Why Randomize? Minimax Optimality under Permutation Invariance

Yuehao Bai
Department of Economics
University of Michigan
yuehaob@umich.edu

October 25, 2020

Abstract

This paper studies finite sample minimax optimal randomization schemes and estimation schemes in estimating parameters including the average treatment effect, when treatment effects are heterogeneous. A randomization scheme is a distribution over a group of permutations of a given treatment assignment vector. An estimation scheme is a joint distribution over assignment vectors, linear estimators, and permutations of assignment vectors. The key element in the minimax problem is that the worst case is over a class of distributions of the data which is invariant to a group of permutations. First, I show that given any assignment vector and any estimator, the uniform distribution over the same group of permutations, namely the complete randomization scheme, is minimax optimal. Second, under further assumptions on the class of distributions and the objective function, I show the minimax optimal estimation scheme involves completely randomizing an assignment vector, while the optimal estimator is the difference-in-means under complete invariance and a weighted average of within-block differences under a block structure, and the numbers of treated and untreated units are determined by Neyman allocations.

KEYWORDS: Experimental design, permutation invariance, average treatment effect

JEL classification codes: C13, C90, C93

1 Introduction

This paper studies finite sample minimax optimal randomization schemes and estimation schemes in estimating parameters including the average treatment effect (ATE), when treatment effects are heterogeneous. A randomization scheme is a distribution over a group of permutations of a given treatment assignment vector. An estimation scheme is a joint distribution over assignment vectors, linear estimators, and permutations of assignment vectors. Randomization is prevalent in the sciences, most notably in randomized controlled trials. See Bruhn and McKenzie (2009) for a review of field experiments and Rosenberger and Lachin (2015) for a review of clinical trials. Randomization is often motivated by the fact that it allows for frequentist inference and it stops subjects from manipulating themselves towards being treated or untreated, but rarely because it helps estimating the ATE. In fact, if the objective is to estimate the ATE precisely, Kasy (2016) recommends experimental designers to not randomize, because in a Bayesian framework with a fixed prior on the distribution of potential outcomes, the optimal assignment scheme never involves randomization.

This paper instead considers a finite sample minimax problem where the objective function satisfies quasiconvexity and the worst case is over a class of conditional distributions given observed covariates. The key assumption is that the class is invariant to a group of permutations. The permutation invariance assumption reflects that the distributions of experimental units could only be distinguished up to some point. When there is absolutely no information on how observed covariates affect the potential outcomes, I assume the class of distributions is invariant to the set of all permutations of the unit indices. When the designer believes the distributions differ according to the values of a discrete covariate, but cannot distinguish the units with the same discrete covariate values, I assume the class of distributions is invariant to the set of permutations which only permute within each block of units with the same value of this discrete covariate. I further assume the objective function is quasiconvex, so it can be the expectation operator, the quantile function, or the survival function. My first main result shows that given any assignment vector and any invariant estimator, the uniform distribution over the same group of permutations, namely the complete randomization scheme, is minimax optimal.

The first main result takes the estimator as given and completely randomizes a given assignment vector, which means that the optimal numbers of treated and control units and the optimal estimator are undetermined. Under further assumptions on the class of distributions and the objective function, my second main result shows the minimax optimal estimation scheme involves completely randomizing a certain assignment vector, while the optimal linear estimator is the difference-in-means estimator under complete invariance and a weighted average of within-block differences under a block structure, and the numbers of treated and control units are determined by Neyman allocations in Neyman (1934), where the ratios depend on the worst case conditional variances.

The study of minimax optimality of randomization schemes dates back to Wu (1981), Li (1983), and Hooper (1989), but all of these papers assume homogeneous treatment effects, and do not have the results we provide on optimal linear estimators. There is a large literature on minimax estimation, including Donoho (1994) and Armstrong and Kolesár (2018), but these papers either don't involve randomization or take the randomization schemes as given. Bai (2020) characterizes optimal stratifications in terms of the MSE of the difference-in-means estimator, and provides an algorithm to calculate minimax matched-pair designs, but again taking as given that estimator. Under the permutation invariance assumption, this paper derives the optimal estimation scheme, which jointly determines the randomization scheme and the linear estimator.

The paper is organized as follows. Section 2 proves the minimax optimality of the complete randomization scheme for any given assignment vector and any invariant estimator. Section 3 derives the optimal estimation scheme. Section 4 concludes. Appendix A contains all the proofs.

2 Optimal Randomization Scheme

Suppose we have n independent units, indexed by $1 \leq i \leq n$. Denote by $Y_i(1)$ the potential outcome of the ith unit if treated and by $Y_i(0)$ if not treated. Let Z_i denote the observed covariates of the ith unit. For $a \in \{0, 1\}$, the potential outcome

$Y_i(a)$ could be decomposed as

$$Y_i(a) = \mu_a + m_a(Z_i) + \epsilon_i(a) \ , \tag{1}$$

where $m_a(Z_i) = E[Y_i(a)|Z_i] - \mu_a$ and $\epsilon_i(a) = Y_i(a) - E[Y_i(a)|Z_i]$. Further define

$$\begin{aligned} W_i &= (Y_i(0), Y_i(1))' \\ \mu &= (\mu_0, \mu_1)' \\ \epsilon_i &= (\epsilon_i(0), \epsilon_i(1))' \\ m(Z_i) &= (m_0(Z_i), m_1(Z_i))' \ . \end{aligned}$$

Define $\mathbf{W} = (W_1, \ldots, W_n)'$ and similarly $\mathbf{Z}$, $\boldsymbol{\epsilon}$, and $\mathbf{m}$. Define $\xi = \mathcal{L}(\boldsymbol{\epsilon}|\mathbf{Z})$, the distribution of $\boldsymbol{\epsilon}$ conditional on $\mathbf{Z}$.

Let H denote a group of permutations of $\{1, \ldots, n\}$. See Section 2.2 of Artin (2011) for the definition of a group. An element of H is denoted by $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$, which maps each integer from 1 to n to an integer in the same set. I identify each $\pi \in H$ as a $n \times n$ permutation matrix so that the $(i, \pi^{-1}(i))$-th element is 1 for $1 \leq i \leq n$ and the rest are 0. Pre-multiplying the vector $(1, \ldots, n)'$ by the matrix π generates a vector with i on the $\pi(i)$-th position. For instance, if $\pi(5) = 1, \pi(3) = 2, \pi(2) = 3, \pi(4) = 4, \pi(1) = 5$, then π is described by

$$\begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \xrightarrow{\pi} \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \\ 2 \\ 4 \\ 1 \end{bmatrix} \ .$$

H doesn't have to include all permutations of $\{1, \ldots, n\}$. For any vector X, $\pi X = (X_{\pi^{-1}(1)}, \ldots, X_{\pi^{-1}(n)})$, so permuting a vector is equivalent to applying the inverse of permutation to its indices. For a matrix $\mathbf{X} = (X_1, \ldots, X_n)'$, define $\pi\mathbf{X} = (X_{\pi^{-1}(1)}, \ldots, X_{\pi^{-1}(n)})'$. Furthermore, if $\xi = \mathcal{L}(\boldsymbol{\epsilon}|\mathbf{Z})$, then I define $\pi\xi = \mathcal{L}(\pi\boldsymbol{\epsilon}|\mathbf{Z})$.

Define $s = (\mu, \mathbf{m}, \xi)$ as the state and $\pi s = (\mu, \pi\mathbf{m}, \pi\xi)$. I make the following assumption on S, the set of possible states.

Assumption 2.1. S is H-invariant, i.e., $s \in S \Rightarrow \pi s \in S$ for all $\pi \in H$.

The following examples demonstrate Assumption 2.1.

Example 2.1 (Complete Invariance)**.** In model (1), denote by M and $\mathcal{E}$ the set of possible values of $\mathbf{m}$ and ξ. Let H be the set of all permutations of $\{1, \ldots, n\}$. Assumption 2.1 implies for any permutation $\pi \in H$,

$$\mathbf{m} \in M \Rightarrow \pi\mathbf{m} \in M$$
$$\xi \in \mathcal{E} \Rightarrow \pi\xi \in \mathcal{E} \ .$$

Complete invariance is natural when the designer doesn't have ex-ante information on either how Z_i affects $Y_i(a)$ for $a \in \{0, 1\}$ or how $\epsilon_i(a)$ is distributed conditional on Z_i. ■

Example 2.2 (Block Model)**.** Suppose $Z_i = (Z_{1i}, Z_{2i})$ where $Z_{1i} \in \{1, \ldots, B\}$. Consider model (1) with

$$m_a(Z_i) = \gamma_{Z_{1i}}(a) + \tilde{m}_a(Z_i) \tag{2}$$

for $\gamma_b(a) = E[Y_i(a)|Z_{1i} = b]$ and $\tilde{m}_a(Z_i) = E[Y_i(a)|Z_i] - \gamma_{Z_{1i}}(a)$. Suppose H consists of all permutations that respects block structure:

$$H = \{\pi : Z_{1\pi^{-1}(i)} = Z_{1i} \text{ for } i = 1, \ldots, n\} \ .$$

This choice of H reflects the belief that in determining the potential outcomes, different values of Z_{1i} implies different conditional means of $Y_i(a)$, but the designer does not know how Z_{2i} affects $Y_i(a)$ and $\epsilon_i(a)$. Note that $\pi\mathbf{m} = (\gamma, \pi\tilde{\mathbf{m}})$ for $\tilde{\mathbf{m}} = (\tilde{m}(Z_1), \ldots, \tilde{m}(Z_n))'$ and $\tilde{m}(Z_i) = (\tilde{m}_0(Z_i), \tilde{m}_1(Z_i))'$. ■

Let A_i denote the treatment status of the ith unit. For convenience of notation, as we will see shortly, we deviate from the convention that A_i is binary, but instead let $A_i \in \{(0, 1)', (1, 0)'\}$. Define $A_i = (0, 1)'$ if the ith unit is treated and $A_i = (1, 0)'$ if untreated. Define the matrix $\mathbf{A} = \sum_{1 \le i \le n} e_{ii} \otimes A_i'$ where e_{ii} is $n \times n$ matrix with

all zeros except the (i,i)-th element being 1. For example, if

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} ,$$

then units $i = 1, 2$ are treated and $i = 3, 4$ are untreated.

The observed outcome for the ith unit is

$$Y_i = A_i' W_i \ , \tag{3}$$

and the matrix version is

$$\mathbf{Y} = \mathbf{A} \operatorname{vec}(\mathbf{W}') = \mathbf{A}(\mathbf{1}_n \otimes \mu + \operatorname{vec}(\mathbf{m}' + \boldsymbol{\epsilon}')) \ , \tag{4}$$

where $\mathbf{1}_n$ is the $n \times 1$ vector of 1's.

I now formally define randomization schemes. A randomization scheme $\mathbf{G}$ is a random variable on H. Recall that $\mathbf{H}$ follows the uniform distribution on a group H if $\pi\mathbf{H} \overset{d}{=} \mathbf{H} \overset{d}{=} \mathbf{H}\pi$ for all $\pi \in H$. I define the complete randomization scheme as a uniformly distributed $\mathbf{H}$ on H. Let Φ denote the set of all randomization schemes.

I use an estimator $\delta \in \mathbb{R}$ which depends only on $(\mathbf{A}, \mathbf{Y})$, the observed data. Since the set of distributions is permutation invariant, a reasonable estimator should be invariant to $\pi \in H$ too:

Assumption 2.2. $\delta(\mathbf{A}, \mathbf{Y}) = \delta(\pi\mathbf{A}, \pi\mathbf{Y})$ for any $\pi \in H$.

In Example 2.1, Assumption 2.2 implies that Z_i should not affect the estimator at all, which is reasonable because we assume all units are permutation invariant. This further implies that all units in the treated group should enter the estimator symmetrically, and the same holds for the control group. In Example 2.2, Assumption 2.2 implies that the estimator may depend on Z_{1i} but not Z_{2i}, also reasonable because we assume all units within a block are permutation invariant. This further implies that all units which are treated in the same block should enter the estimator symmetrically, and the same holds for the control group. We need Assumption 2.2

for Theorem 2.1 below to hold. Theorem 3.1, however, doesn't require Assumption 2.2, but rather often implies that the optimal linear estimator must satisfy Assumption 2.2 as a consequence. See Remark 3.1 for details.

I impose an additional assumption on the loss function L, which represents the designer is only interested in μ instead of $\mathbf{m}$ or ξ.

Assumption 2.3. $L(s,d) = L(\pi s, d)$ for all $\pi \in H$, $s \in S$ and $d \in \mathbb{R}$.

An example that satisfies Assumptions 2.2 and 2.3 is the Average Treatment Effect (ATE), $\delta(\mathbf{A}, \mathbf{Y}) = \bar{Y}_1 - \bar{Y}_0$ with the squared loss $L(s,d) = (\bar{Y}_1 - \bar{Y}_0 - (\mu_1 - \mu_0))^2$, where

$$\bar{Y}_0 = \frac{\sum_{1 \le i \le n} Y_i I\{A_i = (1,0)'\}}{\sum_{1 \le i \le n} I\{A_i = (1,0)'\}} \ , \tag{5}$$

the mean of the control group, and similarly for $\bar{Y}_1$.

Define the risk $r(\pi, \mathbf{A}, \delta; s) = E_s[L(s, \delta(\pi\mathbf{A}, \mathbf{Y}))|\mathbf{Z}]$, where the expectation is based on the state s and conditional on the covariates $\mathbf{Z}$. $r(\mathbf{G}, \mathbf{A}, \delta; s)$ is then a random variable on $\mathbb{R}$. I summarize the distribution of $r(\mathbf{G}, \mathbf{A}, \delta; s)$ by a functional f which maps each distribution on $\mathbb{R}$ to a real number. A typical example of f is the expectation operator, i.e., $f(r(\mathbf{G}, \mathbf{A}, \delta; s)) = E[r(\mathbf{G}, \mathbf{A}, \delta; s)]$, but the results below hold for a general class of f. With some abuse of notation, we let $f(r(\mathbf{G}, \mathbf{A}, \delta; s))$ denote the value of f at the distribution of $r(\mathbf{G}, \mathbf{A}, \delta; s)$.

Assumption 2.4. The functional f on the set of distributions on $\mathbb{R}$ is quasiconvex, i.e., $f(\lambda\mu_1 + (1-\lambda)\mu_2) \le \max\{f(\mu_1), f(\mu_2)\}$ for all distributions μ_1, μ_2 on $\mathbb{R}$ and for all $\lambda \in [0,1]$.

Besides the expectation operator, other examples which satisfy Assumption 2.4 include the quantile function and the survival function $1 - F$, where F is the distribution function. With the survival function the distributions of the risk are compared in terms of First Order Stochastic Dominance.

The following theorem shows the minimax optimality of the complete randomization scheme, for any given assignment vector $\mathbf{A}$ and any estimator δ.

Theorem 2.1. *Suppose Assumptions 2.1–2.4 hold. Then, for any* $\mathbf{A}$ *and* δ, $\mathbf{H}$ *solves*

$$\min_{\mathbf{G} \in \Phi} \max_{s \in S} f(r(\mathbf{G}, \mathbf{A}, \delta; s)) \ . \tag{6}$$

Remark 2.1. In Example 2.1, Theorem 2.1 implies the designer should completely randomize across all units. In Example 2.2, it implies the designer should completely randomize within each block, but not across blocks. ■

Remark 2.2. The intuition why Theorem 2.1 holds is as follows. To begin with, we need only show the conclusion with the worst case over the orbit of each fixed state s_0, $\{\pi s_0 : \pi \in H\}$. But Assumptions 2.1–2.3 imply πs_0 generates the same distribution as $r(\pi^{-1}\mathbf{G}, \mathbf{A}, \delta; s)$. The problem then reduces to finding the maximum of $f(r(\pi^{-1}\mathbf{G}, \mathbf{A}, \delta; s))$ over π. Since $\mathbf{HG} \overset{d}{=} \mathbf{H}$, the average over $\pi \in \mathbf{H}$ of $r(\pi^{-1}\mathbf{G}, \mathbf{A}, \delta; s)$ has the same distribution as $r(\mathbf{H}, \mathbf{A}, \delta; s)$. Since f is quasiconvex by Assumption 2.4, the conclusion follows. ■

Remark 2.3. The result in Theorem 2.1 could be understood as follows. Consider a two-player zero-sum game between the designer and nature, in which the designer chooses the randomization scheme and nature chooses the distribution of the data. If the designer has no idea of the choice of nature, but only the set of choices is permutation invariant, the designer should completely randomize to guard against the worst case, because the objective function is quasiconvex. If $\mathbf{G}$ is a degenerate distribution so that the designer never randomizes, and complete invariance in Example 2.1 holds, then the payoff of the designer is very low at the most adverse distribution of the data. ■

3 Optimal Estimation Scheme

Theorem 2.1 holds under any $\mathbf{A}$ and δ, but it is possible that there does not exist an optimal $\mathbf{A}$ and δ. In this section, I show that optimal $\mathbf{A}$ and δ often exist in very simple forms, under further assumptions on the distributions of the data and the loss function.

I continue to work with the model defined by (1) and (4), but with the quadratic loss function and linear estimators. I assume the parameter of interest is

$$\tau'\mu \text{ for } \tau = (\tau_0, \tau_1)' \,. \tag{7}$$

As an example, if $\tau = (-1, 1)'$, then $\tau'\mu$ is the ATE. Define

$$\mathcal{A}_n = \left\{ \sum_{i=1}^{n} e_{ii} \otimes A_i' : A_i \in \{(1,0)', (0,1)'\} \text{ for } 1 \leq i \leq n \right\} .$$

The estimator is $\delta(\mathbf{A}, \mathbf{Y}) = \beta(\mathbf{A})'\mathbf{Y}$ for some function $\beta : \mathcal{A}_n \to \mathbb{R}^n$ and

$$L(s, d) = \|d - \tau'\mu\|^2 . \tag{8}$$

I formally define an estimation scheme as a joint distribution $\mathcal{L}(\mathbf{A}, \beta, \pi)$ on $\mathcal{A}_n \times \mathbb{R}^n \times H$. I allow $\mathbf{A}$, β, and π to be random, but will see below that the optimal $\mathbf{A}$ is often fixed, and the optimal β is often simply a fixed function of $\mathbf{A}$.

For convenience of notations, I define $U_i(a) = m_a(Z_i) + \epsilon_i(a)$, $U_i = (U_i(0), U_i(1))'$, and $\mathbf{U} = (U_1, \ldots, U_n)'$. Define $\sigma_i^2(a, Z_i) = E[\epsilon_i^2(a)|Z_i]$. Also define $P_{\mathbf{U}} = \mathcal{L}(\mathbf{U}|\mathbf{Z})$ and $\pi P_{\mathbf{U}} = \mathcal{L}(\pi\mathbf{U}|\mathbf{Z})$ if $P_{\mathbf{U}} = \mathcal{L}(\mathbf{U}|\mathbf{Z})$. Denote by $\mathcal{P}_{\mathbf{U}}$ the set of all possible $P_{\mathbf{U}}$'s. Since $\mathbf{U}$ is unobserved, it must be that conditional on $\mathbf{Z}$, $\mathcal{L}(\mathbf{A}, \beta, \pi)$ is independent of $\mathbf{U}$. Assumption 2.1 could be reformulated as follows.

Assumption 3.1. Let H be a group of permutations. $\mathcal{P}_{\mathbf{U}}$ is H-invariant, i.e., $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}} \Rightarrow \pi P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$ for all $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$ and all $\pi \in H$.

Define

$$C = \left\{ c \in \mathbb{R}^{2n} : \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} c' E[(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{H}' \otimes I_2)|\mathbf{Z}]c < \infty \right\} . \tag{9}$$

It is easy to see C is a linear subspace of $\mathbb{R}^{2n}$, so we define Q as the projection matrix on C. Being a projection matrix, Q is symmetric and idempotent. In fact, C is invariant under H and $\pi Q \pi' = Q$ for $\pi \in H$, by Lemmas B.1–B.2.

Theorem 3.1. *Suppose Assumption 3.1 holds. Then, the solution to*

$$\min_{\mathcal{L}(\mathbf{A},\beta,\pi)} \sup_{(\mu, P_{\mathbf{U}}) \in \mathbb{R}^2 \times \mathcal{P}_{\mathbf{U}}} E_{\mathcal{L}(\mathbf{A},\beta,\pi)}[\|\beta'\pi\mathbf{A} \operatorname{vec}(\mathbf{W}') - \tau'\mu\|^2|\mathbf{Z}] . \tag{10}$$

is the same as the solution to

$$\min_{\mathcal{L}(\mathbf{A},\beta)} \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\beta'\mathbf{A}Q(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{H}' \otimes I_2)Q\mathbf{A}'\beta|\mathbf{Z}] \tag{11}$$

$$\text{subject to} \quad Q\mathbf{A}'\beta = \mathbf{A}'\beta, \tag{12}$$

$$\beta'\mathbf{A}(\mathbf{1}_n \otimes (1,0)') = \tau_0 \tag{13}$$

$$\beta'\mathbf{A}(\mathbf{1}_n \otimes (0,1)') = \tau_1 \ , \tag{14}$$

where $\mathcal{L}(\mathbf{A}, \beta)$ *ranges over the set of all distributions of* $(\mathbf{A}, \beta)$.

Remark 3.1. Theorem 3.1 implies that the optimal randomization scheme is again complete randomization, and one need only find the optimal distribution of $\mathbf{A}$ and β. Note that we do not require ex-ante that the estimator to satisfy Assumption 2.2, i.e., be invariant, nor do we require β to be a fixed function of $\mathbf{A}$. We don't even assume $\mathbf{A}$ itself is fixed. But as we will see below, the optimal estimator often turns out to be invariant, taking the form of $\beta(\mathbf{A})'\mathbf{Y}$, where $\mathbf{A}$ is fixed, and $\beta(\mathbf{A})$ is fixed. ■

The constraints of the optimization problem in Theorem 3.1 are easily interpretable in many settings. From now on, suppose the parameter of interest is the ATE in (7). (13) and (14) imply that β_i's of the treated group sum up to 1, while those of the control group sum up to -1. The following two theorems are implications of Theorem 3.1 in Examples 2.1 and 2.2.

Theorem 3.2. *Under the setting of Example 2.1, suppose*

$$\mathcal{P}_{\mathbf{U}} = \{(\mathbf{m}, \xi) : |m_a(Z_i)| \leq m_a, \sigma_i(a, Z_i) \leq \bar{\sigma}_a \text{ for } a \in \{0,1\}\} \tag{15}$$

for some $K_a, \bar{\sigma}_a \geq 0$. *Suppose* $\tau = (-1,1)'$. *Define* $n_0 = \sum_{1 \leq i \leq n} I\{A_i = (1,0)'\}$ *and* $n_1 = n - n_0$. *Then, the solution to* (10) *is defined by*

$$n_0 = \frac{\bar{\sigma}_0}{\bar{\sigma}_0 + \bar{\sigma}_1} n \ , \tag{16}$$

$\delta(\mathbf{A}, \mathbf{Y}) = \bar{Y}_1 - \bar{Y}_0$ *for* $\bar{Y}_0$ *in* (5) *and similarly for* $\bar{Y}_1$, *and* $\mathbf{G} = \mathbf{H}$.

Remark 3.2. Theorem 3.2 implies that the optimal estimation scheme is defined by complete randomization, where the numbers of treated and untreated units are defined by a Neyman allocation in Neyman (1934), and the ratio depends on the worst-case conditional variances of the error terms. The optimal linear estimator is the difference-in-means estimator. ■

Next, we turn to Example 2.2. Let $n(b)$ be the number of units with $Z_{1i} = b$. For convenience, we reorder the units so that $Z_{1i} = 1$ for $1 \leq i \leq n_1$ and $Z_{1i} = b$ for $n(b-1)+1 \leq i \leq n(b)$, $1 \leq b \leq B$. With some abuse of notation, we write $i \in b$ if i is in the bth block.

Theorem 3.3. *Under the setting of Example 2.2, suppose*

$$\mathcal{P}_{\mathbf{U}} = \{(\gamma, \tilde{\mathbf{m}}, \xi) : \gamma \in \mathbb{R}^B, |\tilde{m}_a(Z_{2i})| \leq K_a(Z_{1i}), \sigma_i(a, Z_i) \leq \bar{\sigma}_a(Z_{1i}) \text{ for } a \in \{0,1\}\} \tag{17}$$

for some $K_a(b), \sigma_a(b) \geq 0$ *for* $1 \leq b \leq B$. *Suppose* $\tau = (-1,1)'$. *Define* $n_0(b) = \sum_{1 \leq i \leq n} I\{A_i = (1,0)', i \in b\}$ *and* $n_1(b) = n(b) - n_1(b)$. *Define*

$$\bar{Y}_{0,b} = \frac{1}{n_0(b)} \sum_{1 \leq i \leq n} Y_i I\{A_i = (1,0)', i \in b\}$$

and $\bar{Y}_{1,b}$ *similarly. Then, the solution to* (10) *is defined by*

$$n_0(b) = \frac{\bar{\sigma}_0(b)}{\bar{\sigma}_0(b) + \bar{\sigma}_1(b)} n(b) \ ,$$

the estimator

$$\delta(\mathbf{A}, \mathbf{Y}) = \sum_{1 \leq b \leq B} \tau^*(b)(\bar{Y}_{1,b} - \bar{Y}_{0,b}) \ , \tag{18}$$

where

$$\tau^*(b) = \frac{1/s(b)}{\sum_{1 \leq b \leq B} 1/s(b)} \tag{19}$$

for $s(b) = (\bar{\sigma}_0(b) + \bar{\sigma}_1(b))^2/n(b) + (K_0(b) + K_1(b))^2$, *and* $\mathbf{G} = \mathbf{H}$.

Remark 3.3. Theorem 3.3 implies that the optimal estimation scheme is defined by complete randomization within each block, where the numbers of treated and untreated units are again defined by Neyman allocations in each block. The optimal linear estimator is a weighted average of the difference in means within each block, where the weight is a function of the worst-case conditional variances and unit-specific effects in each block. ■

4 Conclusion

This paper shows the minimax optimality of the complete randomization scheme under permutation invariance. I also characterize the optimal estimation scheme as jointly complete randomization, Neyman allocations, and difference-in-means type estimators. All these results hold in finite sample and under heterogeneous treatment effecs.

It is natural to ask what happens if the invariance assumptions hold only approximately. For instance, the data might not display the exact block structure as in Example 2.2, although there are some discrete variables. Appendix E of Bai (2020) studies the optimal stratified randomization when the designer does not have access to pilot data, where he recommends that they solve a minimax problem in a similar spirit to the one studied here, where the worst case is over a bounded polyhedron. He also provides fast algorithms to search over a restricted set of randomization schemes, the set of all matched-pair designs. The results there, however, take as given the difference-in-means estimator. The problem of jointly determining the optimal estimator and the randomization scheme is left for future work.

Appendix A Proofs of Main Results

A.1 Proof of Theorem 2.1

It is straight forward to see that we need only prove that for any $s_0 \in S$,

$$\min_{\mathbf{G}\in\Phi} \max_{\pi\in H} f(r(\mathbf{G}, \mathbf{A}, \delta; \pi s_0)) = \max_{\pi\in H} f(r(\mathbf{H}, \mathbf{A}, \delta; \pi s_0)) \ , \tag{20}$$

i.e. we need only focus on the orbit of s_0. Note that the orbit could be defined because of Assumption 2.1. Next, for any $\tilde{\pi} \in H$,

$$\begin{aligned} r(\tilde{\pi}, \mathbf{A}, \delta; \pi s_0) &= E_{\pi s_0}[L(\pi s_0, \delta(\tilde{\pi}\mathbf{A}, \tilde{\pi}\mathbf{A}\,\mathrm{vec}(\mathbf{W}')))|\mathbf{Z}] = E_{\pi s_0}[L(s_0, \delta(\tilde{\pi}\mathbf{A}, \tilde{\pi}\mathbf{A}\,\mathrm{vec}(\mathbf{W}')))|\mathbf{Z}] \\ &= E_{s_0}[L(s_0, \delta(\tilde{\pi}\mathbf{A}, \tilde{\pi}\mathbf{A}\,\mathrm{vec}((\pi\mathbf{W})')))|\mathbf{Z}] = r(\pi^{-1}\tilde{\pi}, \mathbf{A}, \delta; s_0) \ , \end{aligned}$$

where the first equality holds by definition, the second equality holds by Assumption 2.3, the third equality holds by (4), and the last follows from Assumption 2.2 which implies

$$\delta(\tilde{\pi}\mathbf{A}, \tilde{\pi}\mathbf{A}\,\mathrm{vec}((\pi\mathbf{W})')) = \delta(\pi^{-1}\tilde{\pi}\mathbf{A}, \pi^{-1}\tilde{\pi}\mathbf{A}\,\mathrm{vec}(\mathbf{W}')) \ .$$

Therefore, for any $\pi \in H$,

$$r(\mathbf{G}, \mathbf{A}, \delta; \pi s_0) \stackrel{d}{=} r(\pi^{-1}\mathbf{G}, \mathbf{A}, \delta; s_0) \tag{21}$$

and

$$r(\mathbf{H}, \mathbf{A}, \delta; \pi s_0) \stackrel{d}{=} r(\pi^{-1}\mathbf{H}, \mathbf{A}, \delta; s_0) \stackrel{d}{=} r(\mathbf{H}, \mathbf{A}, \delta; s_0) \ , \tag{22}$$

since $\mathbf{H}$ is the uniform distribution on H. Note also that

$$\frac{1}{|H|}\sum_{\pi\in H} r(\pi^{-1}\mathbf{G}, \mathbf{A}, \delta; s_0) \stackrel{d}{=} r(\mathbf{H}\mathbf{G}, \mathbf{A}, \delta; s_0) \stackrel{d}{=} r(\mathbf{H}, \mathbf{A}, \delta; s_0) \ , \tag{23}$$

again because $\mathbf{H}$ is the uniform distribution on H. Combining (21), (22) and (23), we have that for any $\tilde{\pi} \in H$,

$$\begin{aligned} &\max_{\pi\in H} f(r(\mathbf{G}, \mathbf{A}, \delta; \pi s_0)) \\ &\geq f\left(\frac{1}{|H|}\sum_{\pi\in H} r(\mathbf{G}, \mathbf{A}, \delta; \pi s_0)\right) \\ &= f(r(\mathbf{H}, \mathbf{A}, \delta; s_0)) \\ &= f(r(\mathbf{H}, \mathbf{A}, \delta; \tilde{\pi} s_0)) \end{aligned} \tag{24}$$

$$= \max_{\pi \in H} f(r(\mathbf{H}, \mathbf{A}, \delta; \pi s_0))$$

for any $\tilde{\pi} \in H$, where the inequality holds by Assumption 2.4, the first equality holds by (21) and (23), the second equality holds by (22) and the last equality holds because (24) holds for each $\tilde{\pi}$ and so maximum over $\tilde{\pi}$ is equal to any one of them. ■

A.2 Proof of Theorem 3.1

Since the distribution $\mathcal{L}(\mathbf{A}, \beta, \pi)$ is completely unrestricted, we could equivalently permute $\mathbf{W}$ instead of $\mathbf{A}$, i.e., (10) is equivalent to

$$\min_{\mathcal{L}(\mathbf{A},\beta,\pi)} \sup_{(\mu, P_{\mathbf{U}}) \in \mathbb{R}^2 \times \mathcal{P}_{\mathbf{U}}} E_{\mathcal{L}(\mathbf{A},\beta,\pi)}[\|\beta' \mathbf{A} \operatorname{vec}((\pi \mathbf{W})') - \tau' \mu\|^2 | \mathbf{Z}] \ . \tag{25}$$

From now on we omit the dependence of the expectation on $\mathcal{L}(\mathbf{A}, \beta, \pi)$. By (4),

$$E[\|\beta' \mathbf{A} \operatorname{vec}((\pi \mathbf{W})') - \tau' \mu\|^2 | \mathbf{Z}] = E[\|(\beta' \mathbf{A}(\mathbf{1}_n \otimes \mu) - \tau' \mu) + \beta' \mathbf{A} \operatorname{vec}(\mathbf{U}' \pi')\|^2 | \mathbf{Z}] \ .$$

Since the maximum is over $\mu \in \mathbb{R}^2$, we get unbounded maximum risk unless $\beta' \mathbf{A}(\mathbf{1}_n \otimes \mu) = \tau' \mu$ for all $\mu \in \mathbb{R}^2$. This is equivalent to the two conditions that

$$\begin{aligned} \beta' \mathbf{A}(\mathbf{1}_n \otimes (1, 0)') &= \tau_0 \\ \beta' \mathbf{A}(\mathbf{1}_n \otimes (0, 1)') &= \tau_1 \ . \end{aligned} \tag{26}$$

Given that (26) holds, we have

$$\begin{aligned} & E[\|\beta' \mathbf{A} \operatorname{vec}((\pi \mathbf{W})') - \tau' \mu\|^2 | \mathbf{Z}] \\ &= E[\|\beta' \mathbf{A} \operatorname{vec}(\mathbf{U}' \pi')\|^2 | \mathbf{Z}] \\ &= E[\|\beta' \mathbf{A}(\pi \otimes I_2) \operatorname{vec}(\mathbf{U}')\|^2 | \mathbf{Z}] \\ &= E[\beta' \mathbf{A}(\pi \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\pi' \otimes I_2) \mathbf{A}' \beta | \mathbf{Z}] \ . \end{aligned}$$

Since $P_{\mathbf{U}}$ is H-invariant by Assumption 3.1, we have for all $\tilde{\pi} \in H$ that

$$\begin{aligned} & \sup_{(\mu, P_{\mathbf{U}}) \in (\mathbb{R}^2 \times \mathcal{P}_{\mathbf{U}})} E[\|\beta' \mathbf{A} \operatorname{vec}((\pi \mathbf{W})') - \tau' \mu\|^2 | \mathbf{Z}] \\ &= \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\beta' \mathbf{A}(\pi \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\pi' \otimes I_2) \mathbf{A}' \beta | \mathbf{Z}] \\ &= \sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[\beta' \mathbf{A}(\tilde{\pi} \pi \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\pi' \tilde{\pi}' \otimes I_2) \mathbf{A}' \beta | \mathbf{Z}] \ . \end{aligned}$$

Since the equality holds for all of $\tilde{\pi} \in H$, it also holds for the average over H. Hence the last expression is equal to

$$\begin{aligned}
&\frac{1}{|H|}\sum_{\tilde{\pi}\in H}\sup_{P_{\mathbf{U}}\in\mathcal{P}_{\mathbf{U}}} E[\beta'\mathbf{A}(\tilde{\pi}\pi\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\pi'\tilde{\pi}'\otimes I_2)\mathbf{A}'\beta|\mathbf{Z}] \\
&\geq \sup_{P_{\mathbf{U}}\in\mathcal{P}_{\mathbf{U}}}\frac{1}{|H|}\sum_{\tilde{\pi}\in H} E[\beta'\mathbf{A}(\tilde{\pi}\pi\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\pi'\tilde{\pi}'\otimes I_2)\mathbf{A}'\beta|\mathbf{Z}] \\
&= \sup_{P_{\mathbf{U}}\in\mathcal{P}_{\mathbf{U}}} E[\beta'\mathbf{A}(\mathbf{H}\pi\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\pi'\mathbf{H}'\otimes I_2)\mathbf{A}'\beta|\mathbf{Z}] \\
&= \sup_{P_{\mathbf{U}}\in\mathcal{P}_{\mathbf{U}}} E[\beta'\mathbf{A}(\mathbf{H}\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\mathbf{H}'\otimes I_2)\mathbf{A}'\beta] \\
&= \sup_{P_{\mathbf{U}}\in\mathcal{P}_{\mathbf{U}}} E[\|\beta'\mathbf{A}\operatorname{vec}(\mathbf{U}'\mathbf{H}')\|^2|\mathbf{Z}] \qquad (27)
\end{aligned}$$

where the second equality holds because $\mathbf{H}$ is the uniform distribution on H so that $\mathbf{H}\pi \overset{d}{=} \mathbf{H}$. I have shown that for each $\mathcal{L}(\mathbf{A},\beta)$, the maximum risk over $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$ is minimized by letting π to be distributed as $\mathbf{H}$ conditional on $(\mathbf{A},\beta)$. Therefore, the solution to (10) must have π distributed as $\mathbf{H}$ conditional on $(\mathbf{A},\beta)$, so $\mathbf{H}$ and $\mathcal{L}(\mathbf{A},\beta)$ are independent. To conclude the proof, note that we must have $\mathbf{A}'\beta \in C$ with probability one under $\mathcal{L}(\mathbf{A},\beta)$ for C in (9), because otherwise (27) is infinite. (27) then becomes

$$\sup_{P_{\mathbf{U}}\in\mathcal{P}_{\mathbf{U}}} E[\beta'\mathbf{A}Q(\mathbf{H}\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\mathbf{H}'\otimes I_2)Q\mathbf{A}'\beta|\mathbf{Z}]\ ,$$

and the result follows. ■

A.3 Proof of Theorem 3.2

To begin with, note that $\mathcal{P}_{\mathbf{U}}$ in (15) clearly satisfies Assumption 3.1. The matrix

$$E[(\mathbf{H}\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\mathbf{H}'\otimes I_2)|\mathbf{Z}]$$

has

$$\frac{1}{n}\sum_{1\leq i\leq n} E[U_iU_i'|Z_i]$$

as the diagonal blocks, and

$$\frac{1}{n(n-1)}\sum_{i\neq j} E[U_iU_j'|Z_i,Z_j]$$

as the off-diagonal blocks. By (15), we immediately see $C = \mathbb{R}^{2n}$ so $Q = I_{2n}$. So we have

$$
\begin{aligned}
&\beta'\mathbf{A}Q(\mathbf{H}\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\mathbf{H}'\otimes I_2)Q\mathbf{A}'\beta \qquad (28)\\
&= \sum_{A_i=(0,1)'}\beta_i^2\frac{1}{n}\sum_{1\le i\le n}U_i^2(0) + \sum_{A_i=(1,0)'}\beta_i^2\frac{1}{n}\sum_{1\le i\le n}U_i^2(1)\\
&\quad + \sum_{i\neq j:A_i=A_j=(1,0)'}\beta_i\beta_j\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}U_i(0)U_j(0)\\
&\quad + \sum_{i\neq j:A_i=A_j=(0,1)'}\beta_i\beta_j\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}U_i(1)U_j(1)\\
&\quad + 2\sum_{A_i=(0,1)',A_j=(1,0)'}\beta_i\beta_j\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}U_i(1)U_j(0)\ .
\end{aligned}
$$

By (13) and (14),

$$
\begin{aligned}
\sum_{A_i=(1,0)',A_j=(0,1)'}\beta_i\beta_j &= \Big(\sum_{A_i=(1,0)'}\beta_i\Big)\Big(\sum_{A_j=(0,1)'}\beta_j\Big) = \tau_0\tau_1\\
\sum_{i\neq j:A_i=A_j=(1,0)'}\beta_i\beta_j &= \Big(\sum_{A_i=(1,0)}\beta_i\Big)^2 - \sum_{A_i=(1,0)'}\beta_i^2 = \tau_0^2 - \sum_{A_i=(1,0)'}\beta_i^2\\
\sum_{i\neq j:A_i=A_j=(0,1)'}\beta_i\beta_j &= \Big(\sum_{A_i=(0,1)}\beta_i\Big)^2 - \sum_{A_i=(0,1)'}\beta_i^2 = \tau_1^2 - \sum_{A_i=(0,1)'}\beta_i^2\ ,
\end{aligned}
$$

so (28) equals

$$
\begin{aligned}
M_n(\mathbf{A},\beta) &= \sum_{A_i=(1,0)'}\beta_i^2\Big(\frac{1}{n}\sum_{1\le i\le n}U_i^2(0) - \frac{1}{n(n-1)}\sum_{i\neq j}U_i(0)U_j(0)\Big)\\
&\quad + \sum_{A_i=(0,1)'}\beta_i^2\Big(\frac{1}{n}\sum_{1\le i\le n}U_i^2(1) - \frac{1}{n(n-1)}\sum_{i\neq j}U_i(1)U_j(1)\Big)\\
&\quad + \tau_0^2\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}U_i(0)U_j(0)\\
&\quad + \tau_1^2\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}U_i(1)U_j(1)\\
&\quad + 2\tau_0\tau_1\frac{1}{n(n-1)}\sum_{1\le i\neq j\le n}U_i(1)U_j(0)\ .
\end{aligned}
$$

Recall $\tau = (-1,1)'$. For any fixed $(\mathbf{A},\beta)$ so that $\sum_{A_i=(1,0)'}\beta_i^2 \le 1$, $\sum_{A_i=(0,1)'}\beta_i^2 \le 1$, it is straightforward to show that $E[M_n(\mathbf{A},\beta)|\mathbf{Z}]$ attains its maximum over $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$ at $m_1(Z_i) = K_1$, $m_0(Z_i) = -K_0$, and $\sigma_i^2(a,Z_i) = \bar{\sigma}_a^2$. If $\sum_{A_i=(1,0)'}\beta_i^2 > 1$ or $\sum_{A_i=(0,1)'}\beta_i^2 > 1$, then the maximum of $E[M_n(\mathbf{A},\beta)|\mathbf{Z}]$ over $P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}$ is at least equal to its value at $m_1(Z_i) = K_1$, $m_0(Z_i) = -K_0$, and $\sigma_i^2(a,Z_i) = \bar{\sigma}_a^2$, which is easily seen to be larger than when $\sum_{A_i=(1,0)'}\beta_i^2 \le 1$ and $\sum_{A_i=(0,1)'}\beta_i^2 \le 1$. As a result, in the solution to (10),

$\sum_{A_i=(1,0)'} \beta_i^2 \leq 1$ and $\sum_{A_i=(0,1)'} \beta_i^2 \leq 1$ with probability one. So

$$\sup_{P_{\mathbf{U}} \in \mathcal{P}_{\mathbf{U}}} E[M_n(\mathbf{A}, \beta)|\mathbf{Z}, \mathbf{A}, \beta] = \sum_{A_i=(1,0)'} \beta_i^2 \bar{\sigma}_0^2 + \sum_{A_i=(0,1)'} \beta_i^2 \bar{\sigma}_1^2 + (K_0 + K_1)^2 \ . \tag{29}$$

(13) and (14) imply $\sum_{A_i=(1,0)'} \beta_i = -1$ and $\sum_{A_i=(0,1)'} \beta_i = 1$. Define n_0 and n_1 as the number of treated and control units. For any fixed $\mathbf{A}$, (11) is solved by $\beta_i = -1/n_0$ if $A_i = (1,0)'$ and $\beta_i = 1/n_1$ if $A_i = (0,1)'$. Minimizing (29) is equivalent to minimizing

$$\frac{\bar{\sigma}_0^2}{n_0} + \frac{\bar{\sigma}_1^2}{n_1} \ . \tag{30}$$

and (16) follows. ■

A.4 Proof of Theorem 3.3

The proof follows similar arguments as those in the proof of Theorem 3.2. N ote that $\mathcal{P}_{\mathbf{U}}$ in (17) clearly satisfies Assumption 3.1. We start with $E[(\mathbf{H} \otimes I_2) \operatorname{vec}(\mathbf{U}') \operatorname{vec}(\mathbf{U}')'(\mathbf{H}' \otimes I_2)|\mathbf{Z}]$. Note that each element in the group H only permutes units within each block. The diagonal block of the matrix is in fact the one in Example 2.1, but with $n = n(b)$. The off-diagonal block at position (b_1, b_2) equals

$$\frac{1}{n_{b_1} n_{b_2}} \sum_{i \in b_1, j \in b_2} U_i U_j' \ .$$

Recall $U_i(a) = \gamma_b(a) + \tilde{m}_a(Z_{2i}) + \epsilon_i(a)$. This time, we choose. Since γ is allowed to be arbitrarily large, we could verify that C in (9) is given by

$$\left\{ c \in \mathbb{R}^{2n} : \sum_{2n(b-1)+1 \leq j \leq 2n(b)} c_j = 0 \right\} \ .$$

As a result, (12) implies that $\sum_{i \in b} \beta_i = 0$ for $1 \leq b \leq B$. The rest of the problem could be solved similarly as in Example 2.1. Indeed, within each block $b \in B$, $\sum_{i \in b} \beta_i = 0$, so we define $\sum_{i \in b: A_i=(1,0)'} \beta_i = \tau(b)$ and $\sum_{i \in b: A_i=(0,1)'} \beta_i = -\tau(b)$. This forces $\tau_1 = -\tau_0$. Recall $\tau = (-1, 1)'$. Given $\tau(b)$, similar arguments as those in the proof of Theorem 3.2 imply

$$n_0(b) = \frac{\bar{\sigma}_0(b)}{\bar{\sigma}_0(b) + \bar{\sigma}_1(b)} n(b) \ ,$$

and $\beta_i = \tau(b)/n_0(b)$ if $A_i = (1,0)'$ and $-\tau(b)/n_1(b)$ otherwise. Finally, we choose $\tau(b)$ to minimize the sum of the counterpart of (29) across $1 \le b \le B$,

$$\sum_{1\le b\le B} \frac{\tau(b)^2}{n(b)}(\bar\sigma_0(b)+\bar\sigma_1(b))^2 + \tau(b)^2(K_0(b)+K_1(b))^2 \ , \tag{31}$$

under the contraint in (13) and (14), both of which are

$$\sum_{b\in B}\tau(b) = 1 \ .$$

The conclude the proof, note it is straightforward to show the minimum of (31) occurs at $\tau^*(b)$ in (19). ■

Appendix B Auxiliary Results

Lemma B.1. *C is H-invariant, i.e., $c \in C \Rightarrow \pi c \in C$ for all $c \in \mathbb{R}^{2n}$ and $\pi \in H$.*

Proof of Lemma B.1. Suppose

$$\sup_{P_{\mathbf{U}}\in\mathcal{P}_{\mathbf{U}}} c'E[(\mathbf{H}\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\mathbf{H}'\otimes I_2)|\mathbf{Z}]c < \infty$$

one need to prove that for any $\pi \in H$,

$$\sup_{P_{\mathbf{U}}\in\mathcal{P}_{\mathbf{U}}} c'\pi'E[(\mathbf{H}\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\mathbf{H}'\otimes I_2)|\mathbf{Z}]\pi'c < \infty \tag{32}$$

However, since $\mathcal{P}_{\mathbf{U}}$ is H-invariant, (32) holds for $\mathbf{U}$ replaced by $\pi\mathbf{U}$ for any $\pi \in H$. Note that

$$\begin{aligned}
&E[(\mathbf{H}\otimes I_2)\operatorname{vec}(\mathbf{U}'\pi')\operatorname{vec}(\mathbf{U}'\pi')'(\mathbf{H}'\otimes I_2)|\mathbf{Z}] \\
&= E_{\tilde\pi\sim\mathbf{H}}[(\tilde\pi\otimes I_2)(\pi\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\pi'\otimes I_2)(\tilde\pi'\otimes I_2)|\mathbf{Z}] \\
&= E_{\tilde\pi\sim\mathbf{H}}[(\tilde\pi\pi\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\pi'\tilde\pi'\otimes I_2)|\mathbf{Z}] \\
&= E_{\tilde\pi\sim\mathbf{H}}[(\tilde\pi\otimes I_2)\operatorname{vec}(\mathbf{U}')\operatorname{vec}(\mathbf{U}')'(\tilde\pi'\otimes I_2)|\mathbf{Z}] \ ,
\end{aligned}$$

where the first equality holds because $\operatorname{vec}(ABC) = (C'\otimes A)\operatorname{vec}(B)$, the second equality holds since $(A\otimes B)(C\otimes D) = AB\otimes CD$, and the last holds since $\mathbf{H}$ is the uniform distribution on H. As a result, (32) holds. ■

Recall that a permutation matrix is orthogonal, i.e. $\pi'\pi = \pi\pi' = I$.

Lemma B.2. *$Q\pi = \pi Q$ or equivalently since π is orthogonal, $\pi Q\pi' = Q$.*

Proof of Lemma B.2. For $c \in C$, $\pi Q\pi^{-1}c = \pi\pi^{-1}c = c$ since C is H-invariant. For an element $b \in C^{\perp}$, $\pi Qb = 0$ since $Qb = 0$. Now, by definition of $C^{\perp}$, $b'c = 0$ for all $c \in C$, and therefore $b'\pi'c = 0$ since $\pi'c \in C$ for all $c \in C$. As a result, $\pi b \in C^{\perp}$ as well, and hence $Q\pi b = 0$. As a result, $\pi Q\pi^{-1}c = Qc$ for all $c \in \mathbb{R}^n$ and hence $\pi Q\pi^{-1} = Q$. ■

References

Armstrong, T. B. and Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, **86** 655–683.

Artin, M. (2011). *Algebra*. Pearson Prentice Hall.

Bai, Y. (2020). Optimality of matched-pair designs in randomized controlled trials. Working paper.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, **1** 200–232. Publisher: American Economic Association.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, **22** 238–270.

Hooper, P. M. (1989). Minimaxity of randomized optimal designs. *The Annals of Statistics*, **17** 1315–1324.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, **24** 324–338.

Li, K.-C. (1983). Minimaxity for randomized designs: Some general results. *The Annals of Statistics*, **11** 225–239.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, **97** 558–625.

Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in clinical trials: Theory and Practice*. John Wiley & Sons.

Wu, C.-F. (1981). On the robustness and efficiency of some randomized designs. *The Annals of Statistics*, **9** 1168–1177.