_{\mathrm{R}}^2$). *Let $N \to \infty$. Assume Conditions 1–4. We have*

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

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

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

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

Theorem 2 above guarantees that the proposed confidence interval in (4.4) for γ attains the nominal coverage probability asymptotically. Below we provide some detailed discussion.

First, the asymptotic regime of Theorem 2 is that $N \to \infty$, which is equivalent to $QN_0 \to \infty$ based on Condition 1. This covers the classical regime where Q is fixed and N_0 converges to ∞. In this regime, enough replications within each arm guarantee that the point estimates for all arms converge jointly to a multivariate normal distribution and that the variance estimators converge in probability. However, when N_0 is small but Q diverges, the joint normality fails. In this case, the asymptotic properties of the point estimates and variance estimators are guaranteed by pooling the outcomes from a large number of treatment combinations due to large Q. Interestingly, both small and large Q regimes are unified by the finite sample probability results provided in Section B.1.

Second, we discuss the condition $\|f^\star\|_\infty = O(Q^{-1})$. The condition requires each element of $f^\star$ has order Q^{-1}, which averages the outcome information over Q treatment combinations. Averaging outcomes across different treatment combinations is especially important for guaranteeing the convergence of the point estimates and variance estimates when Q diverges. This condition holds under many settings and is motivated by some specific examples of f. One special case is that $f = Q^{-1} g_{\mathcal{K}}$ for some $\mathcal{K} \in \mathbb{M}^\star$, which gives $\|f^\star\|_\infty = Q^{-1}$. As another special case, when $f = (1, 0, \ldots, 0)^\top$, we have $\|f^\star\|_\infty = Q^{-1}|\mathbb{M}^\star|$ and the condition holds when $|\mathbb{M}^\star|$ has constant order. This example is important in the application of best arm identification, which we shall discuss in the Supplementary Material A.4.

Third, Theorem 2 allows us to compare the conditions for reaching asymptotic normality of $\widehat{\gamma}$, which we formalize in the following remark.

REMARK 3 (Comparison of conditions for asymptotic normality). Without factor selection, the simple plug-in estimator $\widehat{\gamma}$ in (4.1) satisfies a central limit theorem if

$$N_0^{-1/2} \cdot \frac{\|f\|_\infty}{\|f\|_2} \to 0 \tag{4.5}$$

recalling the definition of N_0 in Condition 1 (See Theorem 1 of Shi and Ding [35]). Condition (4.5) fails when N_0 is small and f is sparse. Besides, it does not incorporate the sparsity information in the structure of factorial effects. With factor selection, however, we can borrow the benefit of a sparse working model and overcome the above drawbacks. Therefore, factor selection broadens the applicability of our proposed estimator $\widehat{\gamma}_{\mathrm{R}}$ by weakening the assumptions for the Wald-type inference.

Fourth, we elaborate on the benefits of conducting forward factorial selection in terms of asymptotic efficiency. We compare the asymptotic variances of $\widehat{\gamma}$ and $\widehat{\gamma}_{\mathrm{R}}$ in Proposition 1 below. In the most general setup, there is no ordering relationship between v_{R}^2 and v^2. That is, the RLS-based estimator may have a higher variance than the unrestricted OLS estimator. This is a known fact due to heteroskedasticity and the sandwich variance [32, 48]. Nevertheless, in many interesting scenarios, we can prove an improvement in efficiency by factor selection. Two conditions are summarized in Proposition 1.

PROPOSITION 1 (Asymptotic relative efficiency comparison between $\widehat{\gamma}$ and $\widehat{\gamma}_{\mathrm{R}}$). *Assume that both $\widehat{\gamma}$ and $\widehat{\gamma}_{\mathrm{R}}$ converge to normal distributions with variances v^2 and v_{R}^2 as $N \to \infty$.*

(i) *If the covariance matrix $V_{\widehat{Y}}$ is $\sigma^2 I_Q$, then*

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

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

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

Proposition 1(i) gives a sufficient condition assuming that the potential outcomes are homoskedastic and uncorrelated. Proposition 1(ii) studies a general heteroskedastic setting with sparse weighting vector f and small working model size $|\mathbb{M}^\star|$. The condition number $\kappa(V_{\widehat{Y}})$ captures the variability of the variances of $\widehat{Y}(z)$ across multiple treatment combination groups in $\mathcal{T}$. When the variability is upper bounded by $\kappa(V_{\widehat{Y}}) < Q/(s^\star|\mathbb{M}^\star|)$, the RLS-based estimator is more efficient than $\widehat{\gamma}$. As an application, in Section A.4 we use Proposition 1(ii) to solve the problem of inferring the best arm in factorial experiments. Moreover, we emphasize that Proposition 1 only provides two sufficient conditions, and $\widehat{\gamma}_{\mathrm{R}}$ can be more efficient than $\widehat{\gamma}$ even if those conditions fail.

There are also interesting examples that demonstrate the advantage of factor selection but are not covered by Proposition 1. One concrete problem of interest is testing the *sharp null hypothesis* of zero effects in uniform factorial designs (with N_0 replications in each arm), that is,

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

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

$$V_{\widehat{Y}} = N_0^{-1}\sigma^2 \cdot I_Q - N^{-1}\sigma^2 \mathbf{1}_Q \mathbf{1}_Q^\top,$$

where $\sigma^2 = (N-1)^{-1}\sum_{i=1}^N (Y_i - \overline{Y}^{\mathrm{obs}})^2$ and $\overline{Y}^{\mathrm{obs}} = N^{-1}\sum_{i=1}^N Y_i$. We can verify that $V_{\widehat{Y}}$ has eigenvalue decomposition

$$V_{\widehat{Y}} = N_0^{-1}\sigma^2 G \,\mathrm{Diag}\{0, 1, \ldots, 1\} G^\top$$

where G is the contrast matrix (2.4). Recalling that $\mathbb{K}$ is the full working model. The variances of the two estimators are

$$v^2 = N_0^{-1}\sigma^2 f^\top G(\cdot, \mathbb{K})G(\cdot, \mathbb{K})^\top f,$$
$$v_{\mathrm{R}}^2 = N_0^{-1}\sigma^2 f^\top G\big(\cdot, \mathbb{M}^\star\big)G\big(\cdot, \mathbb{M}^\star\big)^\top f.$$

This implies that $v^2 \geq v_{\mathrm{R}}^2$ because we always have

$$G(\cdot, \mathbb{K})G(\cdot, \mathbb{K})^\top \succcurlyeq G\big(\cdot, \mathbb{M}^\star\big)G\big(\cdot, \mathbb{M}^\star\big)^\top.$$

Therefore, under $\mathrm{H}_{0\mathrm{F}}$, the proposed RLS-based estimator $\widehat{\gamma}_{\mathrm{R}}$ is more efficient than the plug-in estimator $\widehat{\gamma}$.

Moreover, we can extend Proposition 1 to compare the length of the confidence intervals as well. The conclusion is similar. See Proposition S1 in the Supplementary Material for the details.

5. Post-selection inference under inconsistent selection. Similar to many other consistency results for variable selection, the selection consistency property in Theorem 1 can be difficult to achieve due to the strong regularity conditions. This is because the selection consistency property of forward selection requires the nonzero effects to be well separated from zero. Such a theoretical requirement can be stringent for higher-order factorial effects. In other words, as implied by the hierarchy principle, while main factorial effects and lower-order factorial effects are more likely to have nonnegligible effect sizes, higher-order factorial effects tend to have smaller effect sizes. The selection consistency property is less likely to hold for higher-order effects. More rigorously, when Condition 2(i) is violated, Algorithm 1 may no longer enjoy the selection consistency property.

Statistical inference without selection consistency is not a trivial problem in factorial designs. If we do not impose any restrictions on the factorial selection procedure, the selected model can be arbitrary, even without a stable limit. Classical strategies for post-selection inference [23] have various drawbacks in our current setup. For example, data splitting [40] is a widely used strategy to ensure valid inference after variable selection due to its simplicity. However, it relies on the independent sampling assumption, which is violated in our setting. Also, data splitting faces the conceptual challenge of making inference of a random parameter. Alternatively, selective inference [14] is another widely studied strategy, which can deliver valid inference for data-dependent parameters. However, it cannot be directly applied to analyze data collected in factorial designs, because the selective inference strategy often relies on specific selection methods and parametric modeling assumptions on the outcome.

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

5.1. *Two strategies for inconsistent selection and statistical inference.* We propose two strategies based on the assumption that selection consistency is more plausible for selecting the main factorial effects and lower-order factorial effects up to level $d^\star$ than the higher-order effects. We will add more discussion on $d^\star$ after presenting these two strategies.

In certain research questions, higher-order interactions are nuisance parameters, or there is domain knowledge indicating that higher-order interactions are negligible. Then we can stop our forward selection procedure in Algorithm 1 at $d = d^* < D$ (instead of $d = D$). Such a strategy focuses on recovering a targeted working model $\underline{\mathbb{M}}^\star$ up to level $d^\star$, that is,

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

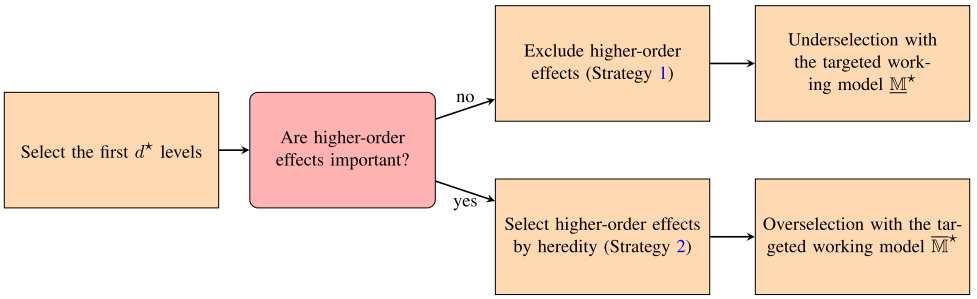

FIG. 1. *Two strategies for factorial selection*: *Strategy* 1 *underselects whereas Strategy* 2 *overselects*, *depending on whether higher-order effects are important or not*.

which leads to an underselected working model. We summarize this strategy below.

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

In Strategy 1, $\alpha_d = \infty$ leads to never rejecting the null hypothesis of zero effects, which excludes the higher-order interactions. Given the selected working model $\widehat{\underline{\mathbb{M}}}$, we can construct an estimator of $\gamma = f^\top \overline{Y}$ (defined in Section 4.1) based on RLS:

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

For Strategy 2, rather than excluding all higher-order interactions with negligible effects, we may further leverage the heredity principle and continue our selection procedure beyond level $d^\star$. This means that instead of selecting the higher-order interactions via marginal t-test and Bonferroni correction, we select the higher-order interaction terms whenever either all of their parent effects are selected (strong heredity), or one of their parent effects is selected (weak heredity). Such a strategy takes higher-order factorial effects into account and targets a working model $\overline{\mathbb{M}}^\star$:

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

where $\mathrm{H}^{(d-d^\star)}(\cdot)$ means applying the $\mathrm{H}(\cdot)$ operator $(d - d^\star)$ times. This strategy is expected to yield an overselected model that includes $\mathbb{M}^\star$. We summarize this strategy as follows.

STRATEGY 2 (Overselection by including higher-order interactions through the heredity principle). *In Algorithm* 1, *set* $\alpha_d = 0$, $d \geq d^\star + 1$ *and apply a heredity principle* (*either weak or strong*, *depending on the knowledge of the structure of the effects*). *Then the higher-order effects beyond level* $d^\star$ *are selected merely by the heredity principle and*

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

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

In Strategy 2, $\alpha_d = 0$ means always rejecting the null hypothesis of zero effect size, which corresponds to always including higher-order interactions. Given the selected working model

$\widehat{\mathbb{M}}$, we can construct an estimator of $\gamma = f^\top \overline{Y}$ based on RLS:

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

In real-world factorial experiments, how should practitioners choose between Strategy 1 and Strategy 2? This relies on the domain knowledge and the research question of interest. Strategy 1 is more suitable when there is domain knowledge that higher-order interactions are negligible, or when the research question only involves lower-order factorial effects. Moreover, Strategy 1 is more suitable when the number of active lower-order interactions is large and Strategy 2 cannot be applied. Meanwhile, Strategy 2 works better when domain knowledge suggests nonnegligible higher-order interactions or the research question targets a more general parameter beyond the factorial effects. Strategy 2 also works well when the number of active lower-order interactions is small, and we can include all the corresponding higher-order terms according to the heredity principle.

Another component that appears in Strategies 1 and 2 is the parameter $d^\star$. In practice, $d^\star$ should be determined by the research question of interest as well as the domain knowledge. A usual choice is $d^\star = 2$, because many research questions typically involve only main effects and two-way interactions. This also reflects the common practice in analyzing factorial experiments, which is assuming away all the higher-order interactions beyond the two-way effects [10, 20, 48]. This is usually supported by the domain knowledge that higher-order interaction signals beyond level $d^\star = 2$ are weak. In general, it is an interesting question to propose some data-adaptive procedure for selecting $d^\star$. We save this as a future effort.

In the following subsection, we study the statistical properties of $(\widehat{\gamma}_{\mathrm{RO}}, \widehat{v}_{\mathrm{RO}})$ and $(\widehat{\gamma}_{\mathrm{RU}}, \widehat{v}_{\mathrm{RU}})$ and demonstrate the trade-offs between the two strategies.

5.2. *Theoretical properties under inconsistent selection.* Throughout this subsection, we discuss the scenario where selection consistency is hard to achieve. We relax Condition 2 as follows.

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

Condition 5 no longer imposes any restriction on the order of the parameters beyond level $d^\star$. Diving into the proof of Theorem 1, we can see that Condition 5 guarantees that Algorithm 1 perfectly screens the first $d^\star$ levels of factorial effects in the sense that $\mathrm{P}\{\widehat{\mathbb{M}}_d = \mathbb{M}^\star_d \text{ for } d = 1, \ldots, d^\star\} \to 1$.

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

THEOREM 3 (Guarantee for Strategy 1). *Recall (4.3) and define $\underline{f}^\star = f[\underline{\mathbb{M}}^\star] = Q^{-1}G(\cdot, \underline{\mathbb{M}}^\star)G(\cdot, \underline{\mathbb{M}}^\star)^\top f$. Assume Conditions 1, 3, 4, 5 and f satisfies the following orthogonality condition*:

$$G\big(\cdot, \mathbb{M}^\star_d\big)^\top f = 0 \quad \textit{for } d^\star + 1 \le d \le K. \tag{5.1}$$

Let $N \to \infty$. We have

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

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

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

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

Now we discuss the key condition (5.1) in Theorem 3. The orthogonality condition in (5.1) restricts the weighting vector f to be orthogonal to the higher-order contrasts. Intuitively, because the higher-order interactions are excluded from the model, making inference on a weighted combination of those excluded interactions infeasible. One set of weighting vectors satisfying (5.1) is the contrast vectors of nonzero canonical lower-order interactions, given by $f = g_{\mathcal{K}}$ for some $\mathcal{K} \in \bigcup_{d=1}^{d^\star} \mathbb{M}_d^\star$. In large K settings, the number of lower-order interactions grows polynomially fast in K, which leads to difficulty for interpretation. As an example, when $K = 10$, for the first two levels of factorial effects without selection, there are more than 50 estimates in total. It can still greatly benefit the analysis and interpretation to filter out the insignificant ones and obtain a parsimonious working model.

Without the condition in (5.1), Strategy 1 could lead to biased estimates when nonzero higher-order interactions are excluded. An inconsistent model $\mathbb{M}_{\mathrm{im}}$ would miss a set of true effects $\mathbb{M}_{\mathrm{miss}} = \mathbb{M}^\star \backslash \mathbb{M}_{\mathrm{im}}$ and lead to the bias:

$$\text{Bias} = Q^{-1} f^\top G(\cdot, \mathbb{M}_{\mathrm{miss}}) \tau(\mathbb{M}_{\mathrm{miss}}). \tag{5.2}$$

From (5.2), the bias is determined by two parts. The first part is the size of the missing nonzero effects, given by $\tau(\mathbb{M}_{\mathrm{miss}})$. If the excluded effects are large, then the bias will be large. The second part depends on $f^\top G(\cdot, \mathbb{M}_{\mathrm{miss}})$. If the linear coefficient vector f is closer to the span of the excluded contrasts $G(\cdot, \mathbb{M}_{\mathrm{miss}})$, the bias will also be larger.

For Strategy 2, we have the following results.

THEOREM 4 (Guarantee for Strategy 2). *Recall* (4.3) *and define* $\overline{f}^\star = f[\overline{\mathbb{M}}^\star] = Q^{-1} G(\cdot, \overline{\mathbb{M}}^\star) G(\cdot, \overline{\mathbb{M}}^\star)^\top f$. *Assume Conditions* 1, 3, 4 *and* 5. *Let* $N \to \infty$. *If* $|\overline{\mathbb{M}}^\star|/N \to 0$, *then*

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

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

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

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

There is an additional technical requirement in Theorem 4 for overselection: $|\overline{\mathbb{M}}^\star|/N \to 0$, which is a sufficient condition for the central limit theorem. The reason is that we need to control the size of the target model $\overline{\mathbb{M}}^\star$ compared with the sample size N to infer a general causal parameter. Nevertheless, when we focus on some specific parameters such as factorial effects, this condition is not necessary.

Both Strategies 1 and 2 have advantages and disadvantages:

- *Estimation bias.* Underselection can induce bias for certain weighting vectors (quantified by (5.2)) while overselection still ensures consistency.

- *Variance.* The constructed estimator typically enjoys a smaller variance with underselection compared with overselection. We provide a discussion here in parallel with Proposition 1. If we consider the homoskedasticity condition that $V_{\widehat{Y}}$ equals $\sigma^2 I_Q$ for some $\sigma > 0$, we can prove $v_{\mathrm{RU}}^2 \leq v_{\mathrm{RO}}^2$. Therefore, in this case, by excluding higher-order terms and pursuing underselection, we can obtain an equal or smaller asymptotic variance compared with overselection. In general, due to heteroskedasticity, the order of v_{RU}^2 and v_{RO}^2 depends on the choice of target weighing vector f. Here, we take a sparse $f = \boldsymbol{e}_1 = (1, 0, \ldots, 0)^\top$ as an example. We can show that

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

  When the variability of $V_{\widehat{Y}}$ between treatment combinations is small in the sense that $\kappa(V_{\widehat{Y}}) < |\overline{\mathbb{M}}^\star|/|\underline{\mathbb{M}}^\star|$, underselection leads to smaller asymptotic variance for inferring $\boldsymbol{e}_1^\top \overline{Y} = \overline{Y}(1)$.
- *Interpretability.* Underselection leads to simpler models that are easier to interpret, while overselection may not be practical if there are too many nonzero lower-order terms, which can result in many redundant terms in the selected model.

If higher-order interactions are not crucial, Strategy 1 can be applied. If higher-order interactions are of interest but hard to select, Strategy 2 can be applied.

6. Simulation. In this section, we use simulation to demonstrate the finite-sample performance of the proposed forward selection framework and the inferential properties of the RLS-based estimator. Our simulation results verify the following properties of the proposed procedure and estimators:

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

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

(G1) echoes our discussion on the comparison of conditions and asymptotic variances for central limit theorems in Remark 3 and Proposition 1. **(G2)** verifies the results in Theorem 1 and 2 and checks the finite sample behaviors of the proposed procedures. For both **(G1)** and **(G2)**, we will vary the sample size and effect size.

6.1. *Simulation setup.* We set up a 2^K factorial experiment, with N_0 units in each treatment combination where K and N_0 are varied across settings. We generate independent potential outcomes from a shifted exponential distribution:

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

where $\mu(z)$ is the superpopulation mean of potential outcomes under treatment combination z. We choose $\mu(z)$ such that the factorial effects satisfy the following structure:

- Main effects: the main effects corresponding to the first five factors, $\tau_{\{1\}}, \ldots, \tau_{\{5\}}$, are nonzero; the rest of the three main effects, $\tau_{\{6\}}, \ldots, \tau_{\{8\}}$, are zero.
- Two-way interactions: the two-way interactions associated with the first five factors are nonzero, that is, $\tau_{\{kl\}} \neq 0$ for $k \neq l, k, l \in [5]$. The rest of the two-way interactions are zero.
- Higher-order interactions: all the higher-order interactions $\tau_{\mathcal{K}}$ are zero if $|\mathcal{K}| \geq 3$.

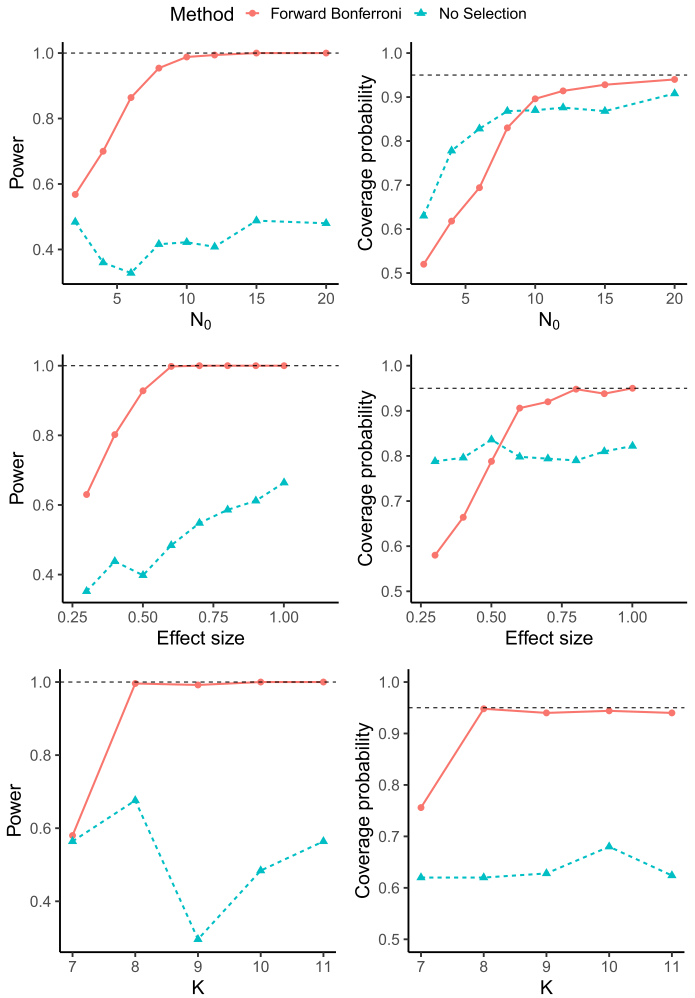

FIG. 2. *Simulation results supporting* **(G1)**. (*i*) *Top left panel*: *power curve with varying* N_0; (*ii*) *Top right panel*: *coverage probability with varying* N_0; (*iii*) *Middle left panel*: *power curve with varying effect size* γ_{target}; (*iv*) *Middle right panel*: *coverage probability with varying effect size* γ_{target}. (*v*) *Bottom left panel*: *power curve with varying number of factors* K; (*vi*) *Bottom right panel*: *coverage probability with varying number of factors* K.

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

6.2. *Simulation results supporting* **(G1)**. In this subsection, we evaluate the performance of the RLS-based estimators $(\widehat{\gamma}_{\mathrm{R}}, \widehat{v}_{\mathrm{R}})$ compared with $(\widehat{\gamma}, \widehat{v})$ for testing a causal effect $\gamma_{\text{target}} = f^{\top}\overline{Y}$ specified by a sparse vector: $f = (0, \ldots, 0, 1)^{\top} \in \mathbb{R}^{\mathcal{Q}}$. Intuitively, γ_{target} measures the average of potential outcomes in the last level. For each estimator, we report: (i) power for testing $\mathrm{H}_0 : \gamma_{\text{target}} = 0$. (ii) coverage probability of the 95% confidence intervals for γ_{target}. Figure 2 summarizes the results.

Figure 2 demonstrates that the RLS-based estimator $\widehat{\gamma}_{\mathrm{R}}$ has much higher power compared with the simple moment estimator $\widehat{\gamma}$ for inferring γ_{target}. This echoes Proposition 1 that the RLS-based estimator has reduced variance compared with the simple moment estimator. Moreover, the RLS-based estimator attains near nominal coverage probability with

reasonably large N_0 and γ_{target}, whereas the simple moment estimator tends to provide undercovered confidence intervals in all cases.

6.3. *Simulation results supporting* (***G2***). In this subsection, we compare the performance of four candidate effect selection methods:

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

For each selection method, we evaluate their performance with three measures: (i) selection consistency probability $\mathrm{P}\{\widehat{\mathbb{M}} = \mathbb{M}^\star\}$, (ii) power of $\widehat{\gamma}_{\mathrm{R}}$ for testing $\mathrm{H}_0 : \gamma_{\text{target}} = 0$ for γ_{target} defined in Section 6.2 and (iii) coverage probability of the RLS-based 95% confidence interval for γ_{target}. The results are summarized in Figure 3.

From Figure 3, all four effect selection methods lead to selection consistency with high probability as N_0 or γ_{target} increases. Nevertheless, with the forward selection procedure, the probability of selection consistency is higher than the naive selection procedure. Besides, forward selection complies with the heredity structure and demonstrates better interpretability than the naive selection methods. In terms of the power of $\widehat{\gamma}_{\mathrm{R}}$ and $\widehat{v}_{\mathrm{R}}$ for testing $\mathrm{H}_0 : \gamma_{\text{target}} = 0$, while all four methods have power approaching one as N_0 and γ_{target} increases, forward selection based procedures possess higher power with small N_0 and γ_{target}. Lastly, we can see an improvement in the coverage probability of the RLS-based confidence intervals with the forward selection procedure.

6.4. *Violations of the conditions*. In this subsection, we discussed the impact of the violation of the conditions on the performance of Algorithm 1.

Small effect size. Condition 2 assumes that the effect sizes should be of a certain order with respect to the sample size N. Small effect sizes will impact the performance of the algorithm, in terms of selection consistency probability, coverage, and power for post-selection inference, etc. For example, the middle panels in Figure 3 show how the selection consistency probability power and coverage for inference vary with effect sizes in a factorial experiment with $K = 8$ factors and $N_0 = 2$ replications on each arm. If the effect sizes are too small compared to the sample size, selection consistency is hard to achieve and post-selection inference is also hurt by the bias generated from model misspecification. Nevertheless, the issue is mitigated as the effect size increases to a larger level.

Failure of heredity. Condition 4 assumes that the factorial effects follow a weak/strong heredity structure. For effects selection, failure of effect heredity typically leads to underselection in the interaction terms, because the effects that violate heredity will be ruled out by the heredity step in Algorithm 1. For post-selection inference, failure of heredity will lead to bias for the RLS-based estimator (4.2) and impact coverage and power for hypothesis testing. If researchers are skeptical about the heredity principle in their studies, they can perform Algorithm 1 without applying the heredity step.

7. Case study: Conjoint survey experiment regarding U.S. presidential candidates. In this section, we apply the forward selection method to a real data example. In particular, we analyze a conjoint survey experiment regarding U.S. citizens' preferences across presidential candidates studied by [18]. The study focuses on how the candidates' traits impact citizens' preferences. The original experiment involves eight attributes of the imaginary candidate profiles: military service (z_1), religion (z_2), college education (z_3), annual income (z_4),

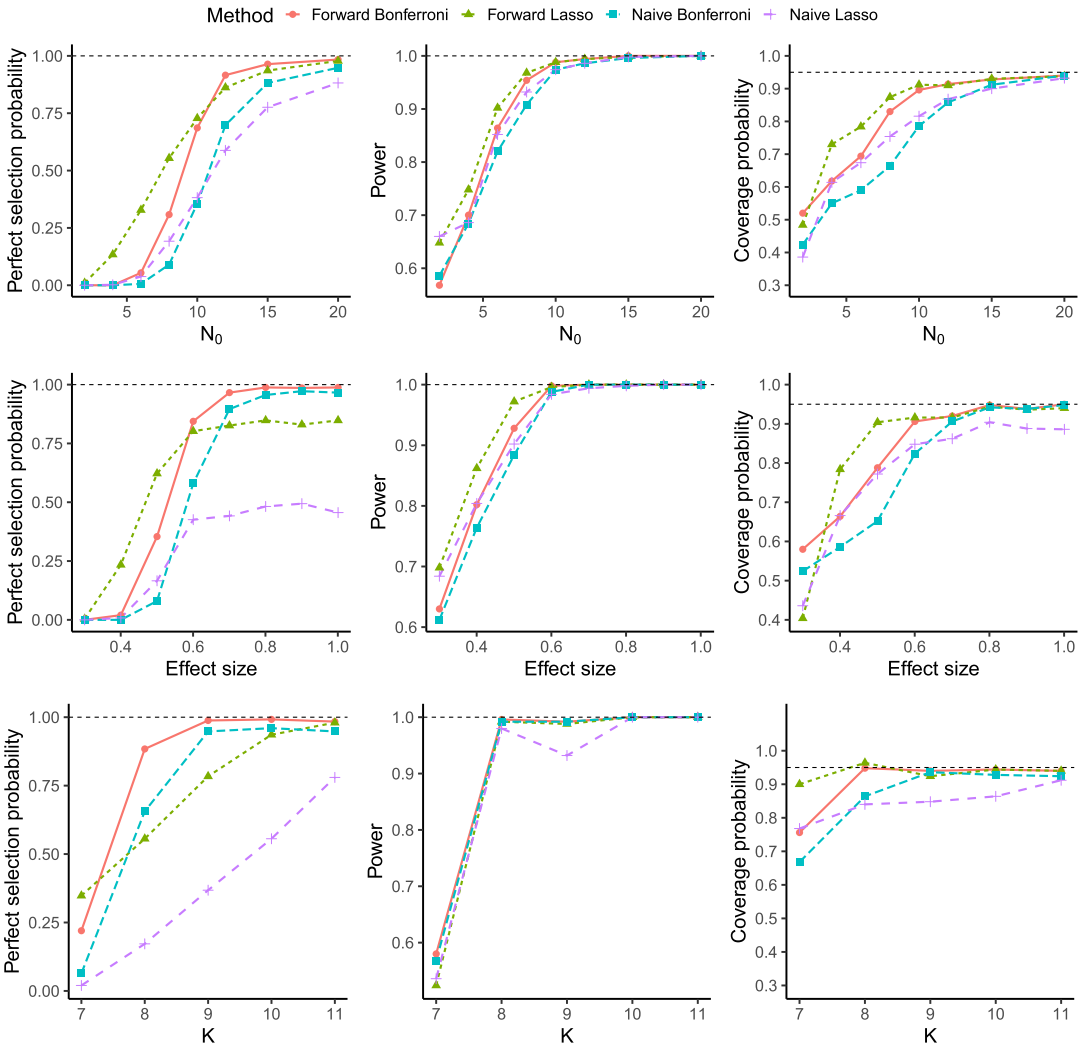

FIG. 3. *Simulation results supporting* **(G2)**. (*i*) *Top row left panel*: *selection consistency probability with a small fixed effect size* $\gamma_{target} = 0.20$, *a fixed number of factors* $K = 8$ *and varying* N_0; (*ii*) *Top row middle panel*: *power curve with a small fixed effect size* $\gamma_{target} = 0.20$, *a fixed number of factors* $K = 8$ *and varying* N_0; (*iii*) *Top row right panel*: *coverage probability with a small fixed effect size* $\gamma_{target} = 0.20$, *a fixed number of factors* $K = 8$ *and varying* N_0; (*iv*) *Middle row left panel*: *selection consistency probability with a small fixed replication* $N_0 = 2$, *a fixed number of factors* $K = 8$ *and varying effect size* γ_{target}; (*v*) *Middle row middle panel*: *power curve with a small fixed replication* $N_0 = 2$, *a fixed number of factors* $K = 8$ *and varying effect size* γ_{target}; (*vi*) *Middle row right panel*: *coverage probability with a small fixed replication* $N_0 = 2$, *a fixed number of factors* $K = 8$ *and varying effect size* γ_{target}. (*vii*) *Bottom row left panel*: *selection consistency probability with a small fixed replication* $N_0 = 2$, *an effect size* $\gamma_{target} = 0.50$ *and a varying number of factors*; (*viii*) *Bottom row middle panel*: *power curve with a small fixed replication* $N_0 = 2$, *an effect size* $\gamma_{target} = 0.50$ *and a varying number of factors*; (*ix*) *Bottom row right panel*: *coverage probability with a small fixed replication* $N_0 = 2$, *an effect size* $\gamma_{target} = 0.50$ *and a varying number of factors*.

racial/ethnic background (z_5), age (z_6), gender (z_7) and profession (z_8), where military service and gender are binary factors while the rest six factors have six levels. To fit into our framework, we drop the profession factor and dichotomize the other six-level factors into binary ones. The outcome is a rating of the candidate profile on a one-to-seven scale, representing the levels of absolute support or opposition to each profile separately. The final data set contains a total of $K = 7$ factors (with $Q = 2^7 = 128$ treatment combinations) and $N = 3456$ profiles. Each treatment combination contains 27 respondents.

TABLE 2
Working model selection results for the presidential candidate experiment based on different strategies

Selection Strategy	Selected Working Model
Forward + Strong Heredity	$\tau_2, \tau_3, \tau_4, \tau_6, \tau_7, \tau_{23}, \tau_{36}, \tau_{46}, \tau_{47}, \tau_{67}, \tau_{467}$
Forward + Weak Heredity	$\tau_2, \tau_3, \tau_4, \tau_6, \tau_7, \tau_{12}, \tau_{23}, \tau_{13}, \tau_{35}, \tau_{36}, \tau_{14}, \tau_{46}, \tau_{47}, \tau_{56}, \tau_{67}, \tau_{57}$
Forward + No Heredity	$\tau_2, \tau_3, \tau_4, \tau_6, \tau_7, \tau_{12}, \tau_{23}, \tau_{13}, \tau_{35}, \tau_{36}, \tau_{14}, \tau_{46}, \tau_{47}, \tau_{56}, \tau_{67}, \tau_{57}$
No Forward	$\tau_3, \tau_6, \tau_{14}$

We applied the forward selection procedure to analyze the data. For each level, we apply Lasso to penalize the factor-based regression and select a working model. Here, we applied Lasso instead of marginal t-tests for the convenience of tuning parameter selection based on existing packages. For the Lasso implementation, we apply cross-validation to decide the level of penalization. Between levels, we incorporate the heredity structure. For comparison, we also report the selected working model from forward selection without heredity as well as that from a full Lasso that does not proceed in a forward style. Table 2 reports the effect selection results based on these strategies.

From Table 2, we can draw the following conclusions:

- *Forward versus nonforward selection.* A forward selection procedure selects more terms into the working model, while the full Lasso procedure selects an overly sparse one. This is because the scale of the factorial effect sizes for different levels can vary, and the forward selection procedure can pick different penalization levels to adapt to the hierarchy. Besides, forward selection can incorporate heredity structure into the selected working model but the full Lasso cannot. Therefore, the forward selection procedure provides a more interpretable view of the role of the factors as well as their interactions.
- *Strong heredity, weak heredity versus no heredity.* In this application, strong heredity produces a more parsimonious working model for the two-way interactions and also discovers one three-way interaction (τ_{467}). Compared with the routine of assuming away three-way or higher-order interactions in practice, this result suggests that forward selection with heredity makes it possible to gain scientific insights beyond lower-order effects. Moreover, weak heredity also leads to a sparse and interpretable working model for two-way interactions. In this case, the result based on the forward selection with weak heredity coincides with that of forward selection with no heredity, which can be viewed as a validation for the plausibility of the heredity principle.

8. Discussion. We have discussed the theory for forward selection and post-selection inference in 2^K factorial designs. The method and theory are especially relevant when the number of factors, K, is large and diverges with the sample size. With a large K, fractional factorial designs [43] are attractive alternatives in the design stage if some higher-order interactions are absent and the designer has correct prior knowledge on them [33, 43]. The trade-off between full and fractional factorial designs is well documented: the fractional factorial design is less costly, whereas the full factorial design allows for exploring higher-order interactions. Moreover, the design-based theory for factor selection and post-selection inference for full factorial designs serves as a stepping stone for the corresponding theory for fractional factorial designs. We leave it to future research.

There are several further directions for exploration. It is conceptually straightforward to extend the theory to general factorial designs with multivalued factors under more complicated notation, and we thus omit the technical details to simplify the theoretical discussion. Another important direction is covariate adjustment in factorial experiments. Lin [27], Lu

[30] and Liu, Ren and Yang [28] demonstrated the efficiency gain of covariate adjustment with small K. Zhao and Ding [49] discussed covariate adjustment in factorial experiments with factors and covariates selected independent of data. Moreover, it is interesting to extend the framework to observational studies [33, 46] by incorporating propensity score and outcome model estimation and exploring the properties of the procedure, such as double robustness, etc. Besides, in the current work, we focus more on the analysis part instead of the design part. If we understand the properties of factor screening, then it is possible to have a better experimental design for factorial experiments.

Funding. Jingshen Wang's acknowledges the support of the U.S. National Science Foundation Grants DMS-2015325, DMS-2239047 and DMS-2220537.

Peng Ding's research is partially supported by the U.S. National Science Foundation (#1945136).

SUPPLEMENTARY MATERIAL

Supplement to "Forward selection and post-selection inference in factorial designs" (DOI: 10.1214/24-AOS2454SUPP; .pdf). The Supplementary Material contains additional results and proofs.

REFERENCES

[1] ANDREWS, I., KITAGAWA, T. and MCCLOSKEY, A. (2019). Inference on winners Technical Report, National Bureau of Economic Research.

[2] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2010). Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing Strength*: *Theory Powering Applications—a Festschrift for Lawrence D. Brown*. *Inst. Math. Stat.* (*IMS*) *Collect*. **6** 56–69. IMS, Beachwood, OH. MR2798511

[3] BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A lasso for hierarchical interactions. *Ann. Statist*. **41** 1111–1141. MR3113805 https://doi.org/10.1214/13-AOS1096

[4] BLACKWELL, M. and PASHLEY, N. E. (2023). Noncompliance and instrumental variables for 2^K factorial experiments. *J. Amer. Statist. Assoc*. **118** 1102–1114. MR4595480 https://doi.org/10.1080/01621459.2021.1978468

[5] BLONIARZ, A., LIU, H., ZHANG, C.-H., SEKHON, J. S. and YU, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proc. Natl. Acad. Sci. USA* **113** 7383–7390. MR3531136 https://doi.org/10.1073/pnas.1510506113

[6] BOX, G. E. P., HUNTER, J. S. and HUNTER, W. G. (2005). *Statistics for Experimenters*: *Design*, *Innovation*, *and Discovery*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2140250

[7] BRANSON, Z., DASGUPTA, T. and RUBIN, D. B. (2016). Improving covariate balance in 2^K factorial designs via rerandomization with an application to a New York City Department of Education high school study. *Ann. Appl. Stat*. **10** 1958–1976. MR3592044 https://doi.org/10.1214/16-AOAS959

[8] CAUGHEY, D., KATSUMATA, H. and YAMAMOTO, T. (2019). Item response theory for conjoint survey experiments Technical Report, Working Paper.

[9] DASGUPTA, T., PILLAI, N. S. and RUBIN, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **77** 727–753. MR3382595 https://doi.org/10.1111/rssb.12085

[10] EGAMI, N. and IMAI, K. (2019). Causal interaction in factorial experiments: Application to conjoint analysis. *J. Amer. Statist. Assoc*. **114** 529–540. MR3963160 https://doi.org/10.1080/01621459.2018.1476246

[11] ESPINOSA, V., DASGUPTA, T. and RUBIN, D. B. (2016). A Bayesian perspective on the analysis of unreplicated factorial experiments using potential outcomes. *Technometrics* **58** 62–73. MR3463157 https://doi.org/10.1080/00401706.2015.1006337

[12] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **70** 849–911. MR2530322 https://doi.org/10.1111/j.1467-9868.2008.00674.x

[13] FISHER, R. A. (1935). *The Design of Experiments*: *Oliver and Boyd*, 1st ed. Edinburgh, London.

[14] FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Preprint. Available at arXiv:1410.2597.

[15] FREEDMAN, D. A. (2008). On regression adjustments to experimental data. *Adv. in Appl. Math.* **40** 180–193. MR2388610 https://doi.org/10.1016/j.aam.2006.12.003

[16] GERBER, A. S. and GREEN, D. P. (2012). *Field Experiments*: *Design*, *Analysis*, *and Interpretation*, Norton, New York, NY.

[17] GUO, X., WEI, L., WU, C. and WANG, J. (2021). Sharp inference on selected subgroups in observational studies. Preprint. Available at arXiv:2102.11338.

[18] HAINMUELLER, J., HOPKINS, D. J. and YAMAMOTO, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Polit. Anal.* **22** 1–30.

[19] HAO, N., FENG, Y. and ZHANG, H. H. (2018). Model selection for high-dimensional quadratic regression via regularization. *J. Amer. Statist. Assoc.* **113** 615–625. MR3832213 https://doi.org/10.1080/01621459.2016.1264956

[20] HAO, N. and ZHANG, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **109** 1285–1301. MR3265697 https://doi.org/10.1080/01621459.2014.881741

[21] HARIS, A., WITTEN, D. and SIMON, N. (2016). Convex modeling of interactions with strong heredity. *J. Comput. Graph. Statist.* **25** 981–1004. MR3572025 https://doi.org/10.1080/10618600.2015.1067217

[22] KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York. MR0045368

[23] KUCHIBHOTLA, A. K., KOLASSA, J. E. and KUFFNER, T. A. (2022). Post-selection inference. *Annu. Rev. Stat. Appl.* **9** 505–527. MR4394918 https://doi.org/10.1146/annurev-statistics-100421-044639

[24] LI, W., NACHTSHEIM, C. J., WANG, K., REUL, R. and ALBRECHT, M. (2013). Conjoint analysis and discrete choice experiments for quality improvement. *J. Qual. Technol.* **45** 74–99.

[25] LI, X. and DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* **112** 1759–1769. MR3750897 https://doi.org/10.1080/01621459.2017.1295865

[26] LIM, M. and HASTIE, T. (2015). Learning interactions via hierarchical group-lasso regularization. *J. Comput. Graph. Statist.* **24** 627–654. MR3397226 https://doi.org/10.1080/10618600.2014.938812

[27] LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420 https://doi.org/10.1214/12-AOAS583

[28] LIU, H., REN, J. and YANG, Y. (2024). Randomization-based joint central limit theorem and efficient covariate adjustment in randomized block 2^K factorial experiments. *J. Amer. Statist. Assoc.* **119** 136–150. MR4713881 https://doi.org/10.1080/01621459.2022.2102985

[29] LU, J. (2016). On randomization-based and regression-based inferences for 2^K factorial designs. *Statist. Probab. Lett.* **112** 72–78. MR3475490 https://doi.org/10.1016/j.spl.2016.01.010

[30] LU, J. (2016). Covariate adjustment in randomization-based causal inference for 2^K factorial designs. *Statist. Probab. Lett.* **119** 11–20. MR3555264 https://doi.org/10.1016/j.spl.2016.07.010

[31] LUCE, R. D. and TUKEY, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *J. Math. Psych.* **1** 1–27.

[32] MENG, X.-L. and XIE, X. (2014). I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? *Econometric Rev.* **33** 218–250. MR3170847 https://doi.org/10.1080/07474938.2013.808567

[33] PASHLEY, N. E. and BIND, M.-A. C. (2023). Causal inference for multiple treatments using fractional factorial designs. *Canad. J. Statist.* **51** 444–468. MR4595237 https://doi.org/10.1002/cjs.11734

[34] SHAO, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. MR1466682

[35] SHI, L. and DING, P. (2022). Berry–Esseen bounds for design-based causal inference with possibly diverging treatment levels and varying group sizes. Preprint. Available at arXiv:2209.12345.

[36] SHI, L., WANG, J. and DING, P. (2025). Supplement to "Forward selection and post-selection inference in factorial designs." https://doi.org/10.1214/24-AOS2454SUPP

[37] SPLAWA-NEYMAN, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. MR1092986

[38] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[39] WANG, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.* **104** 1512–1524. MR2750576 https://doi.org/10.1198/jasa.2008.tm08516

[40] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689 https://doi.org/10.1214/08-AOS646

[41] WEI, W., ZHOU, Y., ZHENG, Z. and WANG, J. (2024). Inference on the best policies with many covariates. *J. Econometrics* **239** Paper No. 105460, 14. MR4708623 https://doi.org/10.1016/j.jeconom.2022.06.013

[42] WIECZOREK, J. and LEI, J. (2022). Model selection properties of forward selection and sequential cross-validation for high-dimensional regression. *Canad. J. Statist.* **50** 454–470. MR4429847 https://doi.org/10.1002/cjs.11635

[43] WU, C. F. J. and HAMADA, M. S. (2009). *Experiments*: *Planning*, *Analysis*, *and Optimization*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2583259

[44] WU, Y., ZHENG, Z., ZHANG, G., ZHANG, Z. and WANG, C. (2022). Non-stationary a/b tests: Optimal variance reduction, bias correction, and valid inference. Bias Correction, and Valid Inference (May 20, 2022).

[45] YATES, F. (1937). The design and analysis of factorial experiments Technical Report No. Technical Communication 35, Imperial Bureau of Soil Science, London, U. K.

[46] YU, R. and DING, P. (2023). Balancing weights for causal inference in observational factorial studies. Preprint. Available at arXiv:2310.04660.

[47] YUAN, M., JOSEPH, V. R. and LIN, Y. (2007). An efficient variable selection approach for analyzing designed experiments. *Technometrics* **49** 430–439. MR2414515 https://doi.org/10.1198/004017007000000173

[48] ZHAO, A. and DING, P. (2022). Regression-based causal inference with factorial experiments: Estimands, model specifications and design-based properties. *Biometrika* **109** 799–815. MR4472849 https://doi.org/10.1093/biomet/asab051

[49] ZHAO, A. and DING, P. (2023). Covariate adjustment in multiarmed, possibly factorial experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 1–23. MR4726956 https://doi.org/10.1093/jrsssb/qkac003

[50] ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. MR2549566 https://doi.org/10.1214/07-AOS584

[51] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

[52] ZHAO, S., WITTEN, D. and SHOJAIE, A. (2021). In defense of the indefensible: A very naïve approach to high-dimensional inference. *Statist. Sci.* **36** 562–577. MR4323053 https://doi.org/10.1214/20-sts815

[53] ZHIRKOV, K. (2022). Estimating and using individual marginal component effects from conjoint experiments. *Polit. Anal.* **30** 236–249.