Bai, Yuehao; Liu, Jizhou; Tabord-Meehan, Max

Article

Inference for matched tuples and fully blocked factorial designs

Quantitative Economics

Provided in Cooperation with:
The Econometric Society

Suggested Citation: Bai, Yuehao; Liu, Jizhou; Tabord-Meehan, Max (2024) : Inference for matched tuples and fully blocked factorial designs, Quantitative Economics, ISSN 1759-7331, The Econometric Society, New Haven, CT, Vol. 15, Iss. 2, pp. 279-330, https://doi.org/10.3982/QE2354

This Version is available at:
https://hdl.handle.net/10419/320307

Quantitative Economics 15 (2024), 279–330 1759-7331/20240279

Inference for matched tuples and fully blocked factorial designs

Yuehao Bai
Department of Economics, University of Southern California

Jizhou Liu
Booth School of Business, University of Chicago

Max Tabord-Meehan
Department of Economics, University of Chicago

This paper studies inference in randomized controlled trials with multiple treatments, where treatment status is determined according to a "matched tuples" design. If there are $|\mathcal{D}|$ possible treatments, then by a matched tuples design, we mean an experimental design where units are sampled i.i.d. from the population of interest, grouped into "homogeneous" blocks of size $|\mathcal{D}|$, and finally, within each block, exactly one individual is randomly assigned to each of the $|\mathcal{D}|$ treatments. We first study estimation and inference for matched tuples designs in the general setting where the parameter of interest is a vector of linear contrasts over the collection of average potential outcomes for each treatment. Parameters of this form include standard average treatment effects used to compare one treatment relative to another, but also include parameters that may be of interest in the analysis of factorial designs. We first establish conditions under which a sample analog estimator is asymptotically normal and construct a consistent estimator of its corresponding asymptotic variance. Combining these results establish the asymptotic exactness of tests based on these estimators. In contrast, we show that, for two common testing procedures based on t-tests constructed from linear regressions, one test is generally conservative while the other is generally invalid. We go on to apply our results to study the asymptotic properties of what we call "fully-blocked" 2^K factorial designs, which are simply matched tuples designs applied to a full factorial experiment. Leveraging our previous results, we establish that our estimator achieves a lower asymptotic variance under the fully-blocked design than that under any stratified factorial design, which stratifies the experimental sample into a finite number of "large" strata. A simulation study and empirical application illustrate the practical relevance of our results.

Yuehao Bai: yuehao.bai@usc.edu
Jizhou Liu: jliu32@chicagobooth.edu
Max Tabord-Meehan: maxtm@uchicago.edu

We thank the anonymous referees, as well as seminar participants at Columbia University, Duke University, Indiana University, Michigan State University, Penn State University, UCLA, University of Pennsylvania, University of Pittsburgh, University of Southern California, UW Milwaukee, and Yale University for helpful comments. We thank Jiehan Xu for excellent research assistance. The third author acknowledges support from NSF grant SES-2149408.

 https://doi.org/10.3982/QE2354

Keywords. Randomized controlled trials, matched tuples, matched pairs, multiple treatments, factorial designs.

JEL classification. C12, C14.

1. Introduction

This paper studies inference in randomized controlled trials with multiple treatments, where treatment status is determined according to a "matched tuples" design. If there are $|\mathcal{D}|$ possible treatments, then by a matched tuples design, we mean an experimental design where units are sampled i.i.d. from the population of interest, grouped into "homogeneous" blocks of size $|\mathcal{D}|$, and finally, within each block, exactly one individual is randomly assigned to each of the $|\mathcal{D}|$ treatments. As such, matched tuples designs generalize the concept of matched pairs designs to settings with more than two treatments. Matched tuples designs are commonly used in the social sciences; see Bold, Kimenyi, Mwabu, Ng'ang'a, and Sandefur (2018), Brown and Andrabi (2020), de Mel et al. (2013), and Fafchamps, McKenzie, Quinn, and Woodruff (2014) for examples in economics, and are often motivated using the simulation evidence presented in Bruhn and McKenzie (2009). However, we are not aware of any formal results that establish valid asymptotically exact methods of inference for matched tuples designs. Accordingly, in this paper we establish general results about estimation and inference for matched tuples designs, and then apply these results to study the asymptotic properties of what we call "fully-blocked" 2^K factorial designs.

We first study estimation and inference for matched tuples designs in the general setting where the parameter of interest is a vector of linear contrasts over the collection of average outcomes for each treatment. Parameters of this form include standard average treatment effects (ATEs) used to compare one treatment relative to another, but as we explain below also include more complicated parameters that may be of interest, for instance, in the analysis of factorial designs. We first establish conditions under which a sample analog estimator is asymptotically normal and construct a consistent estimator of its corresponding asymptotic variance. Combining these results establish the asymptotic validity of tests based on these estimators. We then consider the asymptotic properties of two commonly recommended inference procedures. The first is based on a linear regression with block fixed effects. Importantly, we find the t-test based on such a regression is in general not valid for testing the null hypothesis that a pairwise ATE is equal to a prespecified value. The second is based on a linear regression with cluster-robust standard errors, where clusters are defined at the block level. Here, we find that the corresponding t-test is generally valid but conservative, and that this conservativeness increases in the number of treatments.

Next, we apply our results to study the asymptotic properties of "fully-blocked" 2^K factorial designs. Factorial designs are classical experimental designs (see Wu and Hamada (2011) for a textbook treatment), which are increasingly being used in the social sciences (see, for instance, Alatas, Banerjee, Hanna, Olken, and Tobias (2012), Besedeš, Deck, Sarangi, and Shor (2012), DellaVigna, List, Malmendier, and Rao (2016), Kaur, Kremer, and Mullainathan (2015), Karlan, Osei, Osei-Akoto, and Udry (2014)). In a 2^K factorial design, each treatment is a combination of multiple "factors," where each factor

can take two distinct values, or "levels." As a consequence, a full 2^K factorial design can be thought of as a randomized experiment with 2^K distinct treatments (importantly, however, the analysis of factorial designs typically considers factorial effects as the parameters of interest; see Section 3.3 for a definition). A fully-blocked factorial design is then simply a matched tuples design with blocks of size 2^K. Leveraging our previous results, we establish that our estimator achieves a lower asymptotic variance under the fully-blocked design than under any stratified factorial design, which stratifies the experimental sample into a finite number of "large" strata (such designs include complete randomization as a special case). We also consider settings where only one factor may be of primary interest, and establish that even in such cases it is more efficient to perform a fully-blocked design than to perform a matched pairs design that exclusively focuses on the primary factor of interest.

In a simulation study, we find that although our inference results are asymptotically exact, our proposed tests may be conservative in finite samples when the experiment features many treatments or many blocking variables. Accordingly, we also study the behavior of a matched tuples design with "replicates," where we form blocks of size *two* times the number of treatments, and each treatment is assigned exactly *twice* at random within each block. Although we find that such a design results in an estimator with slightly larger mean squared error, the rejection probabilities of our proposed tests become much closer to the nominal level, which may result in improved power. Further discussion is provided in Section 3.2 below.

Although the analysis of matched tuples designs has to our knowledge not received much attention, there are large literatures on both the analysis of matched pairs designs and the analysis of factorial designs. Recent papers, which have analyzed the properties of matched pairs designs, include Athey and Imbens (2017), Bai, Romano, and Shaikh (2021), Bai (2022), de Chaisemartin and Ramirez-Cuellar (2022), Cytrynbaum (2021), Imai, King, and Nall (2009), Jiang, Liu, Phillips, and Zhang (2020), Fogarty (2018), and van der Laan, Balzer, and Petersen (2012). Our analysis builds directly on the framework developed in Bai, Romano, and Shaikh (2021), and our Theorems 3.1 and 3.2 nest some of their results when specialized to the setting of a binary treatment. Cytrynbaum (2021) considers a generalization of matched pairs designs, a special case of which he refers to as a matched tuples design. However, his design groups units into homogeneous blocks in order to assign a binary treatment with unequal treatment fractions. In contrast, we consider a setting where units are grouped into homogeneous blocks in order to assign multiple treatments.

Recent papers which have analyzed factorial designs include Branson, Dasgupta, and Rubin (2016), Dasgupta, Pillai, and Rubin (2015), Li, Ding, and Rubin (2020), Muralidharan, Romero, and Wüthrich (2019), Pashley and Bind (2019), and Liu, Ren, and Yang (2022). Our setup and notation for 2^K factorial designs mirrors the framework introduced in Dasgupta, Pillai, and Rubin (2015), although our setup differs in that we consider a "super-population" framework where potential outcomes are modeled as random, whereas they maintain a finite population framework where potential out-

comes are modeled as fixed.[1] Borrowing the framework from Dasgupta, Pillai, and Rubin (2015), Branson, Dasgupta, and Rubin (2016) and Li, Ding, and Rubin (2020) propose rerandomization designs for factorial experiments, which are shown to have favorable efficiency properties relative to a completely randomized design. Although we do not provide formal results comparing our fully-blocked design to these rerandomization designs, our simulation evidence suggests that, at least in the inferential framework considered in our paper, the fully-blocked design can improve efficiency relative to these rerandomization designs. Also, closely related to our paper is Liu, Ren, and Yang (2022), who extend the results in Dasgupta, Pillai, and Rubin (2015) to general stratified randomized designs. Their results on variance estimation specifically exclude the setting where each treatment is assigned exactly once per block, which is the primary setting that we consider in this paper.

The rest of the paper is organized as follows. In Section 2, we describe our setup and notation. Section 3 presents the main results. In Section 4, we examine the finite sample behavior of various experimental designs via simulation in the context of 2^K factorial experiments. Finally, in Section 5 we illustrate our proposed inference methods in an empirical application based on the experiment conducted in Fafchamps et al. (2014). We conclude with recommendations for empirical practice in Section 6.

2. Setup and notation

Let $Y_i \in \mathbf{R}$ denote the observed outcome of interest for the ith unit. Let $D_i \in \mathcal{D}$ denote treatment status for the ith unit, where $\mathcal{D}$ denotes a finite set of values of the treatment. We assume $\mathcal{D} = \{1, \ldots, |\mathcal{D}|\}$. Generally, we use $D_i = 1$ to indicate the ith unit is untreated, but such a restriction is not necessary for our results. Let X_i denote the observed baseline covariates for the ith unit, and denote its dimension by $\dim(X_i)$. For $d \in \mathcal{D}$, let $Y_i(d)$ denote the potential outcome for the ith unit if its treatment status were d. The observed outcome and potential outcomes are related to treatment status by the expression

$$Y_i = \sum_{d \in \mathcal{D}} Y_i(d) I\{D_i = d\}. \tag{1}$$

We suppose our sample consists of $J_n := (|\mathcal{D}|)n$ i.i.d. units. For any random variable indexed by i, for example D_i, we denote by $D^{(n)}$ the random vector $(D_1, D_2, \ldots, D_{J_n})$. Let P_n denote the distribution of the observed data $Z^{(n)}$ where $Z_i = (Y_i, D_i, X_i)$, and Q_n denote the distribution of $W^{(n)}$, where $W_i = (Y_i(1), Y_i(2), \ldots, Y_i(|\mathcal{D}|), X_i)$. We assume that $W^{(n)}$ consists of J_n i.i.d observations, so that $Q_n = Q^{J_n}$, where Q is the marginal distribution of W_i. Given Q_n, P_n is then determined by (1) and the mechanism for determining treatment assignment. We thus state our assumptions in terms of assumptions on Q and the treatment assignment mechanism.

[1] The finite population "design-based" perspective may be particularly attractive in settings where the experimental sample is not explicitly drawn from a larger population. In Appendix D.2 in the Supplemental Appendix (Bai, Liu, and Tabord-Meehan (2024)), we provide some preliminary simulation evidence that our proposed estimators may be relevant in such a setting as well; however, given the simulation evidence in de Chaisemartin and Ramirez-Cuellar (2022) and our currently incomplete understanding of the design-based properties of our estimators, we do not make any general claims in this paper.

Our object of interest will generically be defined as a vector of linear contrasts over the collection of expected potential outcomes across treatments. Formally, let

$$\Gamma(Q) := \big(\Gamma_1(Q), \ldots, \Gamma_{|\mathcal{D}|}(Q)\big)',$$

where $\Gamma_d(Q) := E_Q[Y_i(d)]$ for $d \in \mathcal{D}$. Let ν be a real-valued $m \times |\mathcal{D}|$ matrix. We define

$$\Delta_\nu(Q) := \nu\Gamma(Q) \in \mathbf{R}^m,$$

as our generic parameter of interest. For example, in the special case where $\mathcal{D} = \{1, 2\}$ and $\nu = (-1, 1)$, $\Delta_\nu(Q) = E_Q[Y_i(2) - Y_i(1)]$ corresponds to the familiar average treatment effect for a binary treatment. Further examples of $\Delta_\nu(Q)$ are provided in Examples 3.1 and 3.2 below.

We now describe our assumptions on Q. Our first assumption imposes restrictions on the (conditional) moments of the potential outcomes.

ASSUMPTION 2.1. *The distribution Q is such that*

(a) $0 < E[\mathrm{Var}[Y_i(d)|X_i]]$ *for* $d \in \mathcal{D}$.

(b) $E[Y_i^2(d)] < \infty$ *for* $d \in \mathcal{D}$.

(c) $E[Y_i(d)|X_i = x]$, $E[Y_i^2(d)|X_i = x]$, *and* $\mathrm{Var}[Y_i(d)|X_i]$ *are Lipschitz for* $d \in \mathcal{D}$.

Assumption 2.1(a) is a mild restriction imposed to rule out degenerate situations and Assumption 2.1(b) is another mild restriction that permits the application of suitable laws of large numbers and central limit theorems. Assumption 2.1(c), on the other hand, is a smoothness requirement that ensures that units that are "close" in terms of their baseline covariates are also "close" in terms of their potential outcomes. Assumption 2.1(c) is a key assumption for establishing the asymptotic exactness of our proposed tests, since it allows us to argue that certain intermediate quantities in the derivations of our variance estimators vanish asymptotically (see, for instance, the proof of Lemma C.2). Similar smoothness requirements are also imposed in Bai, Romano, and Shaikh (2021).

Next, we specify our assumptions on the mechanism determining treatment status. In words, we consider treatment assignments, which first stratify the experimental sample into n blocks of size $|\mathcal{D}|$ using the observed baseline covariates $X^{(n)}$, and then assign one unit to each treatment uniformly at random within each block. We call such a design a *matched tuples* design. Formally, let

$$\lambda_j = \lambda_j\big(X^{(n)}\big) \subseteq \{1, \ldots, J_n\}, \quad 1 \le j \le n$$

denote n sets each consisting of $|\mathcal{D}|$ elements that form a partition of $\{1, \ldots, J_n\}$.

We assume treatment is assigned as follows.

ASSUMPTION 2.2. *Treatments are assigned so that* $\{Y^{(n)}(d) : d \in \mathcal{D}\} \perp\!\!\!\perp D^{(n)}|X^{(n)}$ *and, conditional on* $X^{(n)}$,

$$\big\{(D_i : i \in \lambda_j) : 1 \le j \le n\big\}$$

are i.i.d. and each uniformly distributed over all permutations of $(1, 2, \ldots, |\mathcal{D}|)$.

We further require that the units in each block be "close" in terms of their baseline covariates in the following sense.

ASSUMPTION 2.3. *The blocks satisfy*

$$\frac{1}{n}\sum_{1\le j\le n}\max_{i,k\in\lambda_j}\|X_i - X_k\|^2 \xrightarrow{P} 0.$$

We will also sometimes require that the distances between units in adjacent blocks be "close" in terms of their baseline covariates.

ASSUMPTION 2.4. *The blocks satisfy*

$$\frac{1}{n}\sum_{1\le j\le \lfloor n/2\rfloor}\max_{i\in\lambda_{2j-1},k\in\lambda_{2j}}\|X_i - X_k\|^2 \xrightarrow{P} 0.$$

We provide three examples of blocking algorithms, which satisfy Assumptions 2.3–2.4:

1. Univariate covariate: When $\dim(X_i) = 1$, we can order units from smallest to largest according to X_i and then block adjacent units into blocks of size $|\mathcal{D}|$. It then follows from Theorem 4.1 of Bai, Romano, and Shaikh (2021) that Assumptions 2.3–2.4 are satisfied as long as $E[X_i^2] < \infty$.

2. Pre-stratification: Suppose we have a covariate vector $\tilde{X}_i = (\tilde{X}_{1i}, \tilde{X}_{2i})$, where $\dim(\tilde{X}_{2i}) = 1$. Let S be a function that maps from the support of $\tilde{X}_{1i}$ to a discrete set $\mathcal{S} = \{1, \ldots, |\mathcal{S}|\}$. Define $S_{1i} = S(\tilde{X}_{1i})$. For all units with the same value of S_i, order the units from smallest to largest according to $\tilde{X}_{2i}$ and then block adjacent units into blocks of size $|\mathcal{D}|$.[2] It follows from Theorem 4.1 of Bai, Romano, and Shaikh (2021) that the resulting blocks satisfy Assumptions 2.3–2.4 with $X_i = (S_{1i}, \tilde{X}_{2i})$ as long as $E[\tilde{X}_{2i}^2] < \infty$. As an example, suppose $\tilde{X}_1 =$ (gender, education level) and $\tilde{X}_2 =$ income. In this case, the blocks could be formed by first stratifying according to gender and education level and then blocking on income. A similar blocking procedure is used in the experiment conducted by Fafchamps et al. (2014), which we revisit in our empirical application in Section 5.

3. Recursive pairing: When $\dim(X_i) > 1$ and $|\mathcal{D}| = 2^K$ for some K, we could form blocks by repeatedly implementing the "pairs-of-pairs" algorithm in Section 4 of Bai, Romano, and Shaikh (2021) to successively larger groups of size 2^k for $k = 0, 1, \ldots, K$. To do this, units would first be matched into pairs (using for instance the nonbipartite matching algorithm from the `R` package `nbpMatching`). Next, these matched pairs would themselves be matched into "pairs-of-pairs" using the average value of the covariates in each pair, in order to generate groups of size four. Continuing in this fashion, we would match pairs of groups until obtaining groups

[2] If the number of units in a stratum is not divisible by $|\mathcal{D}|$, we could simply assign the remaining units at random or drop them from the experiment.

of size 2^K. This is the algorithm we employ in our simulation designs. Such an algorithm could again be shown to satisfy Assumptions 2.3–2.4.

3. Main results

3.1 *Inference for matched tuples designs*

In this section, we study estimation and inference for a general m-dimensional parameter $\Delta_\nu(Q)$ under a matched tuples design. For a pre-specified $\ell \times 1$ column vector Δ_0 and $\ell \times m$ matrix Ψ of rank ℓ, the testing problem of interest is

$$H_0 : \Psi\Delta_\nu(Q) = \Delta_0 \quad \text{versus} \quad H_1 : \Psi\Delta_\nu(Q) \neq \Delta_0 \tag{2}$$

at level $\alpha \in (0, 1)$. First, we describe our estimator of $\Delta_\nu(Q)$. For $d \in \mathcal{D}$, define

$$\hat{\Gamma}_n(d) := \frac{1}{n} \sum_{1 \leq i \leq J_n} I\{D_i = d\} Y_i,$$

and let $\hat{\Gamma}_n = (\hat{\Gamma}_n(1), \ldots, \hat{\Gamma}_n(|\mathcal{D}|))'$. In words, $\hat{\Gamma}_n(d)$ is simply the sample mean of the observations with treatment status $D_i = d$, and $\hat{\Gamma}_n$ is the vector of sample means across all treatments $d \in \mathcal{D}$. With $\hat{\Gamma}_n$ in hand, our estimator of $\Delta_\nu(Q)$ is then given by

$$\hat{\Delta}_{\nu,n} := \nu\hat{\Gamma}_n.$$

In what follows, it will be useful to define $\Gamma_d(X_i) := E[Y_i(d)|X_i]$. Our first result derives the limiting distribution of $\hat{\Delta}_{\nu,n}$ under our maintained assumptions.

Theorem 3.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.3. Then*

$$\sqrt{n}\big(\hat{\Delta}_{\nu,n} - \Delta_\nu(Q)\big) \xrightarrow{d} N(0, \mathbb{V}_\nu),$$

where $\mathbb{V}_\nu := \nu\mathbb{V}\nu'$, with

$$\begin{aligned} \mathbb{V} &:= \mathbb{V}_1 + \mathbb{V}_2, \\ \mathbb{V}_1 &:= \operatorname{diag}\big(E\big[\operatorname{Var}\big[Y_i(d)|X_i\big]\big] : d \in \mathcal{D}\big), \\ \mathbb{V}_2 &:= \left[\frac{1}{|\mathcal{D}|}\operatorname{Cov}\big[\Gamma_d(X_i), \Gamma_{d'}(X_i)\big]\right]_{d,d' \in \mathcal{D}}. \end{aligned} \tag{3}$$

To construct our test, we next define a consistent estimator for the asymptotic variance matrix $\mathbb{V}_\nu$. To begin, note by the law of total variance that

$$E\big[\operatorname{Var}\big[Y_i(d)|X_i\big]\big] = \operatorname{Var}\big[Y_i(d)\big] - E\big[E\big[Y_i(d)|X_i\big]^2\big] + E\big[Y_i(d)\big]^2.$$

Therefore, in order to estimate $\mathbb{V}_1$ consistently, it suffices to provide consistent estimators for $E[E[Y_i(d)|X_i]^2]$, $E[Y_i(d)]$, and $\operatorname{Var}[Y_i(d)]$. A similar remark applies to $\mathbb{V}_2$. In light

of this, define

$$\hat{\rho}_n(d,d) := \frac{2}{n}\sum_{1\le j\le \lfloor n/2\rfloor}\Big(\sum_{i\in\lambda_{2j-1}} Y_i I\{D_i=d\}\Big)\Big(\sum_{i\in\lambda_{2j}} Y_i I\{D_i=d\}\Big),$$

$$\hat{\rho}_n\big(d,d'\big) := \frac{1}{n}\sum_{1\le j\le n}\Big(\sum_{i\in\lambda_j} Y_i I\{D_i=d\}\Big)\Big(\sum_{i\in\lambda_j} Y_i I\big\{D_i=d'\big\}\Big) \quad \text{if } d\neq d',$$

$$\hat{\sigma}_n^2(d) := \frac{1}{n}\sum_{1\le i\le J_n}\big(Y_i-\hat{\Gamma}_n(d)\big)^2 I\{D_i=d\}.$$

To understand the construction, note that in order to estimate $E[E[Y_i(d)|X_i]^2]$ consistently, we would ideally average over the products of the outcomes of two units with similar values of X_i and both with treatment status d. By construction, however, only one unit in each block has treatment status d. To overcome this problem, note that Assumption 2.4 ensures that in the limit units in adjacent blocks also have similar values of X_i. Therefore, to construct our estimator of $E[E[Y_i(d)|X_i]^2]$, denoted by $\hat{\rho}_n(d,d)$, we average over the product of the outcomes of the units with treatment status d in two adjacent blocks. $\hat{\rho}_n(d,d)$ is analogous to the "pairs-of-pairs" variance estimator in Bai, Romano, and Shaikh (2021). A similar construction has also been used in Abadie and Imbens (2008) in a related setting. On the other hand, for $d\neq d'$, we have distinct units with treatment status d and d' within each block and, therefore, our estimator of $E[E[Y_i(d)|X_i]E[Y_i(d')|X_i]]$, denoted $\hat{\rho}_n(d,d')$, can be estimated using units within the same block.

Our estimator for $\mathbb{V}_\nu$ is then given by $\hat{\mathbb{V}}_{\nu,n} := \nu\hat{\mathbb{V}}_n\nu'$, where

$$\hat{\mathbb{V}}_n := \hat{\mathbb{V}}_{1,n}+\hat{\mathbb{V}}_{2,n},$$

$$\hat{\mathbb{V}}_{1,n} := \operatorname{diag}\big(\hat{\mathbb{V}}_{1,n}(d): d\in\mathcal{D}\big),$$

$$\hat{\mathbb{V}}_{2,n} := \big[\hat{\mathbb{V}}_{2,n}\big(d,d'\big)\big]_{d,d'\in\mathcal{D}},$$

with

$$\hat{\mathbb{V}}_{1,n}(d) := \hat{\sigma}_n^2(d)-\big(\hat{\rho}_n(d,d)-\hat{\Gamma}_n^2(d)\big),$$

$$\hat{\mathbb{V}}_{2,n}\big(d,d'\big) := \frac{1}{|\mathcal{D}|}\big(\hat{\rho}_n\big(d,d'\big)-\hat{\Gamma}_n(d)\hat{\Gamma}_n\big(d'\big)\big).$$

Given this estimator, our test is given by

$$\phi_n^\nu\big(Z^{(n)}\big) = I\big\{T_n^\nu\big(Z^{(n)}\big)>c_{1-\alpha}\big\},$$

where

$$T_n^\nu\big(Z^{(n)}\big) = n(\Psi\hat{\Delta}_{\nu,n}-\Psi\Delta_0)'\big(\Psi\hat{\mathbb{V}}_{\nu,n}\Psi'\big)^{-1}(\Psi\hat{\Delta}_{\nu,n}-\Psi\Delta_0),$$

and $c_{1-\alpha}$ is the $1-\alpha$ quantile of the χ^2_ℓ distribution. Our next result establishes the consistency of $\hat{\mathbb{V}}_n$ for $\mathbb{V}$ and the asymptotic validity of the above test.

THEOREM 3.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.4. Then*

$$\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V}.$$

Therefore, for the problem of testing (2) at level $\alpha \in (0,1)$, $\phi_n^\nu(Z^{(n)})$ satisfies

$$\lim_{n\to\infty} E\big[\phi_n^\nu\big(Z^{(n)}\big)\big] = \alpha,$$

under the null hypothesis.

EXAMPLE 3.1 (Inference for Matched Triples). Consider the setting where $\mathcal{D} = \{1,2,3\}$, where we consider $d = 1$ as a control arm and $d = 2,3$ as treatment subarms. See, for example, Bold et al. (2018) and Brown and Andrabi (2020). Suppose our parameter of interest is the vector of average treatment effects for the treatments $d = 2,3$ versus control $d = 1$. In this case, the parameter of interest is given by $\Delta_\nu(Q)$, where

$$\nu = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

It follows from Theorem 3.1 that

$$\sqrt{n}\big(\hat{\Delta}_{\nu,n} - \Delta_\nu(Q)\big) \xrightarrow{d} N(0, \mathbb{V}_\nu),$$

where

$$\mathbb{V}_\nu = \begin{pmatrix} \sigma^2_{\nu,1,1} & \sigma^2_{\nu,1,2} \\ \sigma^2_{\nu,1,2} & \sigma^2_{\nu,2,2} \end{pmatrix},$$

and

$$\sigma^2_{\nu,1,1} = E\big[\mathrm{Var}\big[Y_i(1)|X_i\big]\big] + E\big[\mathrm{Var}\big[Y_i(2)|X_i\big]\big] + \frac{1}{3}E\big[\big((\Gamma_1(X_i) - \Gamma_1) - (\Gamma_2(X_i) - \Gamma_2)\big)^2\big],$$

$$\sigma^2_{\nu,2,2} = E\big[\mathrm{Var}\big[Y_i(1)|X_i\big]\big] + E\big[\mathrm{Var}\big[Y_i(3)|X_i\big]\big] + \frac{1}{3}E\big[\big((\Gamma_1(X_i) - \Gamma_1) - (\Gamma_3(X_i) - \Gamma_3)\big)^2\big],$$

$$\sigma^2_{\nu,1,2} = E\big[\mathrm{Var}\big[Y_i(1)|X_i\big]\big] + \frac{1}{3}E\big[\big((\Gamma_1(X_i) - \Gamma_1) - (\Gamma_2(X_i) - \Gamma_2)\big) \times \big((\Gamma_1(X_i) - \Gamma_1) - (\Gamma_3(X_i) - \Gamma_3)\big)\big],$$

where we recall $\Gamma_d(X_i) = E[Y_i(d)|X_i]$. These variance formulas imply the following two observations: first, by decomposing $\sigma^2_{\nu,1,1}$ using the law of total variance, we can show that the commonly used two-sample t-test is conservative when testing the null hypothesis on the contrast of any two treatment levels in a matched tuples design. A similar observation was made in the special case of a matched-pair design in Bai, Romano, and Shaikh (2021). Second, the adjusted t-test developed in Bai, Romano, and Shaikh (2021) is also conservative for testing such hypotheses. Specifically, Bai, Romano, and Shaikh (2021) study inference for $E[Y(2) - Y(1)]$ in a matched-pair design when $|\mathcal{D}| = 2$ and

the sample size is $2n$. In a matched triples experiment with $|\mathcal{D}| = 3$ and sample size $3n$, researchers may be tempted to apply the variance estimator from Theorem 3.3 in Bai, Romano, and Shaikh (2021) to the subsample with $D_i \in \{1, 2\}$. However, it can be shown in our framework that the limit of the variance estimator from Bai, Romano, and Shaikh (2021) is given by replacing $\frac{1}{3}$ in the last term of $\sigma^2_{\nu,1,1}$ with $\frac{1}{2}$. Therefore, the test that studentizes using the variance estimator from Bai, Romano, and Shaikh (2021) would be asymptotically conservative in our setting.

Next, we study the properties of two commonly recommended inference procedures in the analysis of matched tuple designs. The first procedure is a t-test obtained from a linear regression of outcomes on treatment indicators and block fixed effects. Specifically, we consider a t-test obtained from the following regression:

$$Y_i = \sum_{d \in \mathcal{D}\setminus\{1\}} \beta(d) I\{D_i = d\} + \sum_{1 \le j \le n} \delta_j I\{i \in \lambda_j\} + \epsilon_i, \tag{4}$$

which we interpret as the projection of Y on the indicators for treatment status and block fixed effects. Let $\hat{\beta}_n(d)$, $d \in \mathcal{D}\setminus\{1\}$ and $\hat{\delta}_{j,n}$, $1 \le j \le n$ denote the OLS estimators of $\beta(d)$, $d \in \mathcal{D}\setminus\{1\}$ and δ_j, $1 \le j \le n$. It is common in practice to use $\hat{\beta}_n(d)$ as an estimator for the pairwise average treatment effect between treatment d and treatment 1. See, for instance, de Mel et al. (2013) and Fafchamps et al. (2014). Furthermore, researchers often conduct inference on the pairwise ATEs using the heteroskedasticity-robust variance estimator obtained from (4). Formally, for $d \in \mathcal{D}\setminus\{1\}$ and $\Delta_0 \in \mathbf{R}$, consider the problem of testing

$$E_Q\big[Y_i(d)\big] - E_Q\big[Y_i(1)\big] = \Delta_0 \quad \text{versus} \quad H_1 : E_Q\big[Y_i(d)\big] - E_Q\big[Y_i(1)\big] \neq \Delta_0 \tag{5}$$

at level $\alpha \in (0, 1)$. Let $\kappa_j \cdot \hat{\mathbb{V}}^{\text{sfe}}_n(d, 1)$ denote the "HCj" heteroskedasticity-robust variance estimator of $\hat{\beta}_n(d)$ from the linear regression in (4), where κ_j for $j \in \{0, 1\}$ corresponds to one of two common degrees of freedom corrections (see MacKinnon and White (1985)):

$$\kappa_j = \begin{cases} 1 & \text{if } j = 0, \\ \dfrac{|\mathcal{D}|n}{|\mathcal{D}|n - \big(|\mathcal{D}| - 1 + n\big)} & \text{if } j = 1. \end{cases}$$

The test is then defined as

$$\phi^{\text{sfe}}_n\big(Z^{(n)}\big) = I\big\{\big|T^{\text{sfe}}_n\big(Z^{(n)}\big)\big| > z_{1-\frac{\alpha}{2}}\big\}, \tag{6}$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$-th quantile of the standard normal distribution and

$$T^{\text{sfe}}_n\big(Z^{(n)}\big) = \frac{\hat{\beta}_n(d) - \Delta_0}{\sqrt{\kappa_j \cdot \hat{\mathbb{V}}^{\text{sfe}}_n(d, 1)}}. \tag{7}$$

The following theorem shows that the OLS estimator $\hat{\beta}_n(d)$ is numerically equivalent to the standard difference-in-means estimator. However, it shows that the t-test defined in (6) is not generally valid for testing the null hypothesis defined in (5).

THEOREM 3.3. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 2.2–2.4. Then*

$$\hat{\beta}_n(d) = \hat{\Gamma}_n(d) - \hat{\Gamma}_n(1) \quad \text{for } d \in \mathcal{D}\backslash\{1\}.$$

Moreover,

- *Using estimator* HC0, *the limiting rejection probability of the test defined in (6) could be strictly larger than α.*
- *Using estimator* HC1, *the limiting rejection probability of the test defined in (6) could be strictly larger than α for $|\mathcal{D}| > 2$.*

Bai, Romano, and Shaikh (2021) remark that the test defined in (6) is conservative in the context of a matched-pair design when using HC1. Theorem 3.3 shows that, when considering a matched tuples design with more than two treatments, this is no longer necessarily the case.

REMARK 3.1. An inspection of the proof of Theorem 3.3 reveals that the probability limit of $n \cdot \kappa_1 \hat{\mathbb{V}}_n^{\text{sfe}}(d, 1)$ is given by

$$\begin{aligned}
&\frac{|\mathcal{D}|}{|\mathcal{D}|-1}\left(\operatorname{Var}\left[\Gamma_1(X_i) - \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\Gamma_{d'}(X_i)\right] + \left(1-\frac{1}{|\mathcal{D}|}\right)^2 E\big[\operatorname{Var}\big[Y_i(1)|X_i\big]\big]\right.\\
&\quad + \frac{1}{|\mathcal{D}|^2}\sum_{d'\in\mathcal{D}\backslash\{1\}} E\big[\operatorname{Var}\big[Y_i(d')|X_i\big]\big]\\
&\quad + \operatorname{Var}\left[\Gamma_d(X_i) - \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\Gamma_{d'}(X_i)\right]\\
&\quad \left. + \left(1-\frac{1}{|\mathcal{D}|}\right)^2 E\big[\operatorname{Var}\big[Y_i(d)|X_i\big]\big] + \frac{1}{|\mathcal{D}|^2}\sum_{d'\in\mathcal{D}\backslash\{d\}} E\big[\operatorname{Var}\big[Y_i(d')|X_i\big]\big]\right),
\end{aligned}$$

whereas the true asymptotic variance of $\hat{\beta}_n(d)$ is given by

$$E\big[\operatorname{Var}\big[Y_i(d)|X_i\big]\big] + E\big[\operatorname{Var}\big[Y_i(1)|X_i\big]\big] + \frac{1}{|\mathcal{D}|}E\big[\big((\Gamma_d(X_i)-\Gamma_d) - (\Gamma_1(X_i)-\Gamma_1)\big)^2\big].$$

From these expressions, we can conclude that when $|\mathcal{D}|$ is large it is likely that $\kappa_1 \hat{\mathbb{V}}_n^{\text{sfe}}(d, 1)$ is conservative. However, as shown in the proof of Theorem 3.3, this cannot be guaranteed for finite $|\mathcal{D}| > 2$ in general.

The second procedure is a block-cluster robust t-test, which modifies a recent proposal in de Chaisemartin and Ramirez-Cuellar (2022) to the setting with multiple treatments. Specifically, we consider a cluster-robust t-test constructed from a regression of outcomes on a constant and treatment indicators:

$$Y_i = \gamma(1) + \sum_{d\in\mathcal{D}\backslash\{1\}} \gamma(d) I\{D_i = d\} + \epsilon_i,$$

where clusters are defined at the level of *blocks* of units $\{\lambda_j\}_{1\leq j\leq \mathcal{D}}$. Let $\hat{\gamma}_n(d)$, $d \in \mathcal{D}\backslash\{1\}$ denote the OLS estimator of $\gamma(d)$; it then follows immediately that $\hat{\gamma}_n(d) = \hat{\Gamma}_n(d) - \hat{\Gamma}_n(1)$. We then consider the problem of testing (5) at level $\alpha \in (0, 1)$ using a test defined by

$$\phi_n^{\text{bcve}}\big(Z^{(n)}\big) = I\big\{\big|T_n^{\text{bcve}}\big(Z^{(n)}\big)\big| > z_{1-\frac{\alpha}{2}}\big\},$$

where $z_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-th quantile of the standard normal distribution and

$$T_n^{\text{bcve}}\big(Z^{(n)}\big) = \frac{\hat{\gamma}_n(d) - \Delta_0}{\sqrt{\hat{\mathbb{V}}_n^{\text{bcve}}(d)}}, \tag{8}$$

with $\hat{\mathbb{V}}_n^{\text{bcve}}(d)$ denoting the dth diagonal element of the block-cluster variance estimator defined as

$$\hat{\mathbb{V}}_n^{\text{bcve}} = \Bigg(\sum_{1\leq j\leq n}\sum_{i\in\lambda_j} C_i C_i'\Bigg)^{-1}\Bigg(\sum_{1\leq j\leq n}\Bigg(\sum_{i\in\lambda_j}\hat{\epsilon}_i C_i\Bigg)\Bigg(\sum_{i\in\lambda_j}\hat{\epsilon}_i C_i\Bigg)'\Bigg)\Bigg(\sum_{1\leq j\leq n}\sum_{i\in\lambda_j} C_i C_i'\Bigg)^{-1}, \tag{9}$$

where $C_i = (1, I\{D_i = 2\}, \ldots, I\{D_i = |\mathcal{D}|\})'$ and $\hat{\epsilon}_i = \sum_{d\in\mathcal{D}\backslash\{1\}}(Y_i - \hat{\gamma}_n(d))I\{D_i = d\} + Y_i I\{D_i = 1\} - \hat{\gamma}_n(1)$.

The following theorem shows that the t-test defined in (8) is generally conservative for testing the null hypothesis defined in (5).

THEOREM 3.4. *Consider the block-cluster variance estimator* $\hat{\mathbb{V}}_n^{\text{bcve}}$ *as defined in* (9). *Then the* d*th diagonal element of this estimator is equal to*

$$n\cdot\hat{\mathbb{V}}_n^{\text{bcve}}(d) = \frac{1}{n}\sum_{1\leq j\leq n}\Bigg(\sum_{i\in\lambda_j} Y_i I\{D_i = d\} - \sum_{i\in\lambda_j} Y_i I\{D_i = 1\}\Bigg)^2 - \big(\hat{\Gamma}_n(d) - \hat{\Gamma}_n(1)\big)^2.$$

Moreover, under Assumptions 2.1–2.3,

$$n\cdot\hat{\mathbb{V}}_n^{\text{bcve}}(d) \xrightarrow{p} E\big[\text{Var}\big[Y_i(d)|X_i\big]\big] + E\big[\text{Var}\big[Y_i(1)|X_i\big]\big] + E\big[\big(\big(\Gamma_d(X_i) - \Gamma_d\big) - \big(\Gamma_1(X_i) - \Gamma_1\big)\big)^2\big].$$

It thus follows that the test defined in (8) *is conservative for testing the null hypothesis defined in* (5) *unless*

$$E\big[\big(\big(\Gamma_d(X_i) - \Gamma_d\big) - \big(\Gamma_1(X_i) - \Gamma_1\big)\big)^2\big] = 0. \tag{10}$$

REMARK 3.2. An inspection of the proof of Theorem 3.4 reveals that unless (10) holds the difference between the probability limit of $n\cdot\hat{\mathbb{V}}_n^{\text{bcve}}(d)$ and the asymptotic variance of $\hat{\Gamma}_n(d) - \hat{\Gamma}_n(1)$ is equal to

$$\Big(1 - \frac{1}{|\mathcal{D}|}\Big)E\big[\big(\big(\Gamma_d(X_i) - \Gamma_d\big) - \big(\Gamma_1(X_i) - \Gamma_1\big)\big)^2\big].$$

It thus follows that the test defined in (8) in fact becomes more conservative for testing (5) as the number of treatments $|\mathcal{D}|$ increases.

3.2 *Inference for "replicate" designs*

Our analysis so far has focused on the setting where $J_n = |\mathcal{D}|n$ units are blocked into n blocks of size $|\mathcal{D}|$, and each treatment $d \in \mathcal{D}$ is assigned exactly once in each block. In this section, we consider a modification of this design where units are grouped into blocks of size $2|\mathcal{D}|$ and each treatment status $d \in \mathcal{D}$ is assigned exactly *twice* in each block. Formally, for the remainder of this section, suppose n is even, and let

$$\tilde{\lambda}_j = \tilde{\lambda}_j\big(X^{(n)}\big) \subseteq \{1, \ldots, J_n\}, \quad 1 \le j \le n/2$$

denote $n/2$ sets each consisting of $2|\mathcal{D}|$ elements that form a partition of $\{1, \ldots, J_n\}$.

We assume treatment is assigned as follows.

ASSUMPTION 3.1. *Treatments are assigned so that* $\{Y^{(n)}(d) : d \in \mathcal{D}\} \perp\!\!\!\perp D^{(n)} | X^{(n)}$ *and, conditional on* $X^{(n)}$,

$$\big\{(D_i : i \in \tilde{\lambda}_j) : 1 \le j \le n/2\big\}$$

are i.i.d. and each uniformly distributed over all permutations of $(1, 1, 2, 2, \ldots, |\mathcal{D}|, |\mathcal{D}|)$.

We further require that the units in each block be "close" in terms of their baseline covariates in the following sense.

ASSUMPTION 3.2. *The blocks satisfy*

$$\frac{1}{n} \sum_{1 \le j \le n/2} \max_{i,k \in \tilde{\lambda}_j} \|X_i - X_k\|^2 \xrightarrow{P} 0.$$

We first establish that the limiting distribution of $\hat{\Delta}_{\nu,n}$ for such a "replicate" design is the same as that for the matched tuples design considered in Theorem 3.1.

THEOREM 3.5. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 3.1–3.2. Then*

$$\sqrt{n}\big(\hat{\Delta}_{\nu,n} - \Delta_\nu(Q)\big) \xrightarrow{d} N(0, \mathbb{V}_\nu),$$

with $\mathbb{V}_\nu$ as defined in Theorem 3.1.

Although the limiting distribution of $\hat{\Delta}_{\nu,n}$ for the standard matched tuples and replicate designs are identical, variance estimation in the replicate design is often understood to be conceptually simpler, because each treatment status is assigned *twice* in each block (see, for instance, the discussion of variance estimation in Athey and Imbens (2017) in the context of matched pair designs). Indeed, in this case an alternative variance estimator can be constructed, which is identical to the estimator proposed in Section 3.1 except that we replace $\hat{\rho}_n(d, d)$ by

$$\tilde{\rho}_n(d, d) = \frac{2}{n} \sum_{1 \le j \le \lfloor n/2 \rfloor} \left(\prod_{i \in \lambda_j} Y_i I\{D_i = d\} \right),$$

which no longer requires averaging over the product of outcomes of units in adjacent blocks. The following theorem establishes the consistency of $\tilde{\rho}_n(d,d)$, where importantly we note that Assumption 2.4, which maintains that adjacent blocks be "close," is no longer required. It is then straightforward to show the consistency of the corresponding variance estimator for $\hat{\Delta}_{\nu,n}$ constructed by replacing $\hat{\rho}_n(d,d)$ in $\hat{\mathbb{V}}_n$ with $\tilde{\rho}_n(d,d)$.

THEOREM 3.6. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumptions 3.1–3.2. Then*

$$\tilde{\rho}_n(d,d) \xrightarrow{P} E\big[E\big[Y_i(d)|X_i\big]^2\big]. \tag{11}$$

We remark that Theorems 3.1–3.2 and Theorems 3.5–3.6, yielding identical conclusions, do not allow us to effectively compare the properties of the standard matched tuples design and matched tuples with replicates. In order to compare these designs, we evaluate their finite sample properties via simulation in Section 4. There, we find that the mean squared error of $\hat{\Delta}_{\nu,n}$ under the replicate design is typically larger than under the standard nonreplicate design. However, we also find that the rejection probabilities of our proposed tests under the replicate design are much closer to the nominal level relative to the nonreplicate design, which can sometimes exhibit rejection probabilities strictly smaller than the nominal level when matching on multiple covariates. As a result, the replicate design is sometimes able to achieve better power relative to the nonreplicate design. We emphasize, however, that our current asymptotic framework is not precise enough to capture these differences. One possible conjecture is that since replicate designs could be thought of as convex combinations of matched tuples designs (see Lemma 2 in Bai (2022)), it is as if we are averaging over multiple matched tuples designs when we estimate the limiting variance. However, we leave a detailed theoretical comparison of these two designs to future work.

3.3 *Asymptotic properties of fully-blocked 2^K factorial designs*

In this section, we apply the results derived in Sections 3.1–3.2 to study the asymptotic properties of what we call "fully-blocked" 2^K factorial designs. Section 3.3.1 introduces 2^K factorial experiments. Section 3.3.2 introduces the fully-blocked factorial design and compares the efficiency properties of fully-blocked factorial designs to some alternative designs.

3.3.1 *Setup and notation for 2^K factorial designs* In this section, we describe the setup of a 2^K factorial experiment, the resulting parameters of interest, and their corresponding estimators (see Wu and Hamada (2011) for a textbook treatment). A 2^K factorial design assigns treatments, which are combinations of multiple "factors," where each factor can take two distinct values, or "levels." For instance, Karlan et al. (2014) study the effect of capital constraints and uninsured risk on the investment decisions of farmers in Ghana. In their setting, each treatment consists of two factors: whether or not a household receives a cash grant, and whether or not a household receives an insurance grant. Our setup and notation mirror the framework introduced in Dasgupta, Pillai, and Rubin

(2015) and Li, Ding, and Rubin (2020). Given K factors each with two treatment levels $\{-1, +1\}$, our set of treatments $\mathcal{D}$ now consists of all possible 2^K factor combinations. For a factor combination $d \in \mathcal{D}$, define $\iota_k(d) \in \{-1, +1\}$ to be the level of factor k under treatment d. The vector $\iota(d) := (\iota_1(d), \iota_2(d), \ldots, \iota_K(d))$ then describes the levels of all K factors associated with factor combination d. This notation allows us to define *factorial effects* as parameters of the form $\Delta_\nu(Q)$ for appropriately constructed contrast vectors ν. For instance, consider the contrast vector defined as

$$\nu_k := \big(\iota_k(1), \iota_k(2), \ldots, \iota_k\big(|\mathcal{D}|\big)\big).$$

Then the parameter $\Delta_{\nu_k}(Q)$ obtained from this contrast can be written as

$$\Delta_{\nu_k}(Q) = \sum_{d\in\mathcal{D}} I\big\{\iota_k(d) = +1\big\}\Gamma_d(Q) - \sum_{d\in\mathcal{D}} I\big\{\iota_k(d) = -1\big\}\Gamma_d(Q).$$

We define the *main effect* of factor k as $2^{-(K-1)}\Delta_{\nu_k}(Q)$. In words, the main effect of factor k measures the average difference between the outcomes of factor combinations under which the kth factorial effect is 1 versus the outcomes of factor combinations under which the kth factorial effect is -1. The rescaling $2^{-(K-1)}$ is introduced because there are 2^{K-1} possible values for all the factor combinations when fixing the kth factor. We call ν_k the *generating vector* for the main effect of factor k.

We can subsequently build on the generating vectors of the main effects in order to define the *interaction effects* between various factors. The interaction effect between a given set of factors is defined using the contrast obtained from taking the elementwise product of the generating vectors for the relevant factors. For instance, the two-factor interaction between factors k and k' is defined as $2^{-(K-1)}\Delta_{\nu_{k,k'}}(Q)$, where $\nu_{k,k'} := \nu_k \odot \nu_{k'}$ and $\odot$ denotes elementwise multiplication. Similarly, the three-factor interaction $2^{-(K-1)}\Delta_{\nu_{k,k',k''}}(Q)$ is defined using the contrast vector $\nu_{k,k',k''} := \nu_k \odot \nu_{k'} \odot \nu_{k''}$. We illustrate these definitions in the special case of a 2^2 factorial design in Example 3.2 below. For simplicity, in what follows, we omit the rescaling by $2^{-(K-1)}$ in our discussions and results.

EXAMPLE 3.2. Here, we illustrate the concept of main and interaction effects in the case of a 2^2 factorial design. Table 1 depicts the 4 factor combinations and their corresponding factor levels.

TABLE 1. Example of a 2^2 factorial design.

Factor Combination	Factor 1	Factor 2	Factor 1/2 Interaction
1	-1	-1	$+1$
2	-1	$+1$	-1
3	$+1$	-1	-1
4	$+1$	$+1$	$+1$

From the column labeled Factor 1, we observe that the generating vector for the main effect of factor one, ν_1, is given by

$$\nu_1 = (-1, -1, +1, +1),$$

so that the main effect of factor one is given by (up to rescaling)

$$\Delta_{\nu_1}(Q) = E_Q\big[Y_i(+1, +1) + Y_i(+1, -1)\big] - E_Q\big[Y_i(-1, +1) + Y_i(-1, -1)\big],$$

where here we have indexed potential outcomes explicitly by their factor levels. Similarly, the column labeled Factor 2 corresponds to the generating vector for the main effect of factor two, ν_2. To define the interaction effect between factors one and two, we construct the relevant contrast by taking the elementwise product of ν_1 and ν_2:

$$\nu_{1,2} = \nu_1 \odot \nu_2 = (+1, -1, -1, +1);$$

this produces the column labeled factor 1/2 interaction. Accordingly, the interaction effect between factors one and two is given by (up to rescaling)

$$\Delta_{\nu_{1,2}}(Q) = E_Q\big[Y_i(+1, +1) - Y_i(-1, +1)\big] - E_Q\big[Y_i(+1, -1) - Y_i(-1, -1)\big].$$

In words, $\Delta_{\nu_{1,2}}(Q)$ measures the difference in the the average difference in potential outcomes over factor one when factor two is set to 1 versus the average difference in potential outcomes over factor one when factor two is set to -1.

Given the above setup, we estimate the factorial effect given by $\Delta_\nu(Q)$ using the estimator $\hat{\Delta}_{\nu,n}$ defined in Section 3.1. Wu and Hamada (2011) and Dasgupta, Pillai, and Rubin (2015) explain that $\hat{\Delta}_{\nu,n}$ is a standard estimator in this context. For instance, the estimator of the main effect of factor k, $2^{-(K-1)}\hat{\Delta}_{\nu_k,n}$, is in fact the difference-in-means estimator over the kth factor:

$$\begin{aligned}
&2^{-(K-1)}\hat{\Delta}_{\nu_k,n} \\
&\quad = \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}} I\{\iota_k(d) = +1\}\hat{\Gamma}_n(d) - \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}} I\{\iota_k(d) = -1\}\hat{\Gamma}_n(d) \\
&\quad = \frac{1}{n2^{K-1}}\sum_{1\le i\le J_n}\sum_{d\in\mathcal{D}} I\{\iota_k(d) = +1\}I\{D_i = d\} \\
&\qquad \times Y_i - \frac{1}{n2^{K-1}}\sum_{1\le i\le J_n}\sum_{d\in\mathcal{D}} I\{\iota_k(d) = -1\}I\{D_i = d\}Y_i \\
&\quad = \frac{1}{n2^{K-1}}\sum_{1\le i\le J_n} I\{\iota_k(D_i) = +1\}Y_i - \frac{1}{n2^{K-1}}\sum_{1\le i\le J_n} I\{\iota_k(D_i) = -1\}Y_i.
\end{aligned}$$

3.3.2 *Efficiency properties of fully-blocked factorial designs* In this section, we compare the asymptotic variance of the estimator $\hat{\Delta}_{\nu,n}$ under what we call a "fully-blocked" factorial design relative to some alternative designs. A fully-blocked factorial design first

blocks the experimental sample into n blocks of size 2^K based on the observable characteristics $X^{(n)}$, and then assigns each of the 2^K factor combinations exactly once in each block. Formally, a fully-blocked factorial design is simply a matched tuples design as defined in Section 2, where $\mathcal{D}$ consists of the set of all possible factor combinations.

Our first result compares the fully-blocked factorial design to completely randomized and stratified factorial designs. Given a 2^K factorial experiment and a sample of size $J_n = n2^K$, a completely randomized factorial design simply assigns n individuals to each of the 2^K factor combinations at random. A stratified factorial design first partitions the covariate space into a finite number of groups, or "strata," and then performs a completely randomized factorial design within each stratum. Formally, let $h : \text{supp}(X) \to \{1, \ldots, S\}$ be a function, which maps covariate values into a set of discrete strata labels. Then a stratified factorial design performs a completely randomized factorial design within each stratum produced by $h(\cdot)$. Note that a completely randomized design is a special case of the stratified factorial design where the codomain of $h(\cdot)$ is a singleton. See Branson, Dasgupta, and Rubin (2016) and Li, Ding, and Rubin (2020) for further discussion of these designs. Theorem 3.7 shows that the asymptotic variance of $\hat{\Delta}_{\nu,n}$ is weakly smaller under a fully-blocked factorial design than that under *any* stratified factorial design as defined above, as long as the potential outcomes satisfy the smoothness assumptions described in Assumption 2.1(c).

THEOREM 3.7. *Suppose Assumptions 2.1(a)–(b) hold and let $h : \text{supp}(X) \to \{1, \ldots, S\}$ be any measurable function, which maps covariate values into a set of discrete strata labels. Let $\Delta_\nu(Q)$ be a factorial effect for some 1×2^K contrast vector ν. Then under a stratified factorial design with strata defined by $h(\cdot)$,*

$$\sqrt{n}\big(\hat{\Delta}_{\nu,n} - \Delta_\nu(Q)\big) \xrightarrow{d} N\big(0, \sigma^2_{h,\nu}\big),$$

where $\sigma^2_{h,\nu} = \nu \mathbb{V}_h \nu'$, with

$$\begin{aligned}
\mathbb{V}_h &:= \mathbb{V}_{h,1} + \mathbb{V}_{h,2},\\
\mathbb{V}_{h,1} &:= \text{diag}\big(E\big[\text{Var}\big[Y_i(d)|h(X_i)\big]\big] : d \in \mathcal{D}\big),\\
\mathbb{V}_{h,2} &:= \left[\frac{1}{|\mathcal{D}|}\text{Cov}\big[E\big[Y_i(d)|h(X_i)\big], E\big[Y_i\big(d'\big)|h(X_i)\big]\big]\right]_{d,d' \in \mathcal{D}}.
\end{aligned}$$

Moreover,

$$\sigma^2_\nu \leq \sigma^2_{h,\nu},$$

where $\sigma^2_\nu = \mathbb{V}_\nu$ (as defined in Theorem 3.1) is the asymptotic variance of $\hat{\Delta}_{\nu,n}$ (under Assumptions 2.1–2.3) for a fully-blocked factorial design.

REMARK 3.3. Branson, Dasgupta, and Rubin (2016) and Li, Ding, and Rubin (2020) propose rerandomization designs in the context of factorial experiments, which are also shown to have favorable properties relative to complete and stratified factorial designs. In Section 4.1, we compare the mean squared error of the fully-blocked design to a rerandomized design via Monte Carlo simulation.

Our next result considers settings where only a subset of the factors are of primary interest to the researcher. For instance, Besedeš et al. (2012) use a factorial design to study how the number of options in an agent's choice set affects their ability to make optimal decisions. Here, the primary factor of interest is the number of options (four or thirteen), but the design also features other secondary factors. In such a case, we might imagine that a matched pairs design, which focuses on the factor of primary interest and assigns the other factors by i.i.d. coin flips may be more efficient for estimating the primary factorial effect than the fully-blocked design that treats all the factors symmetrically. In particular, we consider a setting where we are interested in the average main effect on the kth factor, $\Delta_{\nu_k}(Q)$, and compare the performance of the fully-blocked design to a design, which performs matched pairs over the kth factor while assigning the other factors to individuals at random using i.i.d. Bernoulli$(1/2)$ assignment. We call such a design the "factor k specific" matched pairs design. Formally, let

$$\zeta_j = \zeta_j\big(X^{(n)}\big) \subset \big\{1, \ldots, 2^K n\big\}, \quad 1 \le j \le 2^{K-1} n$$

denote a partition of the set of indices such that each ζ_j contains two units. The "factor k specific" matched pairs design satisfies the following assumption.

ASSUMPTION 3.3. *Treatment status is assigned so that* $\{Y^{(n)}(d) : d \in \mathcal{D}\} \perp\!\!\!\perp D^{(n)} | X^{(n)}$ *and, conditional on* $X^{(n)}$,

$$\big\{\big(\iota_k(D_i) : i \in \zeta_j\big) : 1 \le j \le 2^{K-1} n\big\}$$

are i.i.d. and each uniformly distributed over $\{(-1,+1),(+1,-1)\}$. *Furthermore, independently of* $X^{(n)}$ *and independently across* $1 \le j \le K$, $j \ne k$, $\iota_j(D_i)$ *is i.i.d. across* $1 \le i \le 2^K n$ *and* $P\{\iota_j(D_i) = -1\} = P\{\iota_j(D_i) = +1\} = \frac{1}{2}$.

Theorem 3.8 shows that the asymptotic variance of $\hat{\Delta}_{\nu_1,n}$ is weakly smaller under a fully-blocked design than that under the factor specific matched pairs design.

THEOREM 3.8. *Suppose Assumptions 2.1–2.3 hold and the treatment assignment mechanism satisfies Assumption 3.3. Then*

$$\sqrt{n}\big(\hat{\Delta}_{\nu_k,n} - \Delta_{\nu_k}(Q)\big) \xrightarrow{d} N(0, \mathbb{V}_{\nu_k} + \xi_1 + \xi_0),$$

where $\mathbb{V}_{\nu_k}$ *is defined in Theorem 3.1, and*

$$\xi_1 = \sum_{d \in \mathcal{D}: \iota_k(d) = +1} E\left[\left(\Gamma_d(X_i) - \frac{1}{2^{K-1}} \sum_{d' \in \mathcal{D}: \iota_k(d') = +1} \Gamma_{d'}(X_i)\right)^2\right],$$

$$\xi_0 = \sum_{d \in \mathcal{D}: \iota_k(d) = -1} E\left[\left(\Gamma_d(X_i) - \frac{1}{2^{K-1}} \sum_{d' \in \mathcal{D}: \iota_k(d') = -1} \Gamma_{d'}(X_i)\right)^2\right].$$

REMARK 3.4. In this section, we have presented results for "full" factorial designs, which assign individuals to every possible combination of factors. This is in contrast to "fractional" factorial designs, which assign only a subset of the possible factor combinations

(see, e.g., Wu and Hamada (2011), Pashley and Bind (2019)). We leave possible extensions of our procedure to the fractional case for future work.

4. Simulations

In this section, we examine the finite sample performance of the estimator $\hat{\Delta}_{\nu,n}$ and the test $\phi_n^\nu(Z^{(n)})$ in the context of a 2^K factorial experiment, under various alternative experimental designs. In Sections 4.1 and 4.2, the data generating processes are as specified below (in Section 4.3 we study an alternative design with multiple covariates and factors). For $d = (d^{(1)}, d^{(2)}) \in \{-1, 1\}^2$ and $1 \le i \le 4n$, the potential outcomes are generated according to the equation:

$$Y_i(d) = \mu_d + \mu_d(X_i) + \sigma_d(X_i)\epsilon_i.$$

In each of the specifications, $((X_i, \epsilon_i) : 1 \le i \le 4n)$ are i.i.d.; for $1 \le i \le 4n$, X_i and ϵ_i are independent.

Model 1: $\mu_{1,a}(X_i) = \mu_{-1,a}(X_i) = \gamma X_i$ for $a \in \{-1, 1\}$, where $\gamma = 1$. $\mu_{1,1} = 2\mu_{1,-1} = 4\mu_{-1,1} = 2\tau$ for a parameter $\tau \in \{0, 0.2\}$, $\mu_{-1,-1} = 0$, $\epsilon_i \sim N(0, 1)$ and $X_i \sim N(0, 1)$ for all $d \in \{-1, 1\}^2$ and $\sigma_d(X_i) = 1$.

Model 2: As in Model 1, but $\mu_d(X_i) = X_i + (X_i^2 - 1)/3$.

Model 3: As in Model 1, but $\mu_d(X_i) = \gamma_d X_i + (X_i^2 - 1)/3$. $\gamma_{1,1} = 2$, $\gamma_{-1,1} = 1$, $\gamma_{1,-1} = 1/2$ and $\gamma_{-1,-1} = -1$.

Model 4: As in Model 3, but $\mu_d(X_i) = \sin(\gamma_d X_i)$.

Model 5: As in Model 3, $\mu_d(X_i) = \sin(\gamma_d X_i) + \gamma_d X_i/10 + (X_i^2 - 1)/3$.

Model 6: As in Model 3, but $\sigma_d(X_i) = (1 + d^{(1)} + d^{(2)})X_i^2$.

We consider five parameters of interest as listed in Table 2. $\Delta_{\nu_1}(Q)$ and $\Delta_{\nu_2}(Q)$ correspond to the main factorial effects for the two factors. $\Delta_{\nu_{1,2}}(Q)$ corresponds to the interaction effect between the two factors, as discussed in Example 3.2. $\Delta_{\nu_1^1}(Q)$ and $\Delta_{\nu_{-1}^1}(Q)$ denote the average effect of one factor, keeping the value of the other factor fixed at 1 or -1. All simulations are performed with a sample of size $4n = 1000$.

Table 2. Parameters of interest.

Parameter of Interest	Formula
$\frac{1}{2}\Delta_{\nu_1}(Q)$	$\frac{1}{2}E[Y_i(1,1) - Y_i(-1,1)] + \frac{1}{2}E[Y_i(1,-1) - Y_i(-1,-1)]$
$\frac{1}{2}\Delta_{\nu_2}(Q)$	$\frac{1}{2}E[Y_i(1,1) - Y_i(1,-1)] + \frac{1}{2}E[Y_i(-1,1) - Y_i(-1,-1)]$
$\frac{1}{2}\Delta_{\nu_{1,2}}(Q)$	$\frac{1}{2}E[Y_i(1,1) - Y_i(-1,1)] - \frac{1}{2}E[Y_i(1,-1) - Y_i(-1,-1)]$
$\Delta_{\nu_1^1}(Q)$	$E[Y_i(1,1) - Y_i(-1,1)]$
$\Delta_{\nu_{-1}^1}(Q)$	$E[Y_i(1,-1) - Y_i(-1,-1)]$

4.1 *MSE properties of the matched tuples design*

In this section, we study the mean-squared-error performance of $\hat{\Delta}_{\nu,n}$ across several experimental designs. We analyze and compare the MSE for all five parameters of interest for the following seven experimental designs:

1. **(B-B)** $(D_i^{(1)}, D_i^{(2)})$ are i.i.d. across $1 \leq i \leq 4n$ and the two entries are independently distributed as $2A - 1$, where A follows Bernoulli$(1/2)$.
2. **(C)** $(D_i^{(1)}, D_i^{(2)})$ are jointly drawn from a completely randomized design. We uniformly at random divide the experimental sample of size $4n$ into four groups of size n and assign a different $d \in \{-1, 1\}^2$ for each group.
3. **(MP-B)** A matched-pair design for $D^{(1)}$, where units are ordered and paired according to X_i. For each pair, uniformly at random assign $D_i^{(1)} = 1$ to one of the units. Independently, $(D_i^{(2)} : 1 \leq i \leq 4n)$ are i.i.d. with the distribution of $2A - 1$, where $A \sim$ Bernoulli$(1/2)$.
4. **(MT)** Matched tuples design where units are ordered according to X_i.
5. **(Large-2)** A stratified design, where the experimental sample is divided into two strata using the median of X_i as the cutoff. In each stratum, treatment is assigned as in **C**.
6. **(Large-4)** As in **(Large-2)**, but with four strata.
7. **(RE)** A rerandomization design using a Mahalanobis balance function. As outlined in Branson, Dasgupta, and Rubin (2016), we select the main-effect threshold criterion to be the $100(0.01^{1/K})$ percentile of a χ^2_p distribution with $p = \dim(X_i)$, and select the interaction-effect threshold criterion to be $100(0.01^{1/L})$, where L is the number of interaction effects.

Table 3 displays the ratio of the MSE of each design relative to the MSE of **MT**, computed across 4000 Monte Carlo replications. In each of the designs, we set treatment effects to zero by setting $\tau = 0$. As expected from Theorems 3.7 and 3.8, **MT** outperforms **B-B**, **C**, **MP-B**, **Large-2**, and **Large-4** in every model specification. We also find that **MT** compares favorably to **RE**, with **RE** slightly outperforming **MT** in some cases, but with **MT** outperforming in general. Although we do not have formal results comparing the matched tuples design to rerandomization, we note that rerandomization redraws treatments until the distances between certain features of the covariate distribution across treatment statuses are below certain pre-specified thresholds. In contrast, the matched tuples design attempts to *minimize* these distances by blocking units finely based on the covariates. See also Remark 3 of Bai (2022) for a related observation in the binary treatment setting.

4.2 *Inference*

In this section, we study the finite sample properties of several different tests of the null hypothesis $H_0 : \Delta_\nu = 0$ for various choices of ν, against the alternative hypotheses im-

TABLE 3. Ratio of MSEs relative to MT.

Model	Parameter	B-B	C	MP-B	MT	Large-2	Large-4	RE
1	Δ_{ν_1}	2.099	1.948	1.045	1.000	1.335	1.138	1.031
	Δ_{ν_2}	2.036	2.015	2.113	1.000	1.407	1.179	0.988
	$\Delta_{\nu_{1,2}}$	2.008	2.044	2.016	1.000	1.423	1.091	1.014
	$\Delta_{\nu_1^1}$	2.051	2.014	1.563	1.000	1.402	1.134	1.029
	$\Delta_{\nu_{-1}^1}$	2.057	1.978	1.498	1.000	1.357	1.095	1.017
2	Δ_{ν_1}	2.327	2.168	1.044	1.000	1.546	1.249	1.232
	Δ_{ν_2}	2.254	2.259	2.355	1.000	1.619	1.312	1.209
	$\Delta_{\nu_{1,2}}$	2.249	2.287	2.173	1.000	1.646	1.225	1.250
	$\Delta_{\nu_1^1}$	2.285	2.265	1.634	1.000	1.599	1.260	1.227
	$\Delta_{\nu_{-1}^1}$	2.291	2.190	1.585	1.000	1.593	1.215	1.255
3	Δ_{ν_1}	2.042	1.996	1.792	1.000	1.422	1.206	1.124
	Δ_{ν_2}	1.576	1.527	1.480	1.000	1.221	1.140	1.109
	$\Delta_{\nu_{1,2}}$	3.113	2.982	1.943	1.000	1.900	1.337	1.187
	$\Delta_{\nu_1^1}$	3.401	3.351	2.237	1.000	1.979	1.410	1.225
	$\Delta_{\nu_{-1}^1}$	1.899	1.802	1.619	1.000	1.388	1.166	1.103
4	Δ_{ν_1}	1.311	1.305	1.252	1.000	1.100	1.070	1.194
	Δ_{ν_2}	1.218	1.210	1.167	1.000	1.063	1.064	1.057
	$\Delta_{\nu_{1,2}}$	1.296	1.289	1.152	1.000	1.184	1.084	1.191
	$\Delta_{\nu_1^1}$	1.416	1.401	1.259	1.000	1.158	1.080	1.249
	$\Delta_{\nu_{-1}^1}$	1.201	1.202	1.150	1.000	1.128	1.075	1.140
5	Δ_{ν_1}	1.603	1.606	1.315	1.000	1.280	1.169	1.375
	Δ_{ν_2}	1.444	1.458	1.378	1.000	1.225	1.173	1.235
	$\Delta_{\nu_{1,2}}$	1.607	1.598	1.351	1.000	1.370	1.184	1.390
	$\Delta_{\nu_1^1}$	1.802	1.797	1.415	1.000	1.353	1.192	1.441
	$\Delta_{\nu_{-1}^1}$	1.434	1.434	1.262	1.000	1.301	1.164	1.332
6	Δ_{ν_1}	1.119	1.122	1.116	1.000	1.055	1.021	1.065
	Δ_{ν_2}	1.051	1.042	1.056	1.000	1.026	0.991	0.989
	$\Delta_{\nu_{1,2}}$	1.107	1.104	1.077	1.000	1.074	0.994	1.018
	$\Delta_{\nu_1^1}$	1.096	1.100	1.088	1.000	1.058	1.005	1.051
	$\Delta_{\nu_{-1}^1}$	1.197	1.177	1.137	1.000	1.092	1.017	0.996

plied by setting $\tau = 0.2$. In this section, we restrict our attention to five assignment mechanisms: **B-B**, **C**, **MT**, **Large-2,** and **Large-4**. We exclude **MP-B** because it is a nonstandard experimental design for which we have not developed an inference procedure. We also exclude the rerandomization design (**RE**) because, although it is a widely studied design, the inferential results in Li, Ding, and Rubin (2020) are derived in a finite population framework, which is distinct from our superpopulation framework, and their resulting limiting distribution is nonnormal.

In each case, we perform our hypothesis tests at a significance level of 0.05. For design **B-B**, tests are performed using a standard t-test. For designs **C**, **Large-2,** and **Large-4,** the tests are constructed using the asymptotic normality result from Theorem 3.7 combined with variance estimators constructed using the same plug-in method as in

TABLE 4. Rejection probabilities under the null and alternative hypothesis.

		Under H_0					Under H_1				
Model	Parameter	**B-B**	**C**	**MT**	**Large-2**	**Large-4**	**B-B**	**C**	**MT**	**Large-2**	**Large-4**
1	Δ_{ν_1}	0.057	0.049	0.051	0.050	0.046	0.790	0.803	0.977	0.915	0.963
	Δ_{ν_2}	0.052	0.059	0.046	0.060	0.058	0.371	0.403	0.675	0.534	0.593
	$\Delta_{\nu_{1,2}}$	0.049	0.059	0.049	0.059	0.043	0.081	0.093	0.126	0.100	0.106
	$\Delta_{\nu_1^1}$	0.052	0.043	0.048	0.064	0.040	0.646	0.656	0.921	0.816	0.884
	$\Delta_{\nu_{-1}^1}$	0.056	0.051	0.044	0.057	0.048	0.361	0.333	0.594	0.499	0.545
2	Δ_{ν_1}	0.053	0.043	0.049	0.048	0.045	0.738	0.737	0.976	0.875	0.951
	Δ_{ν_2}	0.056	0.061	0.046	0.059	0.056	0.341	0.377	0.670	0.483	0.551
	$\Delta_{\nu_{1,2}}$	0.052	0.065	0.050	0.060	0.044	0.082	0.091	0.126	0.101	0.095
	$\Delta_{\nu_1^1}$	0.049	0.051	0.046	0.057	0.036	0.597	0.610	0.919	0.758	0.840
	$\Delta_{\nu_{-1}^1}$	0.056	0.051	0.046	0.054	0.048	0.340	0.310	0.598	0.436	0.500
3	Δ_{ν_1}	0.054	0.056	0.050	0.053	0.052	0.571	0.570	0.837	0.705	0.787
	Δ_{ν_2}	0.056	0.057	0.056	0.057	0.059	0.235	0.259	0.361	0.286	0.323
	$\Delta_{\nu_{1,2}}$	0.051	0.051	0.052	0.062	0.047	0.060	0.064	0.116	0.091	0.082
	$\Delta_{\nu_1^1}$	0.048	0.051	0.046	0.061	0.035	0.402	0.421	0.885	0.624	0.762
	$\Delta_{\nu_{-1}^1}$	0.061	0.047	0.060	0.056	0.057	0.255	0.234	0.374	0.310	0.340
4	Δ_{ν_1}	0.049	0.051	0.045	0.045	0.050	0.908	0.905	0.968	0.956	0.957
	Δ_{ν_2}	0.051	0.052	0.051	0.051	0.058	0.488	0.520	0.604	0.569	0.559
	$\Delta_{\nu_{1,2}}$	0.056	0.052	0.049	0.065	0.045	0.092	0.102	0.126	0.117	0.111
	$\Delta_{\nu_1^1}$	0.050	0.048	0.051	0.054	0.045	0.762	0.785	0.908	0.865	0.886
	$\Delta_{\nu_{-1}^1}$	0.044	0.055	0.048	0.052	0.046	0.498	0.472	0.544	0.528	0.523
5	Δ_{ν_1}	0.054	0.054	0.045	0.045	0.043	0.844	0.847	0.964	0.912	0.937
	Δ_{ν_2}	0.053	0.056	0.051	0.048	0.053	0.416	0.445	0.589	0.491	0.505
	$\Delta_{\nu_{1,2}}$	0.052	0.054	0.049	0.059	0.049	0.092	0.099	0.124	0.110	0.099
	$\Delta_{\nu_1^1}$	0.051	0.052	0.049	0.058	0.043	0.674	0.688	0.911	0.810	0.847
	$\Delta_{\nu_{-1}^1}$	0.050	0.062	0.049	0.056	0.049	0.416	0.403	0.523	0.461	0.474
6	Δ_{ν_1}	0.050	0.050	0.043	0.058	0.043	0.129	0.128	0.122	0.115	0.130
	Δ_{ν_2}	0.053	0.059	0.057	0.057	0.051	0.074	0.086	0.088	0.079	0.080
	$\Delta_{\nu_{1,2}}$	0.047	0.046	0.052	0.053	0.044	0.052	0.046	0.052	0.057	0.050
	$\Delta_{\nu_1^1}$	0.049	0.046	0.049	0.051	0.043	0.082	0.083	0.077	0.082	0.081
	$\Delta_{\nu_{-1}^1}$	0.059	0.056	0.058	0.059	0.056	0.140	0.113	0.125	0.131	0.135

Bugni, Canay, and Shaikh (2018) and Bugni, Canay, and Shaikh (2019). For design **MT,** the test is constructed as described in Theorem 3.2. Table 4 displays the rejection probabilities under the null and alternative hypotheses, computed from 2000 Monte Carlo replications. The results show that the rejection probabilities are universally around 0.05 under the null hypothesis, which verifies the validity of our tests across all the designs. Under the alternative hypotheses implied by $\tau = 0.2$, the rejection probabilities vary substantially across the different designs, outcome models, and parameters. However, our matched tuples design displays the highest power for almost all parameters and model specifications.

4.3 *Experiments with more factors and covariates*

In this section, we repeat the previous simulation exercises while varying the number of factors K and the number of observed covariates $\dim(X_i)$. The data generating process is constructed as follows:

$$Y_i(d) = \begin{cases} \tau d^{(1)} + \tilde{X}_i'\beta + \epsilon_i & \text{if } K = 1, \\ \tau \cdot \left(d^{(1)} + \dfrac{1}{K-1} \displaystyle\sum_{2 \le k \le K} d^{(k)} \right) + \gamma_d \tilde{X}_i'\beta + \epsilon_i & \text{if } K \ge 2, \end{cases}$$

where $\tau \in \{0, 0.1\}$, $d = (d^{(1)}, \ldots, d^{(K)})$, and $d^{(k)} \in \{-1, 1\}$ represents the treatment status of the kth factor. We set $\gamma_d = 1$ if $d^{(2)} = 1$, $\gamma_d = -1$ otherwise, in order to ensure the conditional means are heterogeneous in the second factor. $\tilde{X}_i$ contains 9 covariates, out of which the first $\dim(X_i)$ covariates are observed and used for the experimental designs. The distributions of $\tilde{X}_i$, ϵ_i and the values of β are calibrated using data obtained from Branson, Dasgupta, and Rubin (2016), who study the covariate balancing properties of 2^K factorial rerandomization designs using data from the New York Department of Education (NYDE). Details on the empirical context and construction of the data generating process are provided in Appendix D.3 in the Supplemental Appendix.

To construct our matched tuples of size 2^K when $\dim(X_i) > 1$, we employ the recursive pairing algorithm described in Section 2 using the Mahalanobis distance. We emphasize, however, that this approach is not guaranteed to be optimal, and we leave the study of potentially more effective matching algorithms to future work.

In addition to the standard matched tuples design (**MT**), we also include a matched tuples design with a *replicate* for each treatment as described in Section 3.2, denoted by **MT2**. For example, in the **MT2** design with two factors, units are matched into groups of *eight*, and two units receive each factor combination. We also continue to consider the alternative designs (**C**, **Large-4**, **MP-B,** and **RE**) from Section 4.1. When constructing the strata for **Large-4**, we stratify on one covariate drawn at random from the set of available covariates.

In Table 5, we report the ratio of the MSE of each design relative to the MSE of **MT** when $\dim(X_i) = 1$ and $K = 1$ (computed from 4000 Monte Carlo replications). For all experiments in this section, the number of observations is fixed to be 1280 so that we have 20 matched tuples of size 64 when $K = 6$. Our simulation results are consistent with those in Section 4.1: **MT** displays the lowest MSE across almost all model specifications. Although **MT2** generally produces larger MSEs than **MT**, it still performs favorably relative to the other designs. For methods that use an increasing number of covariates when $\dim(X_i)$ increases (**MT**, **MT2**, **MP-B,** and **RE**), we observe that the MSE in fact *increases* with the number of available covariates. We expect this is because (as shown in Appendix D.3 in the Supplemental Appendix) the first covariate is a much stronger predictor of the control outcome than the other available covariates, which are relatively uninformative.

In Table 6, we compute the rejection probabilities when testing the null hypothesis $H_0 : \Delta_{\nu_1} = 0$ against the alternative implied by setting $\tau = 0.1$, for various choices of K

TABLE 5. Ratio of MSEs relative to MT using a single factor and covariate.

$\dim(X_i)$	Method	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	Method	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
1	**MT**	1.000	1.003	1.006	1.113	1.297	1.945	**C**	9.151	8.554	8.642	8.939	9.015	9.181
2		1.027	1.052	1.107	1.180	1.463	2.293		9.120	8.528	8.568	8.867	9.053	9.114
4		1.043	1.130	1.420	1.687	2.170	3.338		8.968	8.364	8.569	8.868	8.949	8.765
6		1.192	1.495	1.763	2.241	3.097	4.304		8.945	8.327	8.588	8.994	9.081	8.853
9		1.284	1.702	2.047	2.781	3.337	4.081		8.934	8.309	8.600	8.788	8.915	8.526
1	**MT2**	1.017	1.049	1.074	1.297	1.916	2.903	**Large-4**	4.393	4.605	4.674	4.634	4.393	4.381
2		1.044	1.086	1.212	1.547	2.200	3.585		6.523	6.926	6.745	6.704	6.521	6.367
4		1.224	1.332	1.620	2.231	3.379	4.799		7.321	8.100	7.407	7.559	7.542	7.399
6		1.451	1.901	2.339	3.061	4.020	5.721		8.143	8.137	7.644	7.801	8.288	7.906
9		1.609	2.140	2.693	3.231	4.387	6.903		8.093	8.075	8.170	7.799	8.129	8.402
1	**MP-B**	0.991	8.693	8.807	8.964	8.991	8.829	**RE**	1.073	1.091	1.296	2.032	3.040	3.640
2		0.978	8.854	8.897	8.863	8.811	9.072		1.090	1.069	1.955	3.284	4.282	5.094
4		0.967	8.970	8.711	9.020	8.855	8.749		1.320	1.410	3.278	4.640	5.504	6.270
6		1.175	9.148	8.753	8.941	8.774	8.596		1.961	1.886	3.976	5.648	6.223	6.759
9		1.227	8.793	8.989	9.444	9.227	8.273		2.515	2.566	4.957	6.265	6.676	7.455

and $\dim(X_i)$ (computed from 1000 Monte Carlo replications). Under the null hypothesis, we observe that our tests under design **MT** become conservative as $\dim(X_i)$ and K increase. In particular, we notice a large difference between $K=4$ and $K=5$. However, despite being conservative, **MT** still displays favorable power properties relative to **C** and **Large-4** for all but the largest choices of K.

Our next observation is that our tests under design **MT2** remain exact even as $\dim(X_i)$ and K both increase. As we explain in Section 3.2, we suspect that our challenges for inference using **MT** come from poor estimation of the variance, which seems to be alleviated in **MT2**, where the number of observations receiving each treatment within a tuple are doubled. As a result of this exactness, **MT2** achieves higher power than **MT** when $\dim(X_i)$ and K are large. To further explore these power improvements, Figure 1 presents power plots for three specific choices of K and $\dim(X_i)$ with τ ranging from 0 to 0.1 (Figure D.1 in the Supplemental Appendix presents power plots for alternatives implied by larger values than $\tau=0.1$). First, when $\dim(X_i)$ and K are small, for instance, $\dim(X_i)=K=1$, we observe no significant difference between the power plots generated by **MT** and **MT2**. However, when the dimension of the covariates and factors are both large, for instance, $\dim(X_i)=6$, $K=4$, **MT2** dominates **MT** for all alternative hypotheses. Therefore, our recommendation to practitioners is to consider a matched tuples design when working with few treatments and covariates, but to consider the replicated design when dealing with a large number of treatments and/or covariates.

TABLE 6. Rejection probabilities when testing $H_0 : \Delta_{\nu_1} = 0$ under the null and alternative hypothesis.

		Under H_0						Under H_1					
Method	$\dim(X_i)$	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
MT	1	0.049	0.045	0.033	0.023	0.009	0.008	0.998	1.000	1.000	0.997	0.980	0.837
	2	0.047	0.043	0.041	0.018	0.008	0.002	0.999	0.998	0.997	0.997	0.935	0.732
	4	0.040	0.029	0.031	0.011	0.009	0.008	1.000	1.000	0.979	0.946	0.794	0.583
	6	0.037	0.018	0.010	0.022	0.010	0.007	0.999	0.989	0.936	0.870	0.668	0.479
	9	0.041	0.026	0.016	0.019	0.014	0.003	0.988	0.961	0.895	0.810	0.674	0.319
MT2	1	0.054	0.054	0.044	0.059	0.047	0.052	1.000	0.999	1.000	0.996	0.973	0.858
	2	0.048	0.053	0.041	0.058	0.039	0.055	1.000	0.999	1.000	0.985	0.943	0.784
	4	0.075	0.048	0.054	0.056	0.060	0.046	0.996	0.993	0.981	0.951	0.843	0.673
	6	0.053	0.067	0.046	0.054	0.045	0.046	0.988	0.967	0.926	0.857	0.744	0.579
	9	0.065	0.050	0.053	0.059	0.060	0.047	0.983	0.944	0.872	0.840	0.704	0.494
C	1	0.062	0.054	0.041	0.056	0.059	0.069	0.437	0.449	0.410	0.445	0.463	0.459
	2	0.063	0.049	0.038	0.051	0.065	0.068	0.434	0.450	0.410	0.442	0.459	0.459
	4	0.064	0.050	0.038	0.048	0.055	0.057	0.425	0.448	0.400	0.443	0.457	0.468
	6	0.066	0.052	0.045	0.048	0.054	0.055	0.430	0.437	0.409	0.436	0.437	0.463
	9	0.063	0.042	0.050	0.033	0.054	0.048	0.417	0.439	0.420	0.433	0.433	0.448
Large-4	1	0.050	0.044	0.059	0.061	0.053	0.057	0.685	0.699	0.701	0.683	0.730	0.770
	2	0.046	0.050	0.043	0.052	0.044	0.065	0.560	0.564	0.575	0.585	0.582	0.634
	4	0.053	0.064	0.039	0.059	0.056	0.062	0.497	0.490	0.486	0.527	0.521	0.577
	6	0.055	0.053	0.049	0.057	0.059	0.071	0.462	0.444	0.495	0.519	0.520	0.553
	9	0.044	0.041	0.056	0.051	0.049	0.076	0.457	0.451	0.493	0.490	0.511	0.571

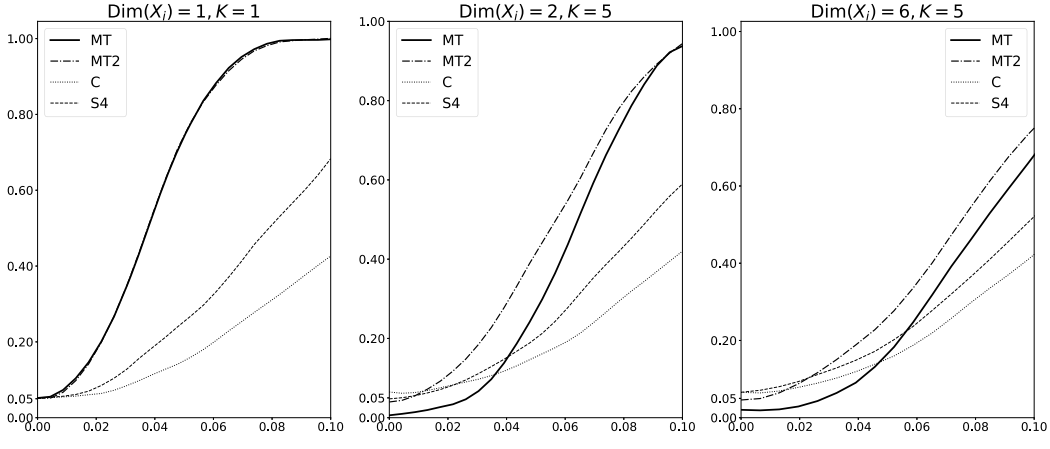

FIGURE 1. Rejection probability under various choices of τ.

5. Empirical application

In this section, we illustrate the inference procedures introduced in Section 3 using the data collected in Fafchamps et al. (2014).[3] Fafchamps et al. (2014) conduct a randomized experiment in order to investigate the effects of several capital aid programs on the profits of small businesses in Ghana. In their experiment, there are three treatment arms, where (in our notation) $D_i = 1$ indicates that the ith firm is untreated, $D_i = 2$ indicates being offered cash, and $D_i = 3$ indicates being offered in-kind grants. The null hypotheses of interest are

$$H_0^d : E\big[Y_i(1)\big] = E\big[Y_i(d)\big] \quad \text{versus} \quad H_1 : E\big[Y_i(1)\big] \neq E\big[Y_i(d)\big] \tag{12}$$

for $d \in \{2, 3\}$, as well as

$$H_0^{2,3} : E\big[Y_i(2)\big] = E\big[Y_i(3)\big] \quad \text{versus} \quad H_1 : E\big[Y_i(2)\big] \neq E\big[Y_i(3)\big]. \tag{13}$$

In their experimental design, blocks are defined by quadruplets, where each quadruplet contains *two* untreated units with $D_i = 1$, one treated unit with $D_i = 2$, and one treated unit with $D_i = 3$. Despite the slight departure from the framework presented in Sections 2–3, in that there are two untreated units in each quadruplet, we show in Appendix A.1 that a slight modification of the variance estimator in Theorem 3.2 produces a valid test for (12)–(13). Specifically, we pretend that there are four treatment levels in each quadruplet, while the first two are in fact controls. Then, by setting generating vectors $\nu^2 = (-1/2, -1/2, 1, 0)$, $\nu^3 = (-1/2, -1/2, 0, 1)$, and $\nu^{2,3} = (0, 0, -1, 1)$ and proceeding with the testing procedure in Theorem 3.2, we obtain valid tests for H_0^d and $H_0^{2,3}$. For each of the hypotheses in (12)–(13), we implement the following tests:

- A t-test based on the OLS estimator in a linear regression of Y on 1, $I\{D_i = 2\}$, and $I\{D_i = 3\}$, together with the usual heteroskedasticity-robust variance estimator.
- The test introduced in Proposition A.1, which implements the test from Theorem 3.2 as described above to accommodate for the fact that there are two untreated units in each block.

We note that Fafchamps et al. (2014) test (12) and (13) using a t-test obtained from a linear regression of outcomes on treatment indicators and block fixed effects. However, as was shown in Theorem 3.3, such a procedure is not guaranteed to be valid. On the other hand, we expect that the t-test obtained from a linear regression without block fixed effects should be conservative for testing (12)–(13) in light of the observations made in Example 3.1 and the fact that this test coincides with a standard two-sample t-test.

Our results are presented in Table 7. The point estimates of the two methods are identical because the OLS estimator coincides with the difference-in-means estimator.

[3]The original paper features six rounds of surveys, which were pooled in the final analysis. We perform our analysis exclusively on the data obtained in the sixth round in order to avoid complications related to time-series dependence across rounds. For simplicity, we additionally drop quadruplets with missing values, and 4 "leftover" groups whose sizes range from 5 to 8 firms. This results in a final sample of 120 quadruplets, or $4n = 480$. Further results on the long-run effects (collected in a seventh survey wave) are contained in Table D.4 in Section D of the Supplemental Appendix.

TABLE 7. Point estimates and standard errors for testing the treatment effects of cash and in-kind grants using different methods (wave 6).

		All Firms (1)	Males (2)	Females (3)	High Initial Profit Women (4)	Low Initial Profit Women (5)
OLS (standard t-test)	Cash treatment	19.64	24.84	16.30	33.09	7.01
		(15.42)	(27.29)	(18.13)	(42.56)	(11.58)
	In-kind treatment	20.26	4.48	30.42	65.36	11.10
		(15.67)	(18.42)	(22.83)	(53.28)	(15.31)
	Cash = in-kind (p-val)	0.975	0.493	0.600	0.610	0.817
Difference-in-means (adjusted t-test)	Cash treatment	19.64	24.84	16.30	33.09	7.01
		(14.24)	(26.05)	(15.21)	(39.27)	(11.15)
	In-kind treatment	20.26	4.48	30.42	65.36	11.10
		(15.24)	(17.79)	(21.97)	(48.27)	(14.99)
	Cash = in-kind (p-val)	0.974	0.468	0.567	0.576	0.815

Note: The results in this table are based on the data from the sixth wave of data collection. For each treatment and each subsample, the number in the first row is the point estimate and that in the second row is the standard error. For testing the equality of the average potential outcomes under the two values of treatment, we report the p-values as in Fafchamps et al. (2014).

However, the standard errors obtained from our variance estimator are always smaller than the heteroskedasticy-robust standard errors. For example, when testing (12) for $d = 3$ among the female subsample, the standard error produced from our variance estimator is 15.21 whereas the heteroskedasticy robust standard error is 18.13. We note that overall the improvements are modest; this suggests that the conditional expectation of the outcomes does not vary substantially with the observable characteristics in this survey wave. This is further corroborated by the calibrated simulations presented in Table D.2 in Appendix D in the Supplemental Appendix.

6. RECOMMENDATIONS FOR EMPIRICAL PRACTICE

We conclude with some recommendations for empirical practice based on our theoretical results as well as the simulation study above. For inference about the linear contrast of expected outcomes given by Δ_ν in a matched tuples design, we recommend the test ϕ_n^ν defined in Section 3.1: Our simulations results show that this test does a good job of controlling size in large samples (approximately 80 blocks). We have shown that tests based on the heteroskedasticity-robust variance estimator from a linear regression of outcomes on treatment and block fixed effects may be *invalid*, in the sense of having rejection probability strictly greater than the nominal level under the null hypothesis. Tests based on the heteroskedasticity-robust variance or block-cluster variance estimators from a linear regression of outcomes on treatment are valid but potentially conservative, which would result in a loss of power relative to our proposed test.

We also find that matched tuples designs have favorable efficiency properties relative to other popular designs (with a specific illustration in the setting of 2^K factorial designs). However, this comes with the caveat that when dealing with a large number of

treatments (in our simulations, this translated to having fewer than 80 blocks) and/or large number of covariates, practitioners may want to consider the replicated matched tuples design introduced in Section 3.2, as our simulations suggest that this design may have more robust size control, which translates to better power in such cases.

Appendix A: Additional details

A.1 *Details for Section 5*

Proposition A.1. *Consider the setting with three treatment statuses* $\{1, 2, 3\}$, *where* 1 *corresponds to being untreated and* 2 *and* 3 *correspond to two treatments. In a matched quadruplets design where each quadruplet has two untreated units and one unit for each treatment, the test introduced in Section* 3.1 *with* $\mathcal{D}' = \{1, 2, 3, 4\}$ *and*

$$\nu = \begin{pmatrix} -1/2 & -1/2 & 1 & 0 \\ -1/2 & -1/2 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

is valid for testing (12)–(13) *at level* $\alpha \in (0, 1)$.

Proof of Proposition A.1. Consider a design of matched quadruplets with two treatments $d = 2, 3$ and two controls $d = 1$, that is, a quadruplet consisting of $(1, 1, 2, 3)$. The difference-in-mean estimator for the effect of the first treatment $d = 2$ is

$$\hat{\Delta}_2 = \frac{1}{n}\sum_{i=1}^{4n} I\{D_i = 2\}Y_i - \frac{1}{2n}\sum_{i=1}^{4n} I\{D_i = 1\}Y_i.$$

Note that

$$\sqrt{n}\big(\hat{\Delta}_2 - \Delta_2(Q)\big) = A_{n,2} + C_{n,2} - (A_{n,1} + C_{n,1}),$$

where

$$A_{n,2} = \frac{1}{\sqrt{n}}\sum_{1\le i\le 4n} I\{D_i = 2\}\big(Y_i(2) - E\big[Y_i(2)|X^{(n)}, D^{(n)}\big]\big),$$

$$C_{n,2} = \frac{1}{\sqrt{n}}\sum_{1\le i\le 4n} I\{D_i = 2\}\big(E\big[Y_i(2)|X^{(n)}, D^{n}\big] - E\big[Y_i(2)\big]\big)$$

and

$$A_{n,1} = \frac{1}{2\sqrt{n}}\sum_{1\le i\le 4n} I\{D_i = 1\}\big(Y_i(1) - E\big[Y_i(1)|X^{(n)}, D^{(n)}\big]\big),$$

$$C_{n,1} = \frac{1}{2\sqrt{n}}\sum_{1\le i\le 4n} I\{D_i = 1\}\big(E\big[Y_i(1)|X^{(n)}, D^{n}\big] - E\big[Y_i(1)\big]\big).$$

Let I_j denote the set of indices for the two untreated units in the jth tuple. Note

$$\begin{aligned}
&\operatorname{Var}\bigl[A_{n,1}|X^{(n)}, D^{(n)}\bigr] \\
&\quad= \frac{1}{2\cdot 2n}\sum_{1\le i\le 4n} I\{D_i = 1\}\operatorname{Var}\bigl[Y_i(1)|X_i\bigr] \\
&\quad= \frac{1}{2\cdot 4n}\sum_{1\le i\le 4n}\operatorname{Var}\bigl[Y_i(1)|X_i\bigr] - \frac{1}{8n}\sum_{1\le j\le n}\frac{1}{2}\sum_{i_j\in I_j}\sum_{k\in\lambda_j: k\notin I_j}\bigl(\operatorname{Var}\bigl[Y_k(1)|X_k\bigr] - \operatorname{Var}[Y_{i_j}|X_{i_j}]\bigr).
\end{aligned}$$

It follows from similar arguments as in the proof of Theorem 3.1 that the second term goes to zero. Therefore,

$$\operatorname{Var}\bigl[A_{n,1}|X^{(n)}, D^{(n)}\bigr] \xrightarrow{P} \frac{1}{2}E\bigl[\operatorname{Var}\bigl[Y_i(1)|X_i\bigr]\bigr].$$

It therefore follows from Lemma S.1.2 of Bai, Romano, and Shaikh (2021) that

$$\gamma\left(\bigl((A_{n,2}, A_{n,1})'\bigl|X^{(n)}, D^{(n)}\bigr), N\left(0, \begin{bmatrix} E\bigl[\operatorname{Var}\bigl[Y_i(2)|X_i\bigr]\bigr] & 0 \\ 0 & \frac{1}{2}E\bigl[\operatorname{Var}\bigl[Y_i(1)|X_i\bigr]\bigr] \end{bmatrix}\right)\right) \xrightarrow{P} 0,$$

where γ is any metric that metrizes weak convergence.

Next, note

$$\begin{aligned}
E\bigl[C_{n,2}|X^{(n)}\bigr] &= \frac{1}{\sqrt{n}}\sum_{1\le i\le 4n}\frac{1}{4}\bigl(E\bigl[Y_i(2)|X^{(n)}\bigr] - E\bigl[Y_i(2)\bigr]\bigr) = \frac{1}{4\sqrt{n}}\sum_{1\le i\le 4n}\bigl(\Gamma_2(X_i) - \Gamma_2\bigr), \\
E\bigl[C_{n,1}|X^{(n)}\bigr] &= \frac{1}{2\sqrt{n}}\sum_{1\le i\le 4n}\frac{1}{2}\bigl(E\bigl[Y_i(1)|X^{(n)}\bigr] - E\bigl[Y_i(1)\bigr]\bigr) = \frac{1}{4\sqrt{n}}\sum_{1\le i\le 4n}\bigl(\Gamma_1(X_i) - \Gamma_1\bigr).
\end{aligned}$$

Therefore,

$$(C_{n,2}, C_{n,1})' \xrightarrow{d} N\left(0, \frac{1}{4}\begin{bmatrix} \operatorname{Var}\bigl(\Gamma_2(X_i)\bigr) & \operatorname{Cov}\bigl(\Gamma_2(X_i), \Gamma_1(X_i)\bigr) \\ \operatorname{Cov}\bigl(\Gamma_2(X_i), \Gamma_1(X_i)\bigr) & \operatorname{Var}\bigl(\Gamma_1(X_i)\bigr)] \end{bmatrix}\right).$$

It then follows from Lemma S.1.2 of Bai, Romano, and Shaikh (2021) that

$$\sqrt{n}\bigl(\hat{\Delta}_2 - \Delta_2(Q)\bigr) \xrightarrow{d} N(0, \mathbb{V}_2),$$

where

$$\begin{aligned}
\mathbb{V}_2 = {} & E\bigl[\operatorname{Var}\bigl[Y_i(2)|X_i\bigr]\bigr] + \frac{1}{2}E\bigl[\operatorname{Var}\bigl[Y_i(1)|X_i\bigr]\bigr] \\
& + \frac{1}{4}\bigl(\operatorname{Var}\bigl(\Gamma_2(X_i)\bigr) + \operatorname{Var}\bigl(\Gamma_1(X_i)\bigr) - 2\operatorname{Cov}\bigl(\Gamma_2(X_i), \Gamma_1(X_i)\bigr)\bigr).
\end{aligned}$$

Now, suppose we pretend the two untreated units are assigned to two distinct treatment levels and denote the two untreated levels and two treated levels by $d \in \{1, 2, 3, 4\}$,

where $d = 1, 2$ actually corresponds to the untreated units. Our estimand can then be defined as

$$\tilde{\Delta}_2(Q) = \Gamma_3(Q) - \frac{1}{2}\big(\Gamma_1(Q) + \Gamma_2(Q)\big).$$

Applying the existing results in Theorem 3.1 with $\nu = (-1/2, -1/2, 1, 0)$. It follows that

$$\sqrt{n}\big(\hat{\Delta}_2 - \tilde{\Delta}_2(Q)\big) \xrightarrow{d} N(0, \tilde{\mathbb{V}}_2),$$

where

$$\begin{aligned}\tilde{\mathbb{V}}_2 &= E\big[\mathrm{Var}\big[Y_i(3)|X_i\big]\big] + \frac{1}{4}\big(E\big[\mathrm{Var}\big[Y_i(1)|X_i\big]\big] + E\big[\mathrm{Var}\big[Y_i(2)|X_i\big]\big]\big) \\ &\quad + \frac{1}{4}\bigg(\mathrm{Var}\big(\Gamma_3(X_i)\big) + \frac{1}{4}\mathrm{Var}\big(\Gamma_1(X_i)\big) + \frac{1}{4}\mathrm{Var}\big(\Gamma_2(X_i)\big) + \frac{1}{2}\mathrm{Cov}\big(\Gamma_2(X_i), \Gamma_1(X_i)\big) \\ &\quad - \mathrm{Cov}\big(\Gamma_3(X_i), \Gamma_1(X_i)\big) - \mathrm{Cov}\big(\Gamma_3(X_i), \Gamma_2(X_i)\big)\bigg) \\ &= \mathbb{V}_2,\end{aligned}$$

where the last equality follows by setting $d = 1, 2, 3$ to $d = 1, 1, 2$. The same argument holds for $v = (-1/2, -1/2, 0, 1)$. As for $\nu = (0, 0, -1, 1)$, the estimation and inference of the third and fourth arms is not affected by treatment status in the first two arms. Therefore, pretending two controls are two different treatment levels yields the true asymptotic variance, meaning that the inference is still valid. □

Appendix B: Proofs of main results

B.1 *Proof of Theorem* 3.1

We derive the limiting distribution of $\sqrt{n}(\hat{\Gamma}_n(d) - \Gamma_d(Q) : d \in \mathcal{D})$, from which the conclusion of the theorem follows by an application of the continuous mapping theorem. Note that

$$\sqrt{n}\big(\hat{\Gamma}_n(d) - \Gamma_d(Q) : d \in \mathcal{D}\big)' = A_n + C_n,$$

where $A_n = (A_{n,d} : d \in \mathcal{D})'$, $C_n = (C_{n,d} : d \in \mathcal{D})'$, and

$$A_{n,d} = \frac{1}{\sqrt{n}} \sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = d\}\big(Y_i(d) - E\big[Y_i(d)|X^{(n)}, D^{(n)}\big]\big),$$

$$C_{n,d} = \frac{1}{\sqrt{n}} \sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = d\}\big(E\big[Y_i(d)|X^{(n)}, D^n\big] - E\big[Y_i(d)\big]\big).$$

Note that conditional on $X^{(n)}$, $D^{(n)}$, $C_{n,d}$'s are constants, and $A_{n,d}$'s are independent. By Assumption 2.2, for $d \in \mathcal{D}$, $E[Y_i(d)|X^{(n)}, D^{(n)}] = E[Y_i(d)|X_i]$. Fix $d \in \mathcal{D}$. Let $i_j \in \lambda_j$ be such that $D_{i_j} = d$. Note

$$\mathrm{Var}\big[A_{n,d}|X^{(n)}, D^{(n)}\big]$$

$$= \frac{1}{n} \sum_{1 \le i \le |\mathcal{D}|n} I\{D_i = d\} \operatorname{Var}\big[Y_i(d)|X_i\big]$$

$$= \frac{1}{|\mathcal{D}|n} \sum_{1 \le i \le |\mathcal{D}|n} \operatorname{Var}\big[Y_i(d)|X_i\big] - \frac{1}{|\mathcal{D}|n} \sum_{1 \le j \le n} \sum_{k \in \lambda_j : k \ne i_j} \big(\operatorname{Var}\big[Y_k(d)|X_k\big] - \operatorname{Var}\big[Y_{i_j}(d)|X_{i_j}\big]\big),$$

where the first equality follows from Assumption 2.2. By Assumption 2.1(b) and the weak law of large numbers,

$$\frac{1}{|\mathcal{D}|n} \sum_{1 \le i \le |\mathcal{D}|n} \operatorname{Var}\big[Y_i(d)|X_i\big] \xrightarrow{P} E\big[\operatorname{Var}\big[Y_i(d)|X_i\big]\big].$$

By Assumptions 2.1(c) and 2.3, we have

$$\begin{aligned}
&\left| \frac{1}{|\mathcal{D}|n} \sum_{1 \le j \le n} \sum_{k \in \lambda_j : k \ne i_j} \big(\operatorname{Var}\big[Y_k(d)|X_k\big] - \operatorname{Var}\big[Y_{i_j}(d)|X_{i_j}\big]\big) \right| \\
&\quad \le \frac{1}{|\mathcal{D}|n} \sum_{1 \le j \le n} \sum_{k \in \lambda_j : k \ne i_j} \big|\operatorname{Var}\big[Y_k(d)|X_k\big] - \operatorname{Var}\big[Y_{i_j}(d)|X_{i_j}\big]\big| \\
&\quad \lesssim \frac{1}{n} \sum_{1 \le j \le n} \sum_{k \in \lambda_j : k \ne i_j} \|X_k - X_{i_j}\| \\
&\quad \le \frac{|\mathcal{D}| - 1}{n} \sum_{1 \le j \le n} \max_{i,k \in \lambda_j} \|X_i - X_k\| \xrightarrow{P} 0.
\end{aligned}$$

Therefore, $\operatorname{Var}[A_{n,d}|X^{(n)}, D^{(n)}] \xrightarrow{P} E[\operatorname{Var}[Y_i(d)|X_i]]$. We can then verify Lindeberg's condition as in the proof of Lemma S.1.4 of Bai, Romano, and Shaikh (2021). It follows that

$$\gamma\big(\big((A_{n,d} : d \in \mathcal{D})' \big| X^{(n)}, D^{(n)}\big)\big), N(0, \mathbb{V}_1)\big) \xrightarrow{P} 0,$$

where $\mathbb{V}_1 = \operatorname{diag}(E[\operatorname{Var}[Y_i(d)|X_i]] : d \in \mathcal{D})$ and γ is any metric that metrizes weak convergence.

Next,

$$E\big[C_{n,d}|X^{(n)}\big] = \frac{1}{|\mathcal{D}|\sqrt{n}} \sum_{1 \le i \le |\mathcal{D}|n} \big(\Gamma_d(X_i) - \Gamma_d\big)$$

and

$$\begin{aligned}
\operatorname{Var}\big[C_{n,d}|X^{(n)}\big] &= \frac{1}{n} \sum_{1 \le j \le n} \sum_{i \in \lambda_j} \frac{1}{|\mathcal{D}|} \left(\Gamma_d(X_i) - \frac{1}{|\mathcal{D}|} \sum_{k \in \lambda_j} \Gamma_d(X_k) \right)^2 \\
&\lesssim \frac{1}{n} \sum_{1 \le j \le n} \max_{i,k \in \lambda_j} \|X_i - X_k\|^2 \xrightarrow{P} 0
\end{aligned}$$

by Assumption 2.1(c) and 2.3. Therefore, by repeating the argument, which establishes (S.24) in the proof of Lemma S.1.4 of Bai, Romano, and Shaikh (2021), it follows that

$$C_{n,d} = \frac{1}{|\mathcal{D}|\sqrt{n}} \sum_{1 \le i \le |\mathcal{D}|n} \big(\Gamma_d(X_i) - \Gamma_d\big) + o_P(1).$$

Therefore,

$$(C_{n,d} : d \in \mathcal{D}) \xrightarrow{d} N(0, \mathbb{V}_2),$$

where $(\mathbb{V}_2)_{d,d'} = \frac{1}{|\mathcal{D}|} \operatorname{Cov}(\Gamma_d(X_i), \Gamma_{d'}(X_i))$. It then follows from Lemma S.1.2 of Bai, Romano, and Shaikh (2021) that

$$\sqrt{n}\big(\hat{\Gamma}_n(d) - \Gamma_d : d \in \mathcal{D}\big)' \xrightarrow{d} N(0, \mathbb{V}_1 + \mathbb{V}_2).$$

The conclusion now follows. □

B.2 *Proof of Theorem 3.2*

The conclusion follows from Lemmas C.1–C.3 together with the continuous mapping theorem. □

B.3 *Proof of Theorem 3.3*

Define

$$C_i = \big(I\{D_i = 2\}, \ldots, I\big\{D_i = |\mathcal{D}|\big\}\big)'.$$

To begin, note it follows from the Frisch–Waugh–Lovell theorem and Assumption 2.2 that

$$\begin{pmatrix} \hat{\beta}_n(2) \\ \vdots \\ \hat{\beta}_n\big(|\mathcal{D}|\big) \end{pmatrix} = \Bigg(\sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i \tilde{C}_i' \Bigg)^{-1} \sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i Y_i,$$

where

$$\tilde{C}_i = \left(I\{D_i = 2\} - \frac{1}{|\mathcal{D}|}, \ldots, I\big\{D_i = |\mathcal{D}|\big\} - \frac{1}{|\mathcal{D}|} \right)'.$$

Next, note for

$$\sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i \tilde{C}_i',$$

the diagonal entries are $\frac{|\mathcal{D}|-1}{|\mathcal{D}|} n$ and the off-diagonal entries are $-\frac{1}{|\mathcal{D}|} n$. It follows from element calculation that the diagonal entries of $(\sum_{1 \le i \le |\mathcal{D}|n} \tilde{C}_i \tilde{C}_i')^{-1}$ are $\frac{2}{n}$ and the off-

diagonal entries are $\frac{1}{n}$. Furthermore,

$$\sum_{1\le i\le|\mathcal{D}|n}\tilde{C}_iY_i=\begin{pmatrix}n\hat{\Gamma}_n(2)-\dfrac{1}{|\mathcal{D}|}\displaystyle\sum_{1\le i\le|\mathcal{D}|n}Y_i\\ \vdots\\ n\hat{\Gamma}_n\big(|\mathcal{D}|\big)-\dfrac{1}{|\mathcal{D}|}\displaystyle\sum_{1\le i\le|\mathcal{D}|n}Y_i\end{pmatrix}.$$

Therefore, for $d\in\mathcal{D}\backslash\{1\}$,

$$\begin{aligned}\hat{\beta}_n(d)&=\frac{2}{n}\bigg(n\hat{\Gamma}_n(d)-\frac{1}{|\mathcal{D}|}\sum_{1\le i\le|\mathcal{D}|n}Y_i\bigg)+\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}\backslash\{1,d\}}\hat{\Gamma}_n\big(d'\big)-\frac{|\mathcal{D}|-2}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n}Y_i\\&=\hat{\Gamma}_n(d)-\hat{\Gamma}_n(1).\end{aligned}$$

The first conclusion of the theorem then follows.

Next, note by the properties of the OLS estimator that

$$\begin{aligned}\hat{\delta}_{j,n}&=\bigg(\sum_{1\le i\le|\mathcal{D}|n}I\{i\in\lambda_j\}\bigg)^{-1}\sum_{1\le i\le|\mathcal{D}|n}I\{i\in\lambda_j\}\bigg(Y_i-\sum_{d\in\mathcal{D}\backslash\{1\}}\hat{\beta}_n(d)I\{D_i=d\}\bigg)\\&=\frac{1}{|\mathcal{D}|}\sum_{i\in\lambda_j}Y_i-\frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}\backslash\{1\}}\hat{\beta}_n(d).\end{aligned}$$

Therefore,

$$\hat{\epsilon}_i=\begin{cases}Y_i-\displaystyle\sum_{1\le j\le n}I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k+\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}\backslash\{1\}}\hat{\beta}_n\big(d'\big) & \text{if } D_i=1,\\ Y_i-\hat{\beta}_n(d)-\displaystyle\sum_{1\le j\le n}I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k+\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}\backslash\{1\}}\hat{\beta}_n\big(d'\big) & \text{if } D_i=d\neq 1.\end{cases}$$

Note it follows from Lemma C.4 that the heteroskedasticity-robust variance estimator of $(\hat{\beta}_n(2),\ldots,\hat{\beta}_n(|\mathcal{D}|))'$ equals

$$\bigg(\sum_{1\le i\le|\mathcal{D}|n}\tilde{C}_i\tilde{C}_i'\bigg)^{-1}\bigg(\sum_{1\le i\le|\mathcal{D}|n}\hat{\epsilon}_i^2\tilde{C}_i\tilde{C}_i'\bigg)\bigg(\sum_{1\le i\le|\mathcal{D}|n}\tilde{C}_i\tilde{C}_i'\bigg)^{-1}.$$

For $d\in\mathcal{D}\backslash\{1\}$, the corresponding $(d-1)$-th diagonal term of $\mathbb{A}=\sum_{1\le i\le|\mathcal{D}|n}\hat{\epsilon}_i^2\tilde{C}_i\tilde{C}_i'$ equals

$$\begin{aligned}\mathbb{A}_d=&\sum_{1\le i\le|\mathcal{D}|n}I\{D_i=1\}\frac{1}{|\mathcal{D}|^2}\hat{\epsilon}_i^2+\sum_{1\le i\le|\mathcal{D}|n}I\{D_i=d\}\frac{\big(|\mathcal{D}|-1\big)^2}{|\mathcal{D}|^2}\hat{\epsilon}_i^2\\&+\sum_{\tilde{d}\in\mathcal{D}\backslash\{1,d\}}\sum_{1\le i\le|\mathcal{D}|n}I\{D_i=\tilde{d}\}\frac{1}{|\mathcal{D}|^2}\hat{\epsilon}_i^2.\end{aligned}$$

For $\tilde{d} \neq \check{d} \in \mathcal{D}\backslash\{1\}$, the correponding $(\tilde{d}-1, \check{d}-1)$-th term of $\sum_{1\leq i\leq|\mathcal{D}|n} \hat{\epsilon}_i^2 \tilde{C}_i \tilde{C}_i'$ equals

$$\begin{aligned}\mathbb{A}_{\tilde{d},\check{d}} &= \sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=1\}\frac{1}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 + \sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=\tilde{d}\}\frac{-(|\mathcal{D}|-1)}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 \\ &\quad + \sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=\check{d}\}\frac{-(|\mathcal{D}|-1)}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 + \sum_{d'\in\mathcal{D}\backslash\{1,\tilde{d},\check{d}\}}\sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=d'\}\frac{1}{|\mathcal{D}|^2}\hat{\epsilon}_i^2.\end{aligned}$$

Therefore,

$$\begin{aligned}&\hat{\mathbb{V}}_n^{\mathrm{sfe}}(d,1) \\ &= \frac{4}{n^2}\mathbb{A}_d + \frac{1}{n^2}\sum_{\tilde{d}\in\mathcal{D}\backslash\{1,d\}}\mathbb{A}_{\tilde{d}} + \frac{4}{n^2}\sum_{\tilde{d}\in\mathcal{D}\backslash\{1,d\}}\mathbb{A}_{d,\tilde{d}} + \frac{2}{n^2}\sum_{\tilde{d}<\check{d}\in\mathcal{D}\backslash\{1,d\}}\mathbb{A}_{\tilde{d},\check{d}} \\ &= \frac{4+|\mathcal{D}|-2+4(|\mathcal{D}|-2)+2(|\mathcal{D}|-2)(|\mathcal{D}|-3)/2}{n^2}\sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=1\}\frac{1}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 \\ &\quad + \frac{1}{n^2}\sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=d\} \\ &\quad \times \frac{4(|\mathcal{D}|-1)^2+|\mathcal{D}|-2-4(|\mathcal{D}|-1)(|\mathcal{D}|-2)+2(|\mathcal{D}|-2)(|\mathcal{D}|-3)/2}{|\mathcal{D}|^2}\hat{\epsilon}_i^2 \\ &\quad + \frac{1}{n^2}\sum_{\tilde{d}\in\mathcal{D}\backslash\{1,d\}}\sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=\tilde{d}\} \\ &\quad \times \big((4+(|\mathcal{D}|-1)^2+|\mathcal{D}|-3-4(|\mathcal{D}|-1)+4(|\mathcal{D}|-3)-2(|\mathcal{D}|-1)(|\mathcal{D}|-3) \\ &\quad + 2(|\mathcal{D}|-3)(|\mathcal{D}|-4)/2)/|\mathcal{D}|^2\big)\hat{\epsilon}_i^2 \\ &= \frac{1}{n^2}\sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=1\}\hat{\epsilon}_i^2 + \frac{1}{n^2}\sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=d\}\hat{\epsilon}_i^2 \\ &= \frac{1}{n^2}\sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=1\}\Big(Y_i - \sum_{1\leq j\leq n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k + \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}\backslash\{1\}}\hat{\beta}_n(d')\Big)^2 \\ &\quad + \frac{1}{n^2}\sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=d\} \\ &\quad \times \Big(Y_i - \hat{\beta}_n(d) - \sum_{1\leq j\leq n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k + \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}\backslash\{1\}}\hat{\beta}_n(d')\Big)^2 \\ &= \frac{1}{n^2}\sum_{1\leq i\leq|\mathcal{D}|n} I\{D_i=1\}\Big(Y_i - \hat{\Gamma}_n(1) - \sum_{1\leq j\leq n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k + \frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat{\Gamma}_n(d')\Big)^2\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=d\}\Big(Y_i-\hat\Gamma_n(d)-\sum_{1\le j\le n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k+\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat\Gamma_n\big(d'\big)\Big)^2\\
&=\frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=1\}\Big(Y_i-\hat\Gamma_n(1)-\sum_{1\le j\le n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k+\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat\Gamma_n\big(d'\big)\Big)^2\\
&\quad+\frac{1}{n^2}\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=d\}\Big(Y_i-\hat\Gamma_n(d)-\sum_{1\le j\le n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k+\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat\Gamma_n\big(d'\big)\Big)^2\\
&=\frac{1}{n^2}\sum_{1\le j\le n}\Big(\sum_{i\in\lambda_j}\Big(I\{D_i=1\}-\frac{1}{|\mathcal{D}|}\Big)Y_i\Big)^2-\frac{1}{n}\Big(\hat\Gamma_n(1)-\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat\Gamma_n\big(d'\big)\Big)^2\\
&\quad+\frac{1}{n^2}\sum_{1\le j\le n}\Big(\sum_{i\in\lambda_j}\Big(I\{D_i=d\}-\frac{1}{|\mathcal{D}|}\Big)Y_i\Big)^2-\frac{1}{n}\Big(\hat\Gamma_n(d)-\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat\Gamma_n\big(d'\big)\Big)^2,
\end{aligned}
$$

where in the last equality we used the fact that for $d\in\mathcal{D}$,

$$
\begin{aligned}
&\sum_{1\le i\le|\mathcal{D}|n} I\{D_i=d\}\Big(Y_i-\sum_{1\le j\le n} I\{i\in\lambda_j\}\frac{1}{|\mathcal{D}|}\sum_{k\in\lambda_j}Y_k\Big)\Big(\hat\Gamma_n(d)-\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat\Gamma_n\big(d'\big)\Big)\\
&\quad=n\Big(\hat\Gamma_n(d)-\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\hat\Gamma_n\big(d'\big)\Big)^2.
\end{aligned}
$$

It follows from Assumptions 2.1 and 2.3 as well as Lemmas C.1–C.3 that as $n\to\infty$,

$$
\begin{aligned}
\hat\Gamma_n(d)&\xrightarrow{P}E\big[Y_i(d)\big]\quad\text{for all } d\in\mathcal{D},\\
\frac{1}{n}\sum_{1\le j\le n}\sum_{i\in\lambda_j} I\{D_i=d\}Y_i^2&\xrightarrow{P}E\big[Y_i^2(d)\big],\\
\frac{1}{n}\sum_{1\le j\le n}\Big(\sum_{i\in\lambda_j} I\{D_i=d\}Y_i\Big)\Big(\sum_{i\in\lambda_j} I\big\{D_i=d'\big\}Y_i\Big)&\xrightarrow{P}E\big[\Gamma_d(X_i)\Gamma_{d'}(X_i)\big]\quad\text{for all } d\neq d'\in\mathcal{D}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
n\hat{\mathbb{V}}_n^{\mathrm{sfe}}(d,1)\xrightarrow{P}&\operatorname{Var}\Big[\Gamma_1(X_i)-\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\Gamma_{d'}(X_i)\Big]\\
&+\Big(1-\frac{1}{|\mathcal{D}|}\Big)^2E\big[\operatorname{Var}\big[Y_i(1)|X_i\big]\big]+\frac{1}{|\mathcal{D}|^2}\sum_{d'\in\mathcal{D}\setminus\{1\}}E\big[\operatorname{Var}\big[Y_i(d')|X_i\big]\big]\\
&+\operatorname{Var}\Big[\Gamma_d(X_i)-\frac{1}{|\mathcal{D}|}\sum_{d'\in\mathcal{D}}\Gamma_{d'}(X_i)\Big]\\
&+\Big(1-\frac{1}{|\mathcal{D}|}\Big)^2E\big[\operatorname{Var}\big[Y_i(d)|X_i\big]\big]+\frac{1}{|\mathcal{D}|^2}\sum_{d'\in\mathcal{D}\setminus\{d\}}E\big[\operatorname{Var}\big[Y_i(d')|X_i\big]\big].
\end{aligned}
$$

Finally, note by Theorem 3.1 that the actual limiting variance for $\hat{\Gamma}_n(d) - \hat{\Gamma}_n(1)$ is

$$E\big[\mathrm{Var}\big[Y_i(d)|X_i\big]\big] + E\big[\mathrm{Var}\big[Y_i(1)|X_i\big]\big] + \frac{1}{|\mathcal{D}|}E\big[\big((\Gamma_d(X_i) - \Gamma_d) - (\Gamma_1(X_i) - \Gamma_1)\big)^2\big].$$

Consider the special case where $E[\mathrm{Var}[Y_i(d')|X_i]]$ are identical across $d' \in \mathcal{D}$ and

$$\Gamma_1(X_i) = \Gamma_d(X_i) = \frac{1}{|\mathcal{D}|}\sum_{d' \in \mathcal{D}} \Gamma_{d'}(X_i) \quad \text{with probability one.}$$

Then the probability limit of $n\hat{\mathbb{V}}_n^{\mathrm{sfe}}(d, 1)$ is clearly strictly smaller than the actual limiting variance for $\hat{\Gamma}_n(d) - \hat{\Gamma}_n(1)$.

For variance estimator HC 1, consider the special case where $\Gamma_d(X_i)$ are identical across $d \in \mathcal{D}$, $E[\mathrm{Var}[Y_i(d)|X_i]] > 0$, $E[\mathrm{Var}[Y_i(1)|X_i]] > 0$, and $E[\mathrm{Var}[Y_i(d')|X_i]]$ is zero for all $d' \in \mathcal{D}\backslash\{1, d\}$. Then

$$\begin{aligned}
&n\hat{\mathbb{V}}_n^{\mathrm{sfe}}(d, 1) \times \frac{|\mathcal{D}|n}{|\mathcal{D}|n - \big(|\mathcal{D}| - 1 + n\big)} \\
&\quad \xrightarrow{P} \frac{|\mathcal{D}|}{|\mathcal{D}| - 1}\left(\left(1 - \frac{1}{|\mathcal{D}|}\right)^2 + \frac{1}{|\mathcal{D}|^2}\right)\big(E\big[\mathrm{Var}\big[Y_i(d)|X_i\big]\big] + E\big[\mathrm{Var}\big[Y_i(1)|X_i\big]\big]\big) \\
&\quad = \frac{|\mathcal{D}|^2 - 2|\mathcal{D}| + 2}{|\mathcal{D}|^2 - |\mathcal{D}|}\big(E\big[\mathrm{Var}\big[Y_i(d)|X_i\big]\big] + E\big[\mathrm{Var}\big[Y_i(1)|X_i\big]\big]\big).
\end{aligned}$$

Note that

$$\frac{|\mathcal{D}|^2 - 2|\mathcal{D}| + 2}{|\mathcal{D}|^2 - |\mathcal{D}|} < 1$$

if and only if $|\mathcal{D}| > 2$. By a continuity argument, the result then follows for the case where $E[\mathrm{Var}[Y_i(d')|X_i]]$ is sufficiently close to zero for all $d' \in \mathcal{D}\backslash\{1, d\}$. □

B.4 *Proof of Theorem 3.4*

First, note that

$$\left(\frac{1}{n}\sum_{1 \le j \le n}\sum_{i \in \lambda_j} C_i C_i'\right)^{-1} = \begin{pmatrix} |\mathcal{D}| & 1 & 1 & \dots & 1 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -1 & -1 & \dots & -1 \\ -1 & 2 & 1 & \dots & 1 \\ -1 & 1 & 2 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 1 & 1 & \dots & 2 \end{pmatrix},$$

and note that

$$\sum_{i\in\lambda_j}\hat{\epsilon}_i C_i = \begin{pmatrix} \sum_{i\in\lambda_j}\sum_{d\in\mathcal{D}\setminus\{1\}}\big(Y_i-\hat{\gamma}_n(d)\big)I\{D_i=d\}+Y_iI\{D_i=1\}-\hat{\gamma}_n(1) \\ \sum_{i\in\lambda_j}\big(Y_i-\hat{\gamma}_n(2)-\hat{\gamma}_n(1)\big)I\{D_i=2\} \\ \sum_{i\in\lambda_j}\big(Y_i-\hat{\gamma}_n(3)-\hat{\gamma}_n(1)\big)I\{D_i=3\} \\ \vdots \\ \sum_{i\in\lambda_j}\big(Y_i-\hat{\gamma}_n\big(|\mathcal{D}|\big)-\hat{\gamma}_n(1)\big)I\big\{D_i=|\mathcal{D}|\big\} \end{pmatrix}.$$

Combining these expressions, it follows that the dth diagonal element of $n\cdot\hat{\mathbb{V}}_n^{\mathrm{bcve}}$ is equal to

$$\begin{aligned} n\cdot\hat{\mathbb{V}}_n^{\mathrm{bcve}}(d) &= \frac{1}{n}\sum_{1\le j\le n}\left(\sum_{i\in\lambda_j}\big(Y_i-\hat{\gamma}_n(d)-\hat{\gamma}_n(1)\big)I\{D_i=d\}-\sum_{i\in\lambda_j}\big(Y_i-\hat{\gamma}_n(1)\big)I\{D_i=1\}\right)^2 \\ &= \frac{1}{n}\sum_{1\le j\le n}\left(\sum_{i\in\lambda_j}Y_iI\{D_i=d\}-\sum_{i\in\lambda_j}Y_iI\{D_i=1\}\right)^2-\big(\hat{\Gamma}_n(d)-\hat{\Gamma}_n(1)\big)^2. \end{aligned}$$

Where the second equality exploits the fact that $\hat{\gamma}_n(d)=\hat{\Gamma}_n(d)-\hat{\Gamma}_n(1)$ for $d\in\mathcal{D}\setminus\{1\}$ and $\hat{\gamma}_n(1)=\hat{\Gamma}_n(1)$. It thus follows from Lemmas C.1–C.2 and the continuous mapping theorem that

$$n\cdot\hat{\mathbb{V}}_n^{\mathrm{bcve}}(d)\xrightarrow{p}E\big[\mathrm{Var}\big[Y_i(d)|X_i\big]\big]+E\big[\mathrm{Var}\big[Y_i(1)|X_i\big]\big]+E\big[\big((\Gamma_d(X_i)-\Gamma_d)-(\Gamma_1(X_i)-\Gamma_1)\big)^2\big].$$

Next, note that by Theorem 3.1, the actual limiting variance of $\hat{\Gamma}_n(d)-\hat{\Gamma}_n(1)$ is given by

$$E\big[\mathrm{Var}\big[Y_i(d)|X_i\big]\big]+E\big[\mathrm{Var}\big[Y_i(1)|X_i\big]\big]+\frac{1}{|\mathcal{D}|}E\big[\big((\Gamma_d(X_i)-\Gamma_d)-(\Gamma_1(X_i)-\Gamma_1)\big)^2\big].$$

Therefore, the test defined in (8) is conservative unless

$$E\big[\big((\Gamma_d(X_i)-\Gamma_d)-(\Gamma_1(X_i)-\Gamma_1)\big)^2\big]=0,$$

as desired. □

B.5 *Proof of Theorem 3.5*

The proof is similar to the proof of Theorem 3.1, with the difference being that two units are assigned to each treatment status in each block. The necessary modification follows from arguing similarly as in Lemma B.3 of Bai (2022) and is omitted. □

B.6 *Proof of Theorem 3.6*

First, note

$$\begin{aligned}
&E\big[\tilde{\rho}_n(d,d)|X^{(n)}\big] \\
&\quad = \frac{2}{n}\sum_{1\le j\le \lfloor n/2\rfloor}\frac{1}{\binom{2|\mathcal{D}|}{2}}\sum_{i<l,i,l\in\lambda_j}E\big[Y_i(d)Y_l(d)|X^{(n)}\big] \\
&\quad = \frac{2}{n}\sum_{1\le j\le \lfloor n/2\rfloor}\frac{1}{\binom{2|\mathcal{D}|}{2}}\sum_{i<l,i,l\in\lambda_j}E\big[Y_i(d)|X_i\big]E\big[Y_l(d)|X_l\big],
\end{aligned}$$

where the first equality follows from the conditional independence assumption in Assumption 2.2 and the fact that in each block, there are $\binom{2|\mathcal{D}|}{2}$ ways to choose 2 units out of $2|\mathcal{D}|$ units and assign them to treatment arm d, and the second equality follows from the fact that conditional on $X^{(n)}$, $Y^{(n)}(d)$ are i.i.d. across units. Equation (11) then follows by arguing similarly as in the proof of Lemma C.3 below (see also Section 4.7 of Bai (2022)). □

B.7 *Proof of Theorem 3.7*

First, we show that

$$\sqrt{n}\big(\hat{\Delta}_{\nu,n}-\Delta_\nu(Q)\big)\xrightarrow{d}N\big(0,\sigma^2_{h,\nu}\big),$$

under the stratified factorial design defined by $h(\cdot)$. To show this, we derive the limiting distribution of $\sqrt{n}(\hat{\Gamma}_n(d)-\Gamma_d(Q):d\in\mathcal{D})$. To that end, note that

$$\sqrt{n}\big(\hat{\Gamma}_n(d)-\Gamma_d(Q):d\in\mathcal{D}\big)'=A_n+B_n+C_n+o_P(1),$$

where $A_n=(A_{n,d}:d\in\mathcal{D})'$, $B_n=(B_{n,d}:d\in\mathcal{D})'$, $C_n=(C_{n,d}:d\in\mathcal{D})'$, with

$$\begin{aligned}
A_{n,d}&=\sqrt{|\mathcal{D}|}\frac{1}{\sqrt{J_n}}\sum_{1\le i\le J_n}\big(Y_i(d)-E\big[Y_i(d)|h(X_i)\big]\big)I\{D_i=d\},\\
B_{n,d}&=\sqrt{|\mathcal{D}|}\frac{1}{\sqrt{J_n}}\sum_{1\le i\le J_n}\big(I\{D_i=d\}-\pi\big)\big(E\big[Y_i(d)|h(X_i)\big]-E\big[Y_i(d)\big]\big),\\
C_{n,d}&=\sqrt{|\mathcal{D}|}\frac{1}{\sqrt{J_n}}\sum_{1\le i\le J_n}\pi\big(E\big[Y_i(d)|h(X_i)\big]-E\big[Y_i(d)\big]\big),
\end{aligned}$$

where $\pi:=\frac{1}{|\mathcal{D}|}$. Rewriting each of these terms using the fact that

$$\begin{aligned}
E\big[Y_i(d)|h(X_i)\big]&=\sum_{1\le s\le S}E\big[Y_i(d)|h(X_i)\big]I\big\{h(X_i)=s\big\}\\
&=\sum_{1\le s\le S}E\big[Y_i(d)|h(X_i)=s\big]I\big\{h(X_i)=s\big\},
\end{aligned}$$

we obtain

$$A_{n,d} = \sqrt{|\mathcal{D}|} \sum_{1\le s\le S} \frac{1}{\sqrt{J_n}} \sum_{1\le i\le J_n} \big(E\big[Y_i(d)|h(X_i)\big] - E\big[Y_i(d)\big]\big) I\big\{D_i = d, h(X_i) = s\big\},$$

$$B_{n,d} = \sqrt{|\mathcal{D}|} \sum_{1\le s\le S} \big(E\big[Y_i(d)|h(X_i) = s\big] - E\big[Y_i(d)\big]\big) \frac{J_n(s)}{J_n} \sqrt{J_n}\left(\frac{J_{n,d}(s)}{J_n(s)} - \pi\right),$$

$$C_{n,d} = \sqrt{|\mathcal{D}|} \sum_{1\le s\le S} \pi\big(E\big[Y_i(d)|h(X_i) = s\big] - E\big[Y_i(d)\big]\big) \sqrt{J_n}\left(\frac{J_n(s)}{J_n} - p(s)\right),$$

where $J_n(s) = \sum_{1\le i\le J_n} I\{h(X_i) = s\}$, $J_{n,d}(s) = \sum_{1\le i\le J_n} I\{h(X_i) = s, D_i = d\}$, $p(s) = P(h(X_i) = s)$, and importantly for $C_{n,d}$ we have used the fact that

$$\sum_{1\le s\le S} \big(E\big[Y_i(d)|h(X_i) = s\big] - E\big[Y_i(d)\big]\big) p(s) = 0,$$

which follows by the law of iterated expectations. By the law of large numbers, $J_n(s)/J_n \xrightarrow{p} p(s)$, and by the properties of stratified block randomization (see Example 3.4 in Bugni, Canay, and Shaikh (2018)),

$$\sqrt{J_n}\left(\frac{J_{n,d}(s)}{J_n(s)} - \pi\right) \xrightarrow{p} 0,$$

and hence we can conclude that $B_{n,d} \xrightarrow{p} 0$ for every $d \in \mathcal{D}$. Using Lemma C.1. in Bugni, Canay, and Shaikh (2019), it can then be shown that

$$\begin{pmatrix} A_n \\ C_n \end{pmatrix} \xrightarrow{d} N\left(0, \begin{bmatrix} \mathbb{V}_{h,1} & 0 \\ 0 & \mathbb{V}_{h,2} \end{bmatrix}\right),$$

and hence the first result follows. Next, let ν be a $1 \times |\mathcal{D}|$ vector of constants. Then it can be shown that

$$\nu \mathbb{V} \nu' = \sum_{d\in\mathcal{D}} \nu_d^2 \mathrm{Var}\big[Y_i(d)\big] - \sum_{d\ne d'\in\mathcal{D}} \frac{1}{|\mathcal{D}|} \mathrm{Var}\big[\nu_d E\big[Y_i(d)|X_i\big] - \nu_{d'} E\big[Y_i\big(d'\big)|X_i\big]\big],$$

and

$$\nu \mathbb{V}_h \nu' = \sum_{d\in\mathcal{D}} \nu_d^2 \mathrm{Var}\big[Y_i(d)\big] - \sum_{d\ne d'\in\mathcal{D}} \frac{1}{|\mathcal{D}|} \mathrm{Var}\big[\nu_d E\big[Y_i(d)|h(X_i)\big] - \nu_{d'} E\big[Y_i\big(d'\big)|h(X_i)\big]\big].$$

It then follows from similar arguments to those used in the proof of Theorem C.2 of Bai (2022) that $\nu \mathbb{V} \nu' \le \nu \mathbb{V}_h \nu'$. In particular, note that

$$\begin{aligned}
&\mathrm{Var}\big[\nu_d E\big[Y_i(d)|X_i\big] - \nu_{d'} E\big[Y_i\big(d'\big)|X_i\big]\big] \\
&\quad = E\big[\big(\nu_d E\big[Y_i(d)|X_i\big] - \nu_{d'} E\big[Y_i\big(d'\big)|X_i\big] - \big(\nu_d E\big[Y_i(d)\big] - \nu_{d'} E\big[Y_i\big(d'\big)\big]\big)\big)^2\big] \\
&\quad = E\big[\big(\nu_d E\big[Y_i(d)|X_i\big] - \nu_{d'} E\big[Y_i\big(d'\big)|X_i\big] - \big(\nu_d E\big[Y_i(d)|h(X_i)\big] - \nu_{d'} E\big[Y_i\big(d'\big)|h(X_i)\big]\big)
\end{aligned}$$

$$
\begin{aligned}
&+\big(\nu_d E\big[Y_i(d)|h(X_i)\big]-\nu_{d'}E\big[Y_i\big(d'\big)|h(X_i)\big]\big)-\big(\nu_d E\big[Y_i(d)\big]-\nu_{d'}E\big[Y_i\big(d'\big)\big]\big)\big)^2\big]\\
&=E\big[\big(\nu_d E\big[Y_i(d)|X_i\big]-\nu_{d'}E\big[Y_i\big(d'\big)|X_i\big]-\big(\nu_d E\big[Y_i(d)|h(X_i)\big]-\nu_{d'}E\big[Y_i\big(d'\big)|h(X_i)\big]\big)\big)^2\big]\\
&\quad+E\big[\big(\nu_d E\big[Y_i(d)|h(X_i)\big]-\nu_{d'}E\big[Y_i\big(d'\big)|h(X_i)\big]\big)-\big(\nu_d E\big[Y_i(d)\big]-\nu_{d'}E\big[Y_i\big(d'\big)\big]\big)\big)^2\big],
\end{aligned}
$$

where the last equality follows because

$$
\begin{aligned}
&E\big[\big(\nu_d E\big[Y_i(d)|X_i\big]-\nu_{d'}E\big[Y_i\big(d'\big)|X_i\big]-\big(\nu_d E\big[Y_i(d)|h(X_i)\big]-\nu_{d'}E\big[Y_i\big(d'\big)|h(X_i)\big]\big)\big)\\
&\quad\times\big(\big(\nu_d E\big[Y_i(d)|h(X_i)\big]-\nu_{d'}E\big[Y_i\big(d'\big)|h(X_i)\big]\big)-\big(\nu_d E\big[Y_i(d)\big]-\nu_{d'}E\big[Y_i\big(d'\big)\big]\big)\big)\big]\\
&=E\big[E\big[\big(\nu_d E\big[Y_i(d)|X_i\big]-\nu_{d'}E\big[Y_i\big(d'\big)|X_i\big]-\big(\nu_d E\big[Y_i(d)|h(X_i)\big]-\nu_{d'}E\big[Y_i\big(d'\big)|h(X_i)\big]\big)\big)\\
&\quad\times\big(\big(\nu_d E\big[Y_i(d)|h(X_i)\big]-\nu_{d'}E\big[Y_i\big(d'\big)|h(X_i)\big]\big)-\big(\nu_d E\big[Y_i(d)\big]-\nu_{d'}E\big[Y_i\big(d'\big)\big]\big)\big)|h(X_i)\big]\big]\\
&=E\big[E\big[\big(\nu_d E\big[Y_i(d)|X_i\big]-\nu_{d'}E\big[Y_i\big(d'\big)|X_i\big]\\
&\quad-\big(\nu_d E\big[Y_i(d)|h(X_i)\big]-\nu_{d'}E\big[Y_i\big(d'\big)|h(X_i)\big]\big)\big)|h(X_i)\big]\\
&\quad\times\big(\big(\nu_d E\big[Y_i(d)|h(X_i)\big]-\nu_{d'}E\big[Y_i\big(d'\big)|h(X_i)\big]\big)-\big(\nu_d E\big[Y_i(d)\big]-\nu_{d'}E\big[Y_i\big(d'\big)\big]\big)\big)\big]\\
&=0,
\end{aligned}
$$

where the last equality follows from the law of iterated expectations. We can thus conclude that the matched tuples design is asymptotically more efficient than the large stratum design, in the sense that the difference in variances between the large stratum and matched tuples designs, $\mathbb{V}_h-\mathbb{V}$, is positive semidefinite. □

B.8 *Proof of Theorem* 3.8

To begin, note that

$$
\begin{aligned}
\hat{\Delta}_{\nu_k,n}=&\frac{1}{n}\sum_{1\le i\le J_n}\sum_{d\in\mathcal{D}}I\big\{\iota_k(d)=+1\big\}I\{D_i=d\}Y_i(d)\\
&-\frac{1}{n}\sum_{1\le i\le J_n}\sum_{d\in\mathcal{D}}I\big\{\iota_k(d)=-1\big\}I\{D_i=d\}Y_i(d).
\end{aligned}
$$

Let A_i, $1\le i\le J_n$ denote a sequence of i.i.d. random vectors, each of which is a $K-1$ vector of i.i.d. Rademacher random variables. Further, assume they are independent of $Y^{(n)}(d)$, $d\in\mathcal{D}$, $D^{(n)}$, and $X^{(n)}$. Define $\iota_{-k}(d)$ as the vector of all entries of $\iota(d)$ except the kth entry. Then we consider the following "averaged" potential outcomes over these $K-1$ factors defined as follows:

$$
\begin{aligned}
\tilde{Y}_i(+1)&:=\sum_{d\in\mathcal{D}}I\big\{\iota_k(d)=+1\big\}I\big\{\iota_{-k}(d)=A_i\big\}Y_i(d),\\
\tilde{Y}_i(-1)&:=\sum_{d\in\mathcal{D}}I\big\{\iota_k(d)=-1\big\}I\big\{\iota_{-k}(d)=A_i\big\}Y_i(d).
\end{aligned}
$$

With this notation, define

$$\tilde{\Delta}_{\nu_k,n} = \frac{1}{n}\sum_{1\le i\le J_n} I\{\iota_k(D_i) = +1\}\tilde{Y}_i(+1) - \frac{1}{n}\sum_{1\le i\le J_n} I\{\iota_k(D_i) = -1\}\tilde{Y}_i(-1).$$

It then follows from the definition of the factor k-specific design that $\tilde{\Delta}_{\nu_k,n}$ has the same distribution as $\hat{\Delta}_{\nu_k,n}$. To see it, note

$$\begin{aligned}&\frac{1}{n}\sum_{1\le i\le J_n}\sum_{d\in\mathcal{D}} I\{\iota_k(d) = +1\}I\{D_i = d\}Y_i(d)\\ &\quad= \frac{1}{n}\sum_{1\le i\le J_n}\sum_{d\in\mathcal{D}} I\{\iota_k(D_i) = +1\}I\{\iota_k(d) = +1\}I\{\iota_{-k}(D_i) = \iota_{-k}(d)\}Y_i(d)\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n}\sum_{1\le i\le J_n} I\{\iota_k(D_i) = +1\}\tilde{Y}_i(+1)\\ &\quad= \frac{1}{n}\sum_{1\le i\le J_n}\sum_{d\in\mathcal{D}} I\{\iota_k(D_i) = +1\}I\{\iota_k(d) = +1\}I\{\iota_{-k}(d) = A_i\}Y_i(d)\end{aligned}$$

and $\iota_{-k}(D_i)$ and A_i follow the same distribution independently of everything else.

Note $\tilde{\Delta}_{\nu_k,n}/2^{K-1}$ can be thought of as the difference-in-means estimator where the treatment has two levels $+1$ and -1 and the potential outcomes are $\tilde{Y}_i(+1)$ and $\tilde{Y}_i(-1)$. The conditions in Lemma S.1.4 in Bai, Romano, and Shaikh (2021) can be verified straightforwardly and, therefore, we have

$$\sqrt{2^{K-1}n}\left(\frac{\hat{\Delta}_{\nu_k,n}}{2^{K-1}} - \frac{\Delta_{\nu_k}(Q)}{2^{K-1}}\right) \xrightarrow{d} N(0, \mathbb{V}_{\nu_k,mp}),$$

where

$$\begin{aligned}\mathbb{V}_{\nu_k,mp} :=\ & E\big[\mathrm{Var}\big[\tilde{Y}_i(+1)|X_i\big]\big] + E\big[\mathrm{Var}\big[\tilde{Y}_i(-1)|X_i\big]\big]\\ &+ \frac{1}{2}E\big[\big(E\big[\tilde{Y}_i(+1)|X_i\big] - E\big[\tilde{Y}_i(+1)\big] - \big(E\big[\tilde{Y}_i(-1)|X_i\big] - E\big[\tilde{Y}_i(-1)\big]\big)\big)^2\big].\end{aligned}$$

Note that by Assumption 2.2,

$$\begin{aligned}E\big[\tilde{Y}_i(+1)|X_i\big] &= E\Big[\sum_{d\in\mathcal{D}} I\{\iota_k(d) = +1\}I\{\iota_{-k}(d) = A_i\}Y_i(d)\Big|X_i\Big]\\ &= \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}} I\{\iota_k(d) = +1\}\Gamma_d(X_i).\end{aligned}$$

Therefore,

$$\frac{1}{2}E\big[\big(E\big[\tilde{Y}_i(+1)|X_i\big] - E\big[\tilde{Y}_i(+1)\big] - \big(E\big[\tilde{Y}_i(-1)|X_i\big] - E\big[\tilde{Y}_i(0)\big]\big)\big)^2\big]$$

$$
\begin{aligned}
&= \frac{1}{2}\cdot\frac{1}{2^{2(K-1)}} \\
&\quad\times E\Bigg[\bigg(\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}\big(\Gamma_d(X_i)-\Gamma_d\big)-\sum_{d\in\mathcal{D}} I\{\iota_k(d)=-1\}\big(\Gamma_d(X_i)-\Gamma_d\big)\bigg)^2\Bigg] \\
&= \frac{1}{2}\cdot\frac{1}{2^{2(K-1)}} E\big[\big(\nu_k'\big(\Gamma_d(X_i)-\Gamma_d : d\in\mathcal{D}\big)\big)^2\big] \\
&= \frac{2^{K-1}}{2^{2(K-1)}}\nu_k' E\Bigg[\frac{1}{2^K}\operatorname{Cov}\big[\Gamma_d(X_i),\Gamma_{d'}(X_i)\big]\Bigg]_{d,d'\in\mathcal{D}}\nu_k \\
&= \frac{1}{2^{(K-1)}}\nu_k'\mathbb{V}_2\nu_k.
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
&\operatorname{Var}\big[\tilde{Y}_i(+1)|X_i\big] \\
&= \operatorname{Var}\Bigg[\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{\iota_{-k}(d)=A_i\}Y_i(d)\Big|X_i\Bigg] \\
&= E\Bigg[\operatorname{Var}\Bigg[\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{\iota_{-k}(d)=A_i\}Y_i(d)\Big|X_i,A_i\Bigg]\Big|X_i\Bigg] \\
&\quad+\operatorname{Var}\Bigg[E\Bigg[\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{\iota_{-k}(d)=A_i\}Y_i(d)\Big|X_i,A_i\Bigg]\Big|X_i\Bigg] \\
&= E\Bigg[\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{\iota_{-k}(d)=A_i\}\operatorname{Var}\big[Y_i(d)|X_i\big]\Big|X_i\Bigg] \\
&\quad+\operatorname{Var}\Bigg[\sum_{d\in\mathcal{D}} I\{\iota_k(d)=+1\}I\{\iota_{-k}(d)=A_i\}\Gamma_d(X_i)\Big|X_i\Bigg] \\
&= \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}:\iota_k(d)=+1}\operatorname{Var}\big[Y_i(d)|X_i\big] \\
&\quad+\frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}:\iota_k(d)=+1}\Bigg(\Gamma_d(X_i)-\frac{1}{2^{K-1}}\sum_{d'\in\mathcal{D}:\iota_k(d')=+1}\Gamma_{d'}(X_i)\Bigg)^2 \\
&= \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}:\iota_k(d)=+1}\Bigg(\operatorname{Var}\big[Y_i(d)|X_i\big]+\Bigg(\Gamma_d(X_i)-\frac{1}{2^{K-1}}\sum_{d'\in\mathcal{D}:\iota_k(d')=+1}\Gamma_{d'}(X_i)\Bigg)^2\Bigg).
\end{aligned}
$$

A similar calculation applies to $\operatorname{Var}[\tilde{Y}_i(-1)|X_i]$. Finally,

$$
\begin{aligned}
\mathbb{V}_{\nu_k,mp} &= \frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}} E\big[\operatorname{Var}\big[Y_i(d)|X_i\big]\big]+\frac{1}{2^{K-1}}\nu_k'\mathbb{V}_2\nu_k \\
&\quad+\frac{1}{2^{K-1}}E\Bigg[\sum_{d\in\mathcal{D}:\iota_k(d)=+1}\Bigg(\Gamma_d(X_i)-\frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}:\iota_k(d)=+1}\Gamma_d(X_i)\Bigg)^2\Bigg]
\end{aligned}
$$

$$+\frac{1}{2^{K-1}}E\Bigg[\sum_{d\in\mathcal{D}:\iota_k(d')=-1}\Bigg(\Gamma_{d'}(X_i)-\frac{1}{2^{K-1}}\sum_{d\in\mathcal{D}::\iota_k(d)=-1}\Gamma_d(X_i)\Bigg)^2\Bigg].$$

The conclusion therefore follows. □

Appendix C: Auxiliary lemmas

Lemma C.1. *Suppose Assumptions 2.1–2.3 hold. Then, for $r=1,2$,*

$$\frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n}Y_i^r(d)I\{D_i=d\}\xrightarrow{P}E\big[Y_i^r(d)\big].$$

Proof of Lemma C.1. We prove the conclusion for $r=1$ only and the proof for $r=2$ follows similarly. To this end, write

$$\begin{aligned}\frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n}Y_i(d)I\{D_i=d\}&=\frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n}\big(Y_i(d)I\{D_i=d\}-E\big[Y_i(d)I\{D_i=d\}|X^{(n)},D^{(n)}\big]\big)\\&\quad+\frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n}E\big[Y_i(d)I\{D_i=d\}|X^{(n)},D^{(n)}\big].\end{aligned}$$

Note

$$\begin{aligned}&\frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n}E\big[Y_i(d)I\{D_i=d\}|X^{(n)},D^{(n)}\big]\\&=\frac{1}{n}\sum_{1\le i\le|\mathcal{D}|n}I\{D_i=d\}E\big[Y_i(d)|X_i\big]\xrightarrow{P}E\big[E\big[Y_i(d)|X_i\big]\big]=E\big[Y_i(d)\big],\end{aligned}$$

where the equality follows from Assumption 2.2 and the convergence in probability follows from Assumption 2.3 and similar arguments to those used in the proof of Theorem 3.1. To complete the proof, we argue

$$\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n}\big(Y_i(d)I\{D_i=d\}-E\big[Y_i(d)I\{D_i=d\}|X^{(n)},D^{(n)}\big]\big)\xrightarrow{P}0.$$

For this purpose, we proceed by verifying the uniform integrability condition in Lemma S.1.3 of Bai, Romano, and Shaikh (2021) conditional on $X^{(n)}$ and $D^{(n)}$. Note for any $m>0$ that

$$\begin{aligned}&\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n}E\big[|Y_i(d)I\{D_i=d\}-E\big[Y_i(d)I\{D_i=d\}|X^{(n)},D^{(n)}\big]\big)\\&\quad\times I\big\{\big|Y_i(d)I\{D_i=d\}-E\big[Y_i(d)I\{D_i=d\}|X^{(n)},D^{(n)}\big]\big|>m\big\}|X^{(n)},D^{(n)}\big]\\&=\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n}E\big[\big|Y_i(d)I\{D_i=d\}-E\big[Y_i(d)|X_i\big]I\{D_i=d\}\big|\end{aligned}$$

$$
\begin{aligned}
&\times I\{|Y_i(d)I\{D_i=d\} - E[Y_i(d)|X_i]I\{D_i=d\}| > m\}|X^{(n)}, D^{(n)}] \\
&\leq \frac{1}{|\mathcal{D}|n} \sum_{1\leq i\leq |\mathcal{D}|n} E[|Y_i(d) - E[Y_i(d)|X_i]|I\{|Y_i(d) - E[Y_i(d)|X_i]| > m\}|X^{(n)}, D^{(n)}] \\
&= \frac{1}{|\mathcal{D}|n} \sum_{1\leq i\leq |\mathcal{D}|n} E[|Y_i(d) - E[Y_i(d)|X_i]|I\{|Y_i(d) - E[Y_i(d)|X_i]| > m\}|X_i] \\
&\xrightarrow{P} E[|Y_i(d) - E[Y_i(d)|X_i]|I\{|Y_i(d) - E[Y_i(d)|X_i]| > m\}],
\end{aligned}
$$

where the first equality holds because of Assumption 2.2, the inequality holds because $0 \leq I\{D_i = d\} \leq 1$, the second equality holds because of Assumption 2.2 again, and the convergence in probability follows from the weak law of large numbers because

$$
\begin{aligned}
&E[|Y_i(d) - E[Y_i(d)|X_i]|I\{|Y_i(d) - E[Y_i(d)|X_i]| > m\}] \\
&\quad \leq E[|Y_i(d) - E[Y_i(d)|X_i]|] \\
&\quad \leq E[|Y_i(d)|] + E[|E[Y_i(d)|X_i]|] \leq E[|Y_i(d)|] + E[E[|Y_i(d)||X_i]] = 2E[|Y_i(d)|].
\end{aligned}
$$

The proof could then be completed using the subsequencing argument as in (S.29) of the proof of Lemma S.1.5 of Bai, Romano, and Shaikh (2021). □

Lemma C.2. *Suppose Assumptions 2.1–2.3 hold. Then* $\hat{\rho}_n(d, d') \xrightarrow{P} E[\Gamma_d(X_i)\Gamma_{d'}(X_i)]$ *as* $n \to \infty$.

Proof of Lemma C.2. To begin with, note

$$
\begin{aligned}
&E[\hat{\rho}_n(d, d')|X^{(n)}] \\
&\quad = \frac{1}{n} \sum_{1\leq j\leq n} \frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)} \sum_{\{i,k\}\subset\lambda_j} \big(\Gamma_d(X_i)\Gamma_{d'}(X_k) + \Gamma_d(X_k)\Gamma_{d'}(X_i)\big) \\
&\quad = \frac{1}{n} \sum_{1\leq j\leq n} \frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)} \sum_{\{i,k\}\subset\lambda_j} \big(\Gamma_d(X_i)\Gamma_{d'}(X_i) + \Gamma_d(X_k)\Gamma_{d'}(X_k) \\
&\qquad - \big(\Gamma_d(X_i) - \Gamma_d(X_k)\big)\big(\Gamma_{d'}(X_i) - \Gamma_{d'}(X_k)\big)\big) \\
&\quad = \frac{1}{|\mathcal{D}|n} \sum_{1\leq i\leq |\mathcal{D}|n} \Gamma_d(X_i)\Gamma_{d'}(X_i) \\
&\qquad - \frac{1}{n} \sum_{1\leq j\leq n} \frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)} \sum_{\{i,k\}\subset\lambda_j} \big(\Gamma_d(X_i) - \Gamma_d(X_k)\big)\big(\Gamma_{d'}(X_i) - \Gamma_{d'}(X_k)\big) \\
&\quad \xrightarrow{P} E[\Gamma_d(X_i)\Gamma_{d'}(X_i)],
\end{aligned}
$$

where the convergence in probability follows from Assumptions 2.1(c) and 2.3. To conclude the proof, we show

$$
\hat{\rho}_n(d, d') - E[\hat{\rho}_n(d, d')|X^{(n)}] \xrightarrow{P} 0. \tag{14}
$$

In order for this, we proceed to verify the uniform integrability condition in Lemma S.1.3 of Bai, Romano, and Shaikh (2021) conditional on $X^{(n)}$. Define

$$\hat{\rho}_{n,j}(d,d') = \Bigg(\sum_{i\in\lambda_j} Y_i I\{D_i = d\}\Bigg)\Bigg(\sum_{i\in\lambda_j} Y_i I\{D_i = d'\}\Bigg).$$

In what follows, we repeatedly use the following inequalities:

$$I\Bigg\{\Bigg|\sum_{1\le j\le k} a_j\Bigg| > \lambda\Bigg\} \le \sum_{1\le j\le k} I\Bigg\{|a_j| > \frac{\lambda}{k}\Bigg\},$$

$$\Bigg|\sum_{1\le j\le k} a_j\Bigg| I\Bigg\{\Bigg|\sum_{1\le j\le k} a_j\Bigg| > \lambda\Bigg\} \le \sum_{1\le j\le k} k|a_j| I\Bigg\{|a_j| > \frac{\lambda}{k}\Bigg\},$$

$$|ab| I\{|ab| > \lambda\} \le a^2 I\{|a| > \sqrt{\lambda}\} + b^2 I\{|b| > \sqrt{\lambda}\}.$$

We will also repeatedly use the facts that $0 \le I\{D_i = d\} \le 1$ and $I\{D_i = d\}I\{D_k = d\} = 0$ for $i \ne k$ in the same stratum. Note

$$\begin{aligned}
&E\big[|\hat{\rho}_{n,j}(d,d') - E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]|I\{|\hat{\rho}_{n,j}(d,d') - E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]| > \lambda\}|X^{(n)}\big] \\
&\le E\Bigg[|\hat{\rho}_{n,j}(d,d')|I\Bigg\{|\hat{\rho}_{n,j}(d,d')| > \frac{\lambda}{2}\Bigg\}\Bigg|X^{(n)}\Bigg] \\
&\quad + E\Bigg[|E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]|I\Bigg\{|E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]| > \frac{\lambda}{2}\Bigg\}\Bigg|X^{(n)}\Bigg] \\
&= E\Bigg[|\hat{\rho}_{n,j}(d,d')|I\Bigg\{|\hat{\rho}_{n,j}(d,d')| > \frac{\lambda}{2}\Bigg\}\Bigg|X^{(n)}\Bigg] \\
&\quad + |E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]|I\Bigg\{|E[\hat{\rho}_{n,j}(d,d')|X^{(n)}]| > \frac{\lambda}{2}\Bigg\} \\
&\le E\Bigg[\Bigg|\sum_{i\in\lambda_j} Y_i(d) I\{D_i = d\} \sum_{i\in\lambda_j} Y_i(d') I\{D_i = d'\}\Bigg| \\
&\quad \times I\Bigg\{\Bigg|\sum_{i\in\lambda_j} Y_i(d) I\{D_i = d\} \sum_{i\in\lambda_j} Y_i(d') I\{D_i = d'\}\Bigg| > \frac{\lambda}{2}\Bigg\}\Bigg|X^{(n)}\Bigg] \\
&\quad + \Bigg|\frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)} \sum_{\{i,k\}\subset\lambda_j} \big(\Gamma_d(X_i)\Gamma_{d'}(X_k) + \Gamma_d(X_k)\Gamma_{d'}(X_i)\big) \\
&\quad \times \Bigg|I\Bigg\{\Bigg|\frac{1}{|\mathcal{D}|(|\mathcal{D}|-1)} \sum_{\{i,k\}\subset\lambda_j} \big(\Gamma_d(X_i)\Gamma_{d'}(X_k) + \Gamma_d(X_k)\Gamma_{d'}(X_i)\big)\Bigg| > \frac{\lambda}{2}\Bigg\} \\
&\lesssim E\Bigg[\sum_{i\in\lambda_j} Y_i^2(d) I\{D_i = d\} I\Bigg\{\Bigg|\sum_{i\in\lambda_j} Y_i(d) I\{D_i = d\}\Bigg| > \sqrt{\frac{\lambda}{2}}\Bigg\}\Bigg|X^{(n)}\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&+E\Bigg[\sum_{i\in\lambda_j} Y_i^2(d')I\{D_i=d'\}I\Bigg\{\Bigg|\sum_{i\in\lambda_j} Y_i(d')I\{D_i=d'\}\Bigg|>\sqrt{\frac{\lambda}{2}}\Bigg\}\Big|X^{(n)}\Bigg]\\
&+\sum_{\{i,k\}\subset\lambda_j}\Bigg(\big|\Gamma_d(X_i)\Gamma_{d'}(X_k)\big|I\Bigg\{\big|\Gamma_d(X_i)\Gamma_{d'}(X_k)\big|>\frac{\lambda}{2}\Bigg\}\\
&+\big|\Gamma_d(X_k)\Gamma_{d'}(X_i)\big|I\Bigg\{\big|\Gamma_d(X_k)\Gamma_{d'}(X_i)\big|>\frac{\lambda}{2}\Bigg\}\Bigg)\\
&\le\sum_{i\in\lambda_j} E\Bigg[Y_i^2(d)I\Bigg\{\big|Y_i(d)\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\Big|X_i\Bigg]+\sum_{i\in\lambda_j} E\Bigg[Y_i^2(d')I\Bigg\{\big|Y_i(d')\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\Big|X_i\Bigg]\\
&+\sum_{i\in\lambda_j}\Gamma_d^2(X_i)I\Bigg\{\big|\Gamma_d(X_i)\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}+\sum_{i\in\lambda_j}\Gamma_{d'}^2(X_i)I\Bigg\{\big|\Gamma_{d'}(X_i)\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{1}{|\mathcal{D}|n}\sum_{1\le j\le n} E\big[\big|\hat\rho_{n,j}(d,d')-E\big[\hat\rho_{n,j}(d,d')\big|X^{(n)}\big]\big|\\
&\quad\times I\big\{\big|\hat\rho_{n,j}(d,d')-E\big[\hat\rho_{n,j}(d,d')\big|X^{(n)}\big]\big|>\lambda\big\}\big|X^{(n)}\big]\\
&\lesssim\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n} E\Bigg[Y_i^2(d)I\Bigg\{\big|Y_i(d)\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\Big|X_i\Bigg]\\
&\quad+\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n} E\Bigg[Y_i^2(d')I\Bigg\{\big|Y_i(d')\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\Big|X_i\Bigg]\\
&\quad+\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n}\Gamma_d^2(X_i)I\Bigg\{\big|\Gamma_d(X_i)\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}+\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n}\Gamma_{d'}^2(X_i)I\Bigg\{\big|\Gamma_d(X_i)\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\\
&\xrightarrow{P} E\Bigg[Y_i^2(d)I\Bigg\{\big|Y_i(d)\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\Bigg]+E\Bigg[Y_i^2(d')I\Bigg\{\big|Y_i(d')\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\Bigg]\\
&\quad+E\Bigg[\Gamma_d^2(X_i)I\Bigg\{\big|\Gamma_d(X_i)\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\Bigg]+E\Bigg[\Gamma_{d'}^2(X_i)I\Bigg\{\big|\Gamma_d(X_i)\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\Bigg],
\end{aligned}
$$

where the convergence in probability follows from the weak law or large numbers. Because $E[Y_i^2(d)]<\infty$, $E[Y_i^2(d')]<\infty$, $E[\Gamma_d^2(X_i)]\le E[Y_i^2(d)]<\infty$, and $E[\Gamma_{d'}^2(X_i)]\le E[Y_i^2(d')]<\infty$, we have

$$
\begin{aligned}
\lim_{\lambda\to\infty} E\Bigg[Y_i^2(d)I\Bigg\{\big|Y_i(d)\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\Bigg]=0,\\
\lim_{\lambda\to\infty} E\Bigg[Y_i^2(d')I\Bigg\{\big|Y_i(d')\big|>\sqrt{\frac{\lambda}{4}}\Bigg\}\Bigg]=0,
\end{aligned}
$$

$$\lim_{\lambda\to\infty} E\left[\Gamma_d^2(X_i) I\left\{\left|\Gamma_d(X_i)\right| > \sqrt{\frac{\lambda}{4}}\right\}\right] = 0,$$

$$\lim_{\lambda\to\infty} E\left[\Gamma_{d'}^2(X_i) I\left\{\left|\Gamma_{d'}(X_i)\right| > \sqrt{\frac{\lambda}{4}}\right\}\right] = 0.$$

It follows from a subsequencing argument as in (S.29) of the proof of Lemma S.1.5 of Bai, Romano, and Shaikh (2021) that (14) holds. The conclusion therefore follows. □

LEMMA C.3. *Suppose Assumptions 2.1–2.4 hold. Then* $\hat{\rho}_n(d,d) \xrightarrow{P} E[\Gamma_d^2(X_i)]$ *as* $n \to \infty$.

PROOF OF LEMMA C.3. For $1 \le j \le \frac{n}{2}$, define

$$\hat{\rho}_{n,j}(d,d) = \sum_{i\in\lambda_{2j-1}} Y_i I\{D_i = d\} \sum_{i\in\lambda_{2j}} Y_i I\{D_i = d\}.$$

By definition, $\hat{\rho}_n(d,d) = \frac{2}{n}\sum_{1\le j\le \frac{n}{2}} \hat{\rho}_{n,j}$. Note by Assumption 2.2,

$$E\big[\hat{\rho}_{n,j}(d,d)|X^{(n)}\big] = \frac{1}{|\mathcal{D}|^2} \sum_{i\in\lambda_{2j-1}, k\in\lambda_{2j}} \Gamma_d(X_i)\Gamma_d(X_k).$$

Further, note

$$\Gamma_d(X_i)\Gamma_d(X_k) = \frac{1}{2}\Gamma_d^2(X_i) + \frac{1}{2}\Gamma_d^2(X_k) - \frac{1}{2}\big(\Gamma_d(X_i) - \Gamma_d(X_k)\big)^2.$$

Therefore,

$$\begin{aligned}
&E\big[\hat{\rho}_n(d,d)|X^{(n)}\big] \\
&\quad= \frac{2}{n}\sum_{1\le j\le\frac{n}{2}} E\big[\hat{\rho}_{n,j}(d,d)|X^{(n)}\big] \\
&\quad= \frac{2}{n}\sum_{1\le j\le\frac{n}{2}} \frac{1}{|\mathcal{D}|^2} \sum_{i\in\lambda_{2j-1}, k\in\lambda_{2j}} \left(\frac{1}{2}\Gamma_d^2(X_i) + \frac{1}{2}\Gamma_d^2(X_k) - \frac{1}{2}\big(\Gamma_d(X_i) - \Gamma_d(X_k)\big)^2\right) \\
&\quad= \frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n} \Gamma_d^2(X_i) - \frac{1}{n|\mathcal{D}|^2}\sum_{1\le j\le\frac{n}{2}} \sum_{i\in\lambda_{2j-1}, k\in\lambda_{2j}} \big(\Gamma_d(X_i) - \Gamma_d(X_k)\big)^2 \\
&\quad\xrightarrow{P} E\big[\Gamma_d^2(X_i)\big],
\end{aligned}$$

where the convergence in probability follows from Assumptions 2.1(c) and 2.4 as well as the weak law of large numbers. To conclude the proof, we show

$$\hat{\rho}_n(d,d) - E\big[\hat{\rho}_n(d,d)|X^{(n)}\big] \xrightarrow{P} 0. \tag{15}$$

In order for this, we proceed to verify the uniform integrability condition in Lemma S.1.3 of Bai, Romano, and Shaikh (2021) conditional on $X^{(n)}$. In what follows, we repeatedly

use the following inequalities:

$$I\left\{\left|\sum_{1\le j\le k} a_j\right| > \lambda\right\} \le \sum_{1\le j\le k} I\left\{|a_j| > \frac{\lambda}{k}\right\},$$

$$\left|\sum_{1\le j\le k} a_j\right| I\left\{\left|\sum_{1\le j\le k} a_j\right| > \lambda\right\} \le \sum_{1\le j\le k} k|a_j| I\left\{|a_j| > \frac{\lambda}{k}\right\},$$

$$|ab| I\{|ab| > \lambda\} \le a^2 I\{|a| > \sqrt{\lambda}\} + b^2 I\{|b| > \sqrt{\lambda}\}.$$

We will also repeatedly use the facts that $0 \le I\{D_i = d\} \le 1$ and $I\{D_i = d\} I\{D_k = d\} = 0$ for $i \ne k$ in the same stratum. Note

$$\begin{aligned}
&E\left[\left|\hat{\rho}_{n,j}(d,d) - E\left[\hat{\rho}_{n,j}(d,d)|X^{(n)}\right]\right| I\left\{\left|\hat{\rho}_{n,j}(d,d) - E\left[\hat{\rho}_{n,j}(d,d)|X^{(n)}\right]\right| > \lambda\right\} \middle| X^{(n)}\right] \\
&\quad \le E\left[\left|\hat{\rho}_{n,j}(d,d)\right| I\left\{\left|\hat{\rho}_{n,j}(d,d)\right| > \frac{\lambda}{2}\right\} \middle| X^{(n)}\right] \\
&\qquad + E\left[\left|E\left[\hat{\rho}_{n,j}(d,d)|X^{(n)}\right]\right| I\left\{\left|E\left[\hat{\rho}_{n,j}(d,d)|X^{(n)}\right]\right| > \frac{\lambda}{2}\right\} \middle| X^{(n)}\right] \\
&\quad = E\left[\left|\hat{\rho}_{n,j}(d,d)\right| I\left\{\left|\hat{\rho}_{n,j}(d,d)\right| > \frac{\lambda}{2}\right\} \middle| X^{(n)}\right] \\
&\qquad + \left|E\left[\hat{\rho}_{n,j}(d,d)|X^{(n)}\right]\right| I\left\{\left|E\left[\hat{\rho}_{n,j}(d,d)|X^{(n)}\right]\right| > \frac{\lambda}{2}\right\} \\
&\quad \le E\left[\left|\sum_{i\in\lambda_{2j-1}} Y_i(d) I\{D_i = d\} \sum_{i\in\lambda_{2j}} Y_i(d) I\{D_i = d\}\right|\right. \\
&\qquad \left. \times I\left\{\left|\sum_{i\in\lambda_{2j-1}} Y_i(d) I\{D_i = d\} \sum_{i\in\lambda_{2j}} Y_i(d) I\{D_i = d\}\right| > \frac{\lambda}{2}\right\} \middle| X^{(n)}\right] \\
&\qquad + \left|\frac{1}{|\mathcal{D}|^2} \sum_{i\in\lambda_{2j-1}, k\in\lambda_{2j}} \Gamma_d(X_i)\Gamma_d(X_k)\right| I\left\{\left|\frac{1}{|\mathcal{D}|^2} \sum_{i\in\lambda_{2j-1}, k\in\lambda_{2j}} \Gamma_d(X_i)\Gamma_d(X_k)\right| > \frac{\lambda}{2}\right\} \\
&\quad \lesssim E\left[\sum_{i\in\lambda_{2j-1}} Y_i^2(d) I\{D_i = d\} I\left\{\left|\sum_{i\in\lambda_{2j-1}} Y_i(d) I\{D_i = d\}\right| > \sqrt{\frac{\lambda}{2}}\right\} \middle| X^{(n)}\right] \\
&\qquad + E\left[\sum_{i\in\lambda_{2j}} Y_i^2(d) I\{D_i = d\} I\left\{\left|\sum_{i\in\lambda_{2j}} Y_i(d) I\{D_i = d\}\right| > \sqrt{\frac{\lambda}{2}}\right\} \middle| X^{(n)}\right] \\
&\qquad + \sum_{i\in\lambda_{2j-1}, k\in\lambda_{2j}} \left|\Gamma_d(X_i)\Gamma_d(X_k)\right| I\left\{\left|\Gamma_d(X_i)\Gamma_d(X_k)\right| > \frac{\lambda}{2}\right\} \\
&\quad \le \sum_{i\in\lambda_{2j-1}} E\left[Y_i^2(d) I\left\{\left|Y_i(d)\right| > \sqrt{\frac{\lambda}{2}}\right\} \middle| X_i\right] + \sum_{i\in\lambda_{2j}} E\left[Y_i^2(d) I\left\{\left|Y_i(d)\right| > \sqrt{\frac{\lambda}{2}}\right\} \middle| X_i\right]
\end{aligned}$$

$$+\sum_{i\in\lambda_{2j-1}}\Gamma_d^2(X_i)I\left\{\left|\Gamma_d(X_i)\right|>\sqrt{\frac{\lambda}{2}}\right\}+\sum_{i\in\lambda_{2j}}\Gamma_d^2(X_i)I\left\{\left|\Gamma_d(X_i)\right|>\sqrt{\frac{\lambda}{2}}\right\}.$$

Therefore,

$$\begin{aligned}
&\frac{1}{|\mathcal{D}|n}\sum_{1\le j\le\frac{n}{2}}E\big[\big|\hat\rho_{n,j}(d,d)-E\big[\hat\rho_{n,j}(d,d)|X^{(n)}\big]\big|\\
&\qquad\times I\big\{\big|\hat\rho_{n,j}(d,d)-E\big[\hat\rho_{n,j}(d,d)|X^{(n)}\big]\big|>\lambda\big\}\big|X^{(n)}\big]\\
&\lesssim\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n}E\left[Y_i^2(d)I\left\{\left|Y_i(d)\right|>\sqrt{\frac{\lambda}{2}}\right\}\Big|X_i\right]\\
&\qquad+\frac{1}{|\mathcal{D}|n}\sum_{1\le i\le|\mathcal{D}|n}\Gamma_d^2(X_i)I\left\{\left|\Gamma_d(X_i)\right|>\sqrt{\frac{\lambda}{2}}\right\}\\
&\xrightarrow{P}E\left[Y_i^2(d)I\left\{\left|Y_i(d)\right|>\sqrt{\frac{\lambda}{2}}\right\}\right]+E\left[\Gamma_d^2(X_i)I\left\{\left|\Gamma_d(X_i)\right|>\sqrt{\frac{\lambda}{2}}\right\}\right],
\end{aligned}$$

where the convergence in probability follows from the weak law or large numbers. Because $E[Y_i^2(d)]<\infty$ and $E[\Gamma_d^2(X_i)]\le E[Y_i^2(d)]<\infty$, we have

$$\lim_{\lambda\to\infty}E\left[Y_i^2(d)I\left\{\left|Y_i(d)\right|>\sqrt{\frac{\lambda}{4}}\right\}\right]=0,$$

$$\lim_{\lambda\to\infty}E\left[\Gamma_d^2(X_i)I\left\{\left|\Gamma_d(X_i)\right|>\sqrt{\frac{\lambda}{4}}\right\}\right]=0.$$

It follows from a subsequencing argument as in the proof of Lemma S.1.5 of Bai, Romano, and Shaikh (2021) that (15) holds. The conclusion therefore follows. □

LEMMA C.4. *Suppose* $(Y_i, X_{1,i}', X_{2,i}')'$, $1\le i\le n$ *is an i.i.d. sequence of random vectors, where* Y_i *takes values in* $\mathbf{R}$, $X_{1,i}$ *takes values in* $\mathbf{R}^{k_1}$, *and* $X_{2,i}$ *takes values in* $\mathbf{R}^{k_2}$. *Consider the linear regression*

$$Y_i=X_{1,i}'\beta_1+X_{2,i}'\beta_2+\epsilon_i.$$

Define $\mathbb{X}=(X_1,\ldots,X_n)'$, $\mathbb{X}_1=(X_{1,1},\ldots,X_{1,n})'$, *and* $\mathbb{X}_2=(X_{2,1},\ldots,X_{2,n})'$. *Define* $\mathbb{P}_2=\mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\mathbb{X}_2'$ *and* $\mathbb{M}_2=\mathbb{I}-\mathbb{P}_2$. *Let* $\hat\beta_{1,n}$ *and* $\hat\beta_{2,n}$ *denote the OLS estimator of* β_1 *and* β_2. *Define* $\hat\epsilon_i=Y_i-X_{1,i}'\hat\beta_{1,n}-X_{2,i}'\hat\beta_{2,n}$. *Define*

$$\tilde{\mathbb{X}}_1=\mathbb{M}_2\mathbb{X}_1.$$

Let

$$\hat\Omega_n=\big(\mathbb{X}'\mathbb{X}\big)^{-1}\big(\mathbb{X}'\mathrm{diag}\big(\hat\epsilon_i^2:1\le i\le n\big)\mathbb{X}\big)\big(\mathbb{X}'\mathbb{X}\big)^{-1}$$

denote the heteroskedasticity-robust variance estimator of $(\hat\beta_{1,n},\hat\beta_{2,n})$. *Then the upper-left* $k_1\times k_1$ *block of* $\hat\Omega_n$ *equals*

$$\big(\tilde{\mathbb{X}}_1'\tilde{\mathbb{X}}_1\big)^{-1}\big(\tilde{\mathbb{X}}_1'\mathrm{diag}\big(\hat\epsilon_i^2:1\le i\le n\big)\tilde{\mathbb{X}}_1\big)\big(\tilde{\mathbb{X}}_1'\tilde{\mathbb{X}}_1\big)^{-1}.$$

PROOF OF LEMMA C.4. By the formula for the inverse of a partitioned matrix, the first k_1 rows of $(\mathbb{X}'\mathbb{X})^{-1}$ equal

$$\begin{pmatrix} (\mathbb{X}_1'\mathbb{M}_2\mathbb{X}_1)^{-1} & -(\mathbb{X}_1'\mathbb{M}_2\mathbb{X}_1)^{-1}\mathbb{X}_1'\mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1} \end{pmatrix}.$$

Furthermore,

$$\mathbb{X}'\mathrm{diag}(\hat{\epsilon}_i^2 : 1 \le i \le n)\mathbb{X} = \begin{pmatrix} \mathbb{X}_1'\mathrm{diag}(\hat{\epsilon}_i^2 : 1 \le i \le n)\mathbb{X}_1 & \mathbb{X}_1'\mathrm{diag}(\hat{\epsilon}_i^2 : 1 \le i \le n)\mathbb{X}_2 \\ \mathbb{X}_2'\mathrm{diag}(\hat{\epsilon}_i^2 : 1 \le i \le n)\mathbb{X}_1 & \mathbb{X}_2'\mathrm{diag}(\hat{\epsilon}_i^2 : 1 \le i \le n)\mathbb{X}_2 \end{pmatrix}.$$

The conclusion then follows from elementary calculations. □

REFERENCES

Abadie, Alberto and Guido W. Imbens (2008), "Estimation of the conditional variance in paired experiments." *Annales d'Économie et de Statistique*, 91/92, 175–187. [286]

Alatas, Vivi, Abhijit Banerjee, Rema Hanna, Benjamin A. Olken, and Julia Tobias (2012), "Targeting the poor: Evidence from a field experiment in Indonesia." *American Economic Review*, 102 (4), 1206–1240. [280]

Athey, Susan and Guido W. Imbens (2017), "The econometrics of randomized experiments." In *Handbook of Economic Field Experiments*, Vol. 1, 73–140, Elsevier. [281, 291]

Bai, Yuehao (2022), "Optimality of matched-pair designs in randomized controlled trials." *American Economic Review*, 112 (12), 3911–3940. https://www.aeaweb.org/articles?id=10.1257/aer.20201856. [281, 292, 298, 315, 316, 317]

Bai, Yuehao, Jizhou Liu, and Max Tabord-Meehan (2024), "Supplement to 'Inference for matched tuples and fully blocked factorial designs'." *Quantitative Economics Supplemental Material*, 15, https://doi.org/10.3982/QE2354. [282]

Bai, Yuehao, Joseph P. Romano, and Azeem M. Shaikh (2021), "Inference in experiments with matched pairs." *Journal of the American Statistical Association*, 0 (ja), 1–37. Publisher: Taylor & Francis. https://doi.org/10.1080/01621459.2021.1883437. [281, 283, 284, 286, 287, 288, 289, 307, 309, 310, 319, 321, 322, 323, 325, 327]

Besedeš, Tibor, Cary Deck, Sudipta Sarangi, and Mikhael Shor (2012), "Age effects and heuristics in decision making." *Review of Economics and Statistics*, 94 (2), 580–595. [280, 296]

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur (2018), "Experimental evidence on scaling up education reforms in Kenya." *Journal of Public Economics*, 168, 1–20. http://www.sciencedirect.com/science/article/pii/S0047272718301518. [280, 287]

Branson, Zach, Tirthankar Dasgupta, and Donald B. Rubin (2016), "Improving covariate balance in 2K factorial designs via rerandomization with an application to a New York City Department of Education High School Study." *The Annals of Applied Statistics*, 10 (4), 1958–1976. https://doi.org/10.1214/16-AOAS959. [281, 282, 295, 298, 301]

Brown, Christina and Tahir Andrabi (2020), "Inducing positive sorting through performance pay: Experimental evidence from Pakistani schools." [280, 287]

Bruhn, Miriam and David McKenzie (2009), "In pursuit of balance: Randomization in practice in development field experiments." *American economic journal: applied economics*, 1 (4), 200–232. [280]

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh (2018), "Inference under covariate-adaptive randomization." *Journal of the American Statistical Association*, 113 (524), 1784–1796. https://doi.org/10.1080/01621459.2017.1375934. PMID: 30906087. [300, 317]

Bugni, Federico A., Ivan A. Canay, and Azeem M. Shaikh (2019), "Inference under covariate-adaptive randomization with multiple treatments." *Quantitative Economics*, 10 (4), 1747–1785. https://onlinelibrary.wiley.com/doi/abs/10.3982/QE1150. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE1150. [300, 317]

Cytrynbaum, Max (2021), "Designing representative and balanced experiments by local randomization." arXiv preprint arXiv:2111.08157. [281]

Dasgupta, Tirthankar, Natesh S. Pillai, and Donald B. Rubin (2015), "Causal inference from 2^k factorial designs by using potential outcomes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77 (4), 727–753. [281, 282, 292, 293, 294]

de Chaisemartin, Clément and Jaime Ramirez-Cuellar (2022), "At what level should one cluster standard errors in paired and small-strata experiments?" https://arxiv.org/abs/1906.00288. [281, 282, 289]

deMel, Suresh, David McKenzie, and Christopher Woodruff (2013), "The demand for, and consequences of, formalization among informal firms in Sri Lanka." *American Economic Journal: Applied Economics*, 5 (2), 122–150. https://www.aeaweb.org/articles?id=10.1257/app.5.2.122. [280, 288]

DellaVigna, Stefano, John A. List, Ulrike Malmendier, and Gautam Rao (2016), "Voting to tell others." *The Review of Economic Studies*, 84 (1), 143–181. [280]

Fafchamps, Marcel, David McKenzie, Simon Quinn, and Christopher Woodruff (2014), "Microenterprise growth and the flypaper effect: Evidence from a randomized experiment in Ghana." *Journal of Development Economics*, 106, 211–226. https://www.sciencedirect.com/science/article/pii/S0304387813001375. [280, 282, 284, 288, 304, 305]

Fogarty, Colin B. (2018), "Regression-assisted inference for the average treatment effect in paired experiments." *Biometrika*, 105 (4), 994–1000. [281]

Imai, Kosuke, Gary King, and Clayton Nall (2009), "The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation." *Statistical Science*, 24 (1), 29–53. [281]

Jiang, Liang, Xiaobin Liu, Peter C. B. Phillips, and Yichong Zhang (2020), "Bootstrap inference for quantile treatment effects in randomized experiments with matched pairs." *The Review of Economics and Statistics*, 1–47. [281]

Karlan, Dean, Robert Osei, Isaac Osei-Akoto, and Christopher Udry (2014), "Agricultural decisions after relaxing credit and risk constraints." *The Quarterly Journal of Economics*, 129 (2), 597–652. [280, 292]

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan (2015), "Self-control at work." *Journal of Political Economy*, 123 (6), 1227–1277. [280]

Li, Xinran, Peng Ding, and Donald B. Rubin (2020), "Rerandomization in 2^K factorial experiments." *The Annals of Statistics*, 48 (1), 43–63. [281, 282, 293, 295, 299]

Liu, Hanzhong, Jiyang Ren, and Yuehan Yang (2022), "Randomization-based joint central limit theorem and efficient covariate adjustment in randomized block 2^k factorial experiments." *Journal of the American Statistical Association*, 1–15. [281, 282]

MacKinnon, James G. and Halbert White (1985), "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties." *Journal of econometrics*, 29 (3), 305–325. [288]

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich (2019), "Factorial designs, model selection, and (incorrect) inference in randomized experiments." Technical Report, National Bureau of Economic Research. [281]

Pashley, Nicole E. and Marie-Abele C. Bind (2019), "Causal inference for multiple treatments using fractional factorial designs." arXiv preprint arXiv:1905.07596. [281, 297]

van der Laan, Mark J., Laura B. Balzer, and Maya L. Petersen (2012), "Adaptive matching in randomized trials and observational studies." *Journal of statistical research*, 46 (2), 113. [281]

Wu, C. F. Jeff and Michael S. Hamada (2011), *Experiments: Planning, Analysis, and Optimization*, Vol. 552. John Wiley & Sons. [280, 292, 294, 297]

Co-editor James Hamilton handled this manuscript.

Manuscript received 14 March, 2023; final version accepted 10 February, 2024; available online 13 February, 2024.

The replication package for this paper is available at https://doi.org/10.5281/zenodo.10601297. The Journal checked the data and codes included in the package for their ability to reproduce the results in the paper and approved online appendices.