Paper review: Covariate-adaptive randomization inference in matched designs

Samuel D. Pimentel and Yaxuan Huang (2024)

Suehyun Kim

20 January 2025

Causal Inference Lab.
Seoul National University

Table of contents

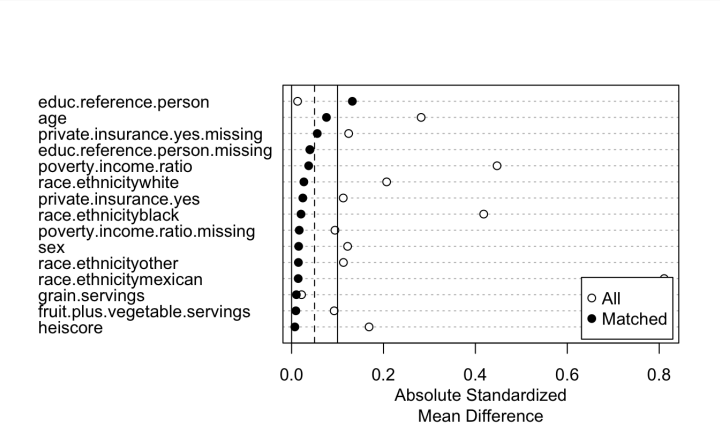
Table of contents

Introduction

- Matched observational studies adjust for observed confounding by grouping each treated unit with similar control units.
- With exact agreement on propensity scores in each matched set, a matched observational study can be treated as a stratified randomized experiment.
- In reality, propensity scores are rarely matched exactly, except in cases where the measured variables are few and discrete. However, uniform randomization inference is still popularly conducted despite such discrepancies.

Introduction

Example. Optimal pair matching with PS calipers (783 treated units)



Match	PS (Tr)	PS (Ctrl)	Abs. Diff.
435	0.579	0.438	0.141
531	0.553	0.436	0.117
695	0.589	0.472	0.117
522	0.634	0.529	0.104
543	0.542	0.441	0.101
503	0.562	0.465	0.098

Would uniform randomization inference remain valid in such cases?

Introduction

This paper addresses the problem via **covariate-adaptive randomization inference** in matched observational studies, which leverages estimated propensity scores to update non-uniform permutation probabilities.

- It retains many advantages of uniform randomization inference, including ease of implementation and compatibility with estimation procedures as well as sensitivity analyses.
- It tends to restore approximate control over Type I error.
- The combination of matched design and covariate-adaptive randomization improves precision while remaining robust to propensity score misspecification.

Table of contents

Notation

Consider a population of individuals $(Y(1), Y(0), Z, X, U)$, under the Neyman-Rubin causal model and the stable unit treatment value assumption (SUTVA).

- Z: Binary indicator for treatment
- $Y(z)$: Potential outcomes
- Y: Observed outcome
- X: Vector of observed covariates
- U: Unobserved covariate
- $\lambda(x) = P(Z \mid X = x)$: Propensity score
- $\pi(x, u) = P(Z \mid X = x, U = u)$: True probability of treatment

Notation

We assume individuals are first sampled independently from the population, and then formed into a **matched design** $\mathcal{M}$ based on observed covariates X, consisting of K matched sets.

- Each matched set ($k = 1, \ldots, K$) contains exactly one treated individual and one or more control individuals.
- Individuals in set k are numbered $k1$ through kn_k.
  - n_k: Number of units in set k
  - $n = \sum_{k=1}^{K} n_k$
  - $\mathbf{Y}(1), \mathbf{Y}(0), \mathbf{Y}, \mathbf{Z}, \mathbf{U} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$: Data in contiguous order
- $\mathcal{Z}_{\mathcal{M}}$: Set of treatment vectors $\mathbf{Z}'$ such that $\sum_{i=1}^{n_k} Z'_{ki} = 1$ for all k.
  - i.e., all possible within-match permuations for a given matched design $\mathcal{M}$.

Uniform randomization inference in matched designs

A matched design is **exact** on a variable or a quantity $\mathbf{v} \in \mathbb{R}^n$ if for any matched set k, $v_{ki} = v_{kj}$ for all $i, j \in \{1, \ldots, n_k\}$.

Consider the distribution of treatment assignments conditional on the matched design, $P(\mathbf{Z} \mid \mathcal{Z}_\mathcal{M}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0))$. This distribution is discrete uniform under two key assumptions:

(i) *No unobserved confounding*, i.e. $\lambda(x) = \pi(x, u)$,

(ii) *Exact matching* on covariates $\mathbf{X}$, or more generally, on true propensity scores $\lambda(\mathbf{X})$.

$\Longrightarrow$ **Uniform randomization inference** with $P(\mathbf{Z} = \mathbf{z} \mid \mathcal{Z}_\mathcal{M}) = 1/|\mathcal{Z}_\mathcal{M}|$

Uniform randomization inference in matched designs

Consider the **sharp null hypothesis** of no effect:

$$H_0 : \mathbf{Y}(1) = \mathbf{Y}(0). \tag{1}$$

- The test statistic $T(Z, Y)$ varies only through $\mathbf{Z}$, which is uniformly distributed over all permutations in $\mathcal{Z}_{\mathcal{M}}$ under H_0.
- We use this null distribution to compute Fisher's exact p-value:
  - Monte Carlo simulations
  - Normal approximation
- Estimates and confidence intervals for a homogeneous additive effect can also be obtained:
  - Hodges-Lehmann point estimates and the corresponding CIs
  - Normal approximation

PS discrepancies disrupting uniform inference

In practice, exact matching is almost never possible. The implications of the lack of fit between the nominal uniform distribution and the true distribution include:

- Slow-shrinking finite sample bias of the difference-in-means estimator (Hansen, 2009),
- Bias and inconsistency in treatment effect estimation (Sävje, 2022),
- Failure of Type I error control for the uniform randomized test, even in infinite samples (Guo & Rothenhäusler, 2023).

PS discrepancies disrupting uniform inference

Some existing remedies are *(i) performing balance tests* as tests for lack of fit and *(ii) using regression adjustments* to remove bias not addressed by close matching.

- The balance test approach has two major shortcomings:
  - It relies on an asymptotic regime in which PS differences within matched sets approach zero as sample size increases, which may not be reasonable in many practical settings.
  - It does not provide a solution in case of balance failure.

- Regression adjustments also rely on strong assumptions on the outcome model and still require the PS discrepancies to shrink to zero within matched sets.

Table of contents

Covariate-adaptive randomization inference

Uniform inference with exact matching can be summarized as follows:

Free of hidden bias + *Exact matching* on $\lambda(\mathbf{X})$

$\Longrightarrow$ Valid uniform inference with $P(\mathbf{Z} = \mathbf{z} \mid \mathcal{Z}_{\mathcal{M}}) = 1/|\mathcal{Z}_{\mathcal{M}}|$.

Covariate-adaptive randomization inference

Uniform inference with exact matching can be summarized as follows:

> Free of hidden bias + *Exact matching* on $\lambda(\mathbf{X})$
> $\Longrightarrow$ Valid uniform inference with $P(\mathbf{Z} = \mathbf{z} \mid \mathcal{Z}_{\mathcal{M}}) = 1/|\mathcal{Z}_{\mathcal{M}}|$.

Covariate-adaptive randomization inference seeks for an analogue by correcting for the actual conditional distribution of treatment:

> Free of hidden bias + *Inexact matching* on $\lambda(\mathbf{X})$
> $\Longrightarrow$ Valid covariate-adaptive inference with *true* $P(\mathbf{Z} = \mathbf{z} \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X})$.

Conditional distribution of treatment status

First, we compute the true conditional distribution of treatment status $P(\mathbf{Z} = \mathbf{z} \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X})$. We assume the absence of unobserved confounding, and express the distribution in terms of true propensity scores.

Since $Z_{ki} \sim \text{Bernoulli}(\lambda(X_{ki}))$, we obtain

$$
\begin{aligned}
&P(Z_{ki} = 1 \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}) \\
&= \frac{P(Z_{k1} = 0, \ldots, Z_{ki} = 1, \ldots, Z_{kn_k} = 0 \mid \lambda(\mathbf{X}_k))}{\sum_{j=1}^{n_k} P(Z_{k1} = 0, \ldots, Z_{kj} = 1, \ldots, Z_{kn_k} = 0 \mid \lambda(\mathbf{X}_k))} \\
&= \frac{\lambda(X_{ki}) \prod_{j \neq i} (1 - \lambda(X_{kj}))}{\sum_{j=1}^{n_k} \lambda(X_{kj}) \prod_{l \neq j} (1 - \lambda(X_{kl}))} = \frac{\text{odds}\{\lambda(X_{ki})\}}{\sum_{j=1}^{n_k} \text{odds}\{\lambda(X_{kj})\}}
\end{aligned}
$$

Denote the scaled PS odds $\frac{\text{odds}\{\lambda(X_{ki})\}}{\sum_{j=1}^{n_k} \text{odds}\{\lambda(X_{kj})\}}$ by p_{ki}.

Conditional distribution of treatment status

Since individuals are sampled independently, the joint conditional probability for a treatment vector $\mathbf{Z}$ is given as follows:

$$P(\mathbf{Z} = \mathbf{z} \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}) = \begin{cases} \prod_{k=1}^{K} \prod_{i=1}^{n_k} p_{ki}^{z_{ki}} & \text{if } \mathbf{z} \in \mathcal{Z}_{\mathcal{M}} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

- In practice, we derive an empirical distribution by substituting an estimate of the propensity score $\hat{\lambda}(\cdot)$ for $\lambda(\cdot)$ in the formula for p_{ki}.
- Note that the distribution above is uniform under exact matching, where $p_{ki} = 1/n_k$ for all k, i.

Conditional distribution of treatment status

Consequently, an investigator may conduct covariate-adaptive randomization inference via a *Monte Carlo approach* by following the steps below:

(i) Compute the estimated PS odds $\hat{p}_{ki}$ for each subject.

(ii) Draw a treated unit from each matched set k from an *independent multinomial distribution* with 1 trial and probabilities $\hat{p}_{ki}$.

(iii) Compute the test statistic $T(\mathbf{Z}, \mathbf{Y})$ for each draw, and repeat the process to approximate the null distribution (Monte Carlo method).

The resulting null distribution can be used to compute exact p-values for the sharp null hypothesis and to construct CIs by inverting the test.

Large-sample distribution of the test statistic

One can also derive the **large-sample distribution of the test statistic** under the null hypothesis. We demonstrate a normal approximation of the standard *difference-in-means* estimator,

$$
\begin{aligned}
T(\mathbf{Z}, \mathbf{Y}) &= \frac{1}{K} \sum_{k=1}^{K} \left\{ \sum_{i=1}^{n_k} Z_{ki} Y_{ki} - \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (1 - Z_{ki}) Y_{ki} \right\} \\
&= \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k Z_{ki} - 1}{n_k - 1}.
\end{aligned}
$$

Large-sample distribution of the test statistic

Since the outcome $\mathbf{Y}$ is fixed under the sharp null hypothesis, the expectation and variance of $T(\mathbf{Z}, \mathbf{Y})$ are calculated only over the distribution of $\mathbf{Z}$:

$$\mathbb{E}[T(\mathbf{Z}, \mathbf{Y}) \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n_k} Y_{ki} \frac{n_k p_{ki} - 1}{n_k - 1},$$

$$\begin{aligned} &\mathrm{Var}[T(\mathbf{Z}, \mathbf{Y}) \mid \mathcal{Z}_{\mathcal{M}}, \mathbf{X}, \mathbf{Y}(1), \mathbf{Y}(0)] \\ &\quad = \frac{1}{K^2} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left(\frac{n_k}{n_k - 1} \right)^2 Y_{ki} p_{ki} \left\{ Y_{ki}(1 - p_{ki}) - \sum_{j \neq i}^{n_k} Y_{kj} p_{kj} \right\}. \end{aligned} \quad (3)$$

Large-sample distribution of the test statistic

Remarks.

(i) The simplest available CLT is with respect to the inner sums $\sum_{i=1}^{n_k} Y_{ki} \frac{n_k Z_{ki} - 1}{n_k - 1}$, which are independent random variables with nonidentical distributions.

- A finite-sample CLT can be applied when n_k is uniformly bounded, and a Lindeberg condition holds on the potential outcomes $Y(0)$.

(ii) The M-estimation framework can be used to account for the variation that arises from the estimation of the propensity score; however the new variance estimates are near-identical or smaller.

(iii) The difference-in-means estimator is no longer an unbiased estimator for a homogeneous treatment effect.

- The IPW estimator, which can also be used as a test statistic, remains unbiased under covariate-adaptive randomization.
- However, it exhibits greater variance.

Estimates and CIs for an additive effect

Suppose we aim to estimate a **constant additive treatment effect** τ such that

$$Y_{ki}(1) = Y_{ki}(0) + \tau \quad \text{for all } k, i. \tag{4}$$

and seek a point estimate $\hat{\tau}$ and a confidence interval for τ.

- The authors recommend using the *maximum p-value estimator*, which is defined as the value of τ_0 with the largest two-sided p-value.
  - It enjoys many of the attractive properties of the traditional Hodges-Lehmann estimator, including median unbiasedness and asymptotic normality.
  - It performs substantially better than the difference-in-means estimator or the IPW estimator.

Estimates and CIs for an additive effect

In order to obtain confidence intervals for τ, one can:

- Invert the test based on Monte Carlo draws of the difference-in-means statistic.
    - May be computationally intensive, although shortcuts exist.
- Use the large-sample normal approximation to the null distribution of the difference-in-means estimator to invert the test.

Table of contents

Impact of PS estimation error

Recall that the true conditional probabilities of treatment are based on *true propensity scores*.

Since we only have access to *estimated propensity scores*, we need to guarantee that the estimation procedure results in only small deviations in nominal and actual test size and CI coverage.

Impact of PS estimation error

Notation

- $\hat{\lambda_N}(\cdot)$: Estimated PS fit on an *external sample* of size N
- F_λ : True conditional distribution of $\mathbf{Z}$ as a function of $\lambda(\cdot)$
- $F_{\hat{\lambda_N}}$: Empirical conditional distribution of $\mathbf{Z}$ as a function of $\hat{\lambda_N}(\cdot)$
- $p_{\hat{\lambda_N},n}$: p-value produced by a nominal level-α test using $F_{\hat{\lambda_N}}$

Impact of PS estimation error

Theorem 1 ensures that the test size is approximately preserved when the propensity score is well-estimated.

Theorem 1

If no unobserved confounding is present and the sharp null hypothesis of no effect is true, then

$$P(p_{\hat{\lambda_N},n} \leq \alpha) \leq \alpha + d_{TV}(F_\lambda, F_{\hat{\lambda_N}})$$

where $d_{TV}(P, Q) = \sup_A |P(A) - Q(A)|$ is the total variation distance between distributions P and Q.

i.e., when the estimated and true propensity scores are close, the nominal Type I error will be close to the true Type I error.

Impact of PS estimation error

Moreover, **Theorem 2** provides a desirable asymptotic property for the case in which the true propensity score follows a *logistic regression model*.

Theorem 2

Suppose that $P(Z = 1 \mid X) = \lambda(X) = \frac{1}{1+\exp(-\beta^T X)}$ and that $\hat{\lambda_N}$ is obtained by estimating this model using maximum likelihood. Suppose furthermore that $\lim_{n\to\infty} n/N = 0$ and that covariates X have compact support. Then under the conditions of Theorem 1,

$$\lim \sup_{n,N\to\infty} P(p_{\hat{\lambda_N},n} \leq \alpha) \leq \alpha.$$

Impact of PS estimation error

Nevertheless, **Theorem 2** comes with a few limitations.

- If the propensity score model is not logistic, the bound on the rate of convergence of Type I error violation toward zero is much weaker.

  $\longrightarrow$ Use **Theorem 1** to directly explore model misspecification.

- The asymptotic regime that assumes a pilot sample with $\lim_{n\to\infty} n/N = 0$ is unlikely to hold in reality.

  $\longrightarrow$ Simulation studies reveal robustness to such a violation.

Table of contents

Review of the sensitivity analysis framework

In matched observational studies, **sensitivity analysis** can be conducted to assess the robustness of the results to hidden bias. Rosenbaum's sensitivity analysis model restricts the true probabilities of treatment using a constant $\Gamma \geq 1$:

$$1/\Gamma \leq \frac{\pi(x,u)(1-\pi(x,u'))}{\pi(x,u')(1-\pi(x,u))} \leq \Gamma \quad \text{for all } x, u, u'. \tag{5}$$

This is equivalent to the following model, with arbitrary $\kappa(\cdot)$, $\gamma = \log(\Gamma)$ and $0 \leq U \leq 1$:

$$\log\left(\frac{\pi(X,U)}{1-\pi(X,U)}\right) = \kappa(X) + \gamma U. \tag{6}$$

Review of the sensitivity analysis framework

Using the sensitivity model, we compute the *worst-case p-values* over all possible values of U.

- Here, we focus on *sum statistics* $T(\mathbf{Z}, \mathbf{Y}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} Z_{ki} f_{ki}(\mathbf{Y})$, where larger values of $T(\mathbf{Z}, \mathbf{Y})$ lead to rejection in a one-sided test.
- Suppose the n_k units in each matched set are arranged in increasing order of f_{ki} so that $f_{k1} \leq \cdots \leq f_{kn_k}$ for all k.
- Consider the following set of n-tuples taking the most extreme values of U:
  - U^+ : Set of $\mathbf{u}$ such that $u_{ki} \in \{0, 1\}$ and $u_{k1} \leq \cdots \leq u_{kn_k}$,
  - U^- : Set of $\mathbf{u}$ such that $u_{ki} \in \{0, 1\}$ and $u_{k1} \geq \cdots \geq u_{kn_k}$.

Review of the sensitivity analysis framework

Let $\alpha_{unif}(\mathbf{Y}, \mathbf{u})$ represent the p-value under uniform randomization for a given $\mathbf{u}$ using $T(\mathbf{Z}, \mathbf{Y})$. When mathcing on the propensity score is *exact*, the significance level can be bounded as follows:

$$\min_{\mathbf{u} \in U^-} \alpha_{unif}(\mathbf{Y}, \mathbf{u}) \leq \alpha_{unif}(\mathbf{Y}, \mathbf{U}) \leq \max_{\mathbf{u} \in U^+} \alpha_{unif}(\mathbf{Y}, \mathbf{u}).$$

This allows us to bound the results of the hypothesis test by searching over a highly structured finite set U^+ and U^- of candidate $\mathbf{u}$-values.

Review of the sensitivity analysis framework

Nevertheless, the sets U^+ and U^- may become large and complex, resulting in computational issues. Gastwirth et al. (2000) proposed a strategy to find *approximate bounds* for the tail probabilities.

Specifically, consider the additive contribution of the k-th matched set $T_k = \sum_{i=1}^{n_k} Z_{ki} f_{ki}(\mathbf{Y})$ to the test statistic T.

- We look for $\mathbf{u}_k$ that maximizes the expectation of T_k.
- If there is a tie, maximize the variance.
- The maximization of T_k is achieved only at $\mathbf{u}_k$ such that $u_{k1} = \cdots = u_{kl} = 0$ and $u_{k(l+1)} = \cdots = u_{kn_k} = 1$ for some l, so that it suffices to consider the n_k expectations and variances μ_{kl} and v_{kl}.

Review of the sensitivity analysis framework

For the k-th matched set,

- Let $\mu_k = \max_{l=1,\ldots,n_k} \mu_{kl}$,
- Let $v_k = \max_{a \in A_k} v_{ka}$, where A_k is the set of values of l that maximize μ_{kl}.

Then the *approximate maximum p-value over* $\mathbf{U}$ is given by

$$1 - \Phi\left(\frac{T - \sum_{k=1}^{K} \mu_k}{\sqrt{\sum_{k=1}^{K} v_k}}\right)$$

which can be used to compute a conservative bound for testing the one-sided hypothesis.

Sensitivity analysis under covariate-adaptive randomization

We extend the aforementioned framework under covariate-adaptive randomization inference. Define $\alpha_{adapt}(\mathbf{Y}, \mathbf{u})$ as the p-value under covariate-adaptive randomization using the true propensity score.

Theorem 3 provides analogous sensitivity bounds for significance levels:

Theorem 3

For any $\mathbf{u} \in [0,1]^n$, we have

$$\min_{\mathbf{u} \in U^-} \alpha_{adapt}(\mathbf{Y}, \mathbf{u}) \leq \alpha_{adapt}(\mathbf{Y}, \mathbf{U}) \leq \max_{\mathbf{u} \in U^+} \alpha_{adapt}(\mathbf{Y}, \mathbf{u}).$$

Sensitivity analysis under covariate-adaptive randomization

The large-sample approach of Gastwirth et al. (2000) can also be extended. In particular, the key quantities μ_{kl} and v_{kl} take the following values:

$$\mu_{kl} = \frac{\sum_{i=1}^{l} p_{ki} f_{ki} + \Gamma \sum_{i=l+1}^{n_k} p_{ki} f_{ki}}{\sum_{i=1}^{l} p_{ki} + \Gamma \sum_{i=l+1}^{n_k} p_{ki}},$$

$$v_{kl} = \frac{\sum_{i=1}^{l} p_{ki} f_{ki}^2 + \Gamma \sum_{i=l+1}^{n_k} p_{ki} f_{ki}^2}{\sum_{i=1}^{l} p_{ki} + \Gamma \sum_{i=l+1}^{n_k} p_{ki}} - \mu_{kl}^2.$$

Sensitivity analysis under covariate-adaptive randomization

Remark. Whereas $\kappa(X)$ has no bearing on sensitivity analysis under exact matching, *we require some bound or estimate of* $\kappa(X)$ to compute sensitivity bounds under covariate-adaptive randomization.

In general, $\lambda(X)$ provides information about $\kappa(X)$ as follows:

$$\lambda(X) = \mathbb{E}[\pi(X, U) \mid X] = \int_0^1 \frac{1}{1 + \exp(-\kappa(X) - \gamma u)} du.$$

Since $\pi(x, u)$ is increasing in u for all x, this implies

$$\kappa(X) \in \left[\log\left(\frac{\lambda(X)}{1 - \lambda(X)}\right) - \gamma, \log\left(\frac{\lambda(X)}{1 - \lambda(X)}\right) \right].$$

Table of contents

Finite-sample performance via simulation

Through a simulation study, we aim to observe:

- The performance of covariate-adaptive randomization inference in matched designs compared to traditional inference procedures, with a focus on *Type I error control and coverage probability*.
- The *role of matched design* in covariate-adaptive randomization inference relative to nonuniform permutation in an unmatched study.

Z-dependence

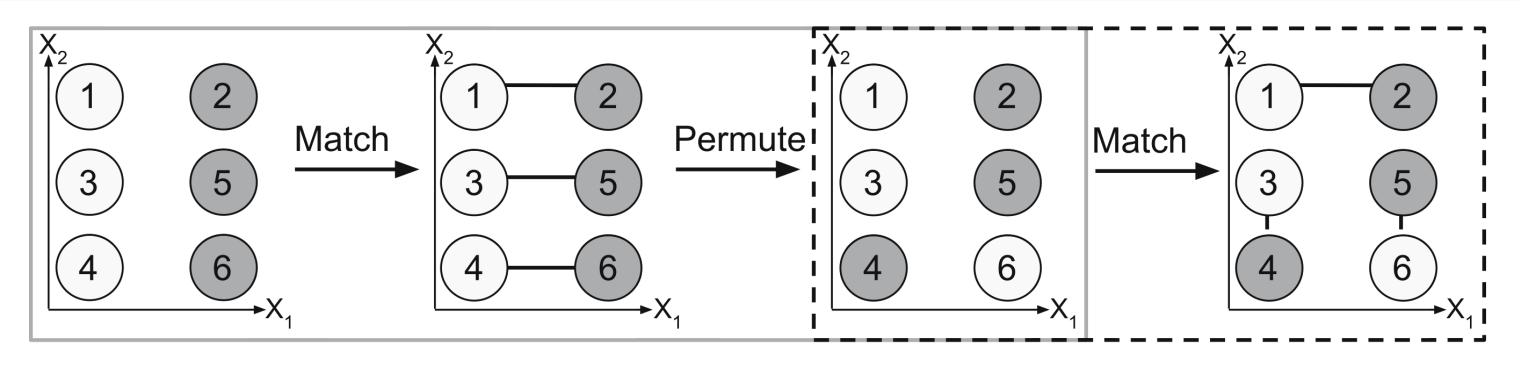
Before we proceed, we need to address the issue of **Z-dependence** in matched designs.

So far, we assumed that the match $\mathcal{M}$ is fixed over all possible $\mathbf{Z}$-values. However, *a matched design is a function of the treatment status*, since it is constructed with reference to both $\mathbf{X}$ and $\mathbf{Z}$.

- In the case of exact matching this issue need not arise, as the matched design $\mathcal{M}$ remains unchanged for all $\mathbf{Z}$.
- When matching is inexact this guarantee no longer holds, and for some $\mathbf{Z} \in \mathcal{Z}_{\mathcal{M}}$ it may be the case that $\mathcal{M}$ is not selected.

Z-dependence

Example. Greedy matching algorithm in a small sample



- Values of $\mathbf{Z}$ that would have produced a different match do not belong in the support of treatment distribution conditional on $\mathcal{Z}_{\mathcal{M}}$.
- The authors refer to this phenomenon of reduced support as *Z-dependence*.

Simulation settings

Data generation:

- The covariates $X \in \mathbb{R}^p$ are drawn from independent standard normal distributions.
  - Number of observations n: 100 or 1,000
  - Number of covariates p: 2, 5 or 10
- The true propensity score is given by one of the following models:
  - Linear: $\text{logit}[P(Z = 1 \mid X)] = \log(0.3/0.7) + \Delta \cdot X_1$
  - Nonlinear: $\text{logit}[P(Z = 1 \mid X)] = \log(0.3/0.7) + \frac{\Delta}{\sqrt{265}}(X_1 + 4X_1^3)$
  - $\Delta = 0.2$ (weak PS signal) or $\Delta = 0.6$ (strong PS signal)
- The outcomes are generated by one of the following models, with additive noise $\varepsilon \sim N(0, 2^2)$:
  - Linear: $Y = X_1 + \varepsilon$
  - Nonlinear: $Y = \frac{1}{\sqrt{265}}(X_1 + 4X_1^3) + \varepsilon$

Simulation settings

Matching procedure and randomized inference:

- Matching is conducted on the joint (X, Z) datasets using a robust Mahalanobis distance.
- We consider cases both with and without calipers, with the caliper width set to 0.2 times the sample standard deviation of the fitted propensity scores.
- If a PS caliper is used, the PS model is fitted with a logistic model with linear additive terms using maximum likehood.
- The null distribution is obtained by 5,000 Monte Carlo draws to test the one-sided hypothesis at a significance level of $\alpha = 0.05$.

Simulation settings

Accounting for Z-dependence: We consider two scenarios for each simulation setting, with and without Z-dependence.

- With Z-dependence: Consider all permutations in $\mathcal{Z}_{\mathcal{M}}$, with respect to the original treatment vector $\mathbf{Z}$.
- Without Z-dependence: Consider $\mathcal{M}$ merely as a means of stratification, and re-draw the treatment vector after matching according to the true conditional distribution of treatment using the true propensity scores.
    - i.e., as if it were a stratified randomized experiment

Simulation 1: Type I error control

In the **first simulation**, we compare (i) covariate-adaptive randomization inference with true PS, (ii) covariate-adaptive randomization inference with estimated PS, and (iii) uniform randomization inference.

- We assess the *control of Type I error* by computing the proportion of rejections in a nominal level-0.05 test across 8,000 iterations.
- We then compare the precision of inference by computing the *average length of the 95% CIs* for cases where the Type I error was controlled at 0.05 or less.

Simulation 1: Type I error control

n=100, p=2

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.055	0.057	0.056	0.054	0.055	0.055	Yes	Yes
		0.041	0.045	0.044	0.054	0.059	0.058	No	Yes
		0.051	0.058	0.058	0.051	0.060	0.060	Yes	No
		0.064	0.083	0.110	0.054	0.073	0.093	No	No
Linear	Nonlinear	0.048	0.048	0.048	0.052	0.051	0.051	Yes	Yes
		0.041	0.040	0.038	0.055	0.051	0.050	No	Yes
		0.044	0.048	0.049	0.052	0.056	0.056	Yes	No
		0.055	0.061	0.078	0.053	0.058	0.066	No	No
Nonlinear	Linear	0.051	0.050	0.051	0.051	0.051	0.050	Yes	Yes
		0.028	0.029	0.028	0.048	0.048	0.046	No	Yes
		0.051	0.052	0.052	0.051	0.053	0.052	Yes	No
		0.060	0.056	0.088	0.048	0.048	0.065	No	No
Linear	Linear	0.050	0.049	0.050	0.053	0.054	0.052	Yes	Yes
		0.042	0.041	0.041	0.052	0.054	0.051	No	Yes
		0.049	0.050	0.052	0.054	0.056	0.057	Yes	No
		0.068	0.065	0.089	0.054	0.053	0.065	No	No

n=100, p=5

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.062	0.071	0.070	0.046	0.052	0.052	Yes	Yes
		0.042	0.053	0.048	0.054	0.072	0.063	No	Yes
		0.044	0.067	0.069	0.048	0.078	0.078	Yes	No
		0.057	0.074	0.131	0.052	0.082	0.115	No	No
Linear	Nonlinear	0.057	0.055	0.056	0.053	0.052	0.051	Yes	Yes
		0.046	0.044	0.040	0.047	0.050	0.045	No	Yes
		0.042	0.054	0.055	0.052	0.073	0.073	Yes	No
		0.062	0.060	0.100	0.048	0.056	0.077	No	No
Nonlinear	Linear	0.058	0.055	0.056	0.050	0.049	0.048	Yes	Yes
		0.031	0.033	0.029	0.048	0.052	0.044	No	Yes
		0.042	0.049	0.050	0.050	0.064	0.063	Yes	No
		0.060	0.052	0.112	0.054	0.058	0.091	No	No
Linear	Linear	0.058	0.056	0.054	0.050	0.048	0.048	Yes	Yes
		0.042	0.045	0.039	0.046	0.051	0.044	No	Yes
		0.041	0.046	0.050	0.048	0.070	0.072	Yes	No
		0.073	0.062	0.127	0.046	0.047	0.078	No	No

n=1000, p=10

		With Z-dependence			Without Z-dependence				
Z−model	Y−model	Oracle	Estimate	Uniform	Oracle	Estimate	Uniform	Caliper?	Regression?
Nonlinear	Nonlinear	0.052	0.080	0.080	0.048	0.074	0.074	Yes	Yes
		0.030	0.073	0.068	0.048	0.118	0.109	No	Yes
		0.041	0.098	0.100	0.052	0.116	0.118	Yes	No
		0.072	0.234	0.521	0.050	0.184	0.449	No	No
Linear	Nonlinear	0.048	0.049	0.050	0.051	0.051	0.051	Yes	Yes
		0.046	0.045	0.044	0.055	0.055	0.052	No	Yes
		0.037	0.059	0.063	0.051	0.081	0.084	Yes	No
		0.097	0.138	0.334	0.054	0.083	0.224	No	No
Nonlinear	Linear	0.044	0.044	0.043	0.051	0.050	0.050	Yes	Yes
		0.025	0.025	0.022	0.049	0.050	0.044	No	Yes
		0.042	0.054	0.059	0.051	0.067	0.070	Yes	No
		0.109	0.103	0.397	0.048	0.046	0.254	No	No
Linear	Linear	0.050	0.050	0.050	0.052	0.051	0.051	Yes	Yes
		0.046	0.047	0.043	0.049	0.050	0.046	No	Yes
		0.048	0.061	0.065	0.050	0.067	0.070	Yes	No
		0.140	0.133	0.436	0.050	0.050	0.225	No	No

Figure 1: Type I error results for uniform and covariate-adaptive inference.

Colors correspond to the magnitude of the error rates, and triangles indicate that a one-sample z-test rejected the null hypothesis that the error rate was 0.05.

All results are under $\Delta = 0.6$ (strong PS signal).

Simulation 1: Type I error control

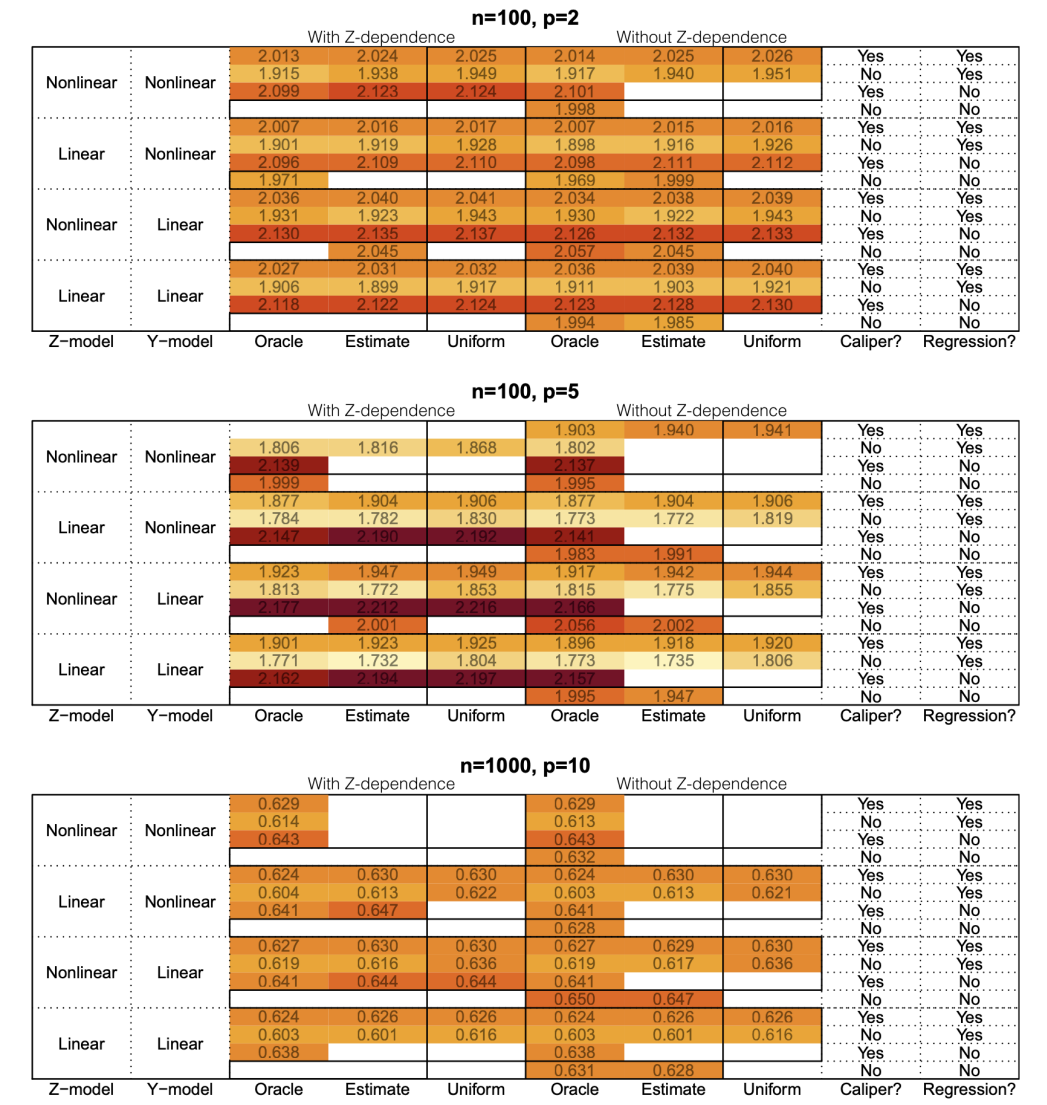


Figure 2: Average CI length for cases with approximate control of Type I error at 0.05 or less.

Darker colors indicate longer CIs.

Simulation 1: Type I error control

Observations:

- The simulations emphasize the importance of taking measures to control covariate discrepancies within matched pairs and guarantee type I error control for post-matching inference.
  - The importance of such measures appears to increase with the size and complexity of the dataset.
  - The type of adjustment seems to matter less than the fact of employing it.
- While covariate-adaptive randomization inference is slightly less succesful than PS calipers or regression adjustments, this appears to be a product of Z-dependence.
- Covariate-adaptive inference holds a small edge over caliper approaches in terms of CI length, because it does not require a reduction in sample size to correct discrepancies.

Simulation 2: The role of matched design

In the **second simulation**, we investigate the role of matched design in covariate-adaptive randomization inference. Specifically, we compare the results to those obtained by applying nonuniform randomization inference in an unmatched study, as described in Branson and Bind (2019).

- We once again examine the control over Type I error and the average CI lengths.
- We also probe the possibility that matching confers robustness to incorrect model assumptions using a new simulation setting with a badly misspecified PS score,

$$\text{logit}[P(Z = 1 \mid X)] = \log(0.3/0.7) + \mathbf{I}(X_1 X_2 \geq 0).$$

Simulation 2: The role of matched design

Specification			Type I Error			Confidence Interval Length		
Z-model	Y-model	Regression?	Unmatched	Matched (Z-dependence?)		Unmatched	Matched (Z-dependence?)	
				With	Without		With	Without
Linear	Linear	No	0.000	0.284	0.045	8.68	3.41	3.43
Linear	Linear	Yes	0.053	0.166	0.047	1.81	5.73	2.01
Linear	Nonlinear	No	0.001	0.037	0.051	7.70	2.11	3.22
Linear	Nonlinear	Yes	0.041	0.014	0.052	4.76	3.57	2.03
Nonlinear	Linear	No	0.000	0.258	0.056	8.85	3.39	5.22
Nonlinear	Linear	Yes	0.048	0.265	0.055	1.78	5.89	1.90
Nonlinear	Nonlinear	No	0.008	0.044	0.064	7.96	2.09	5.16
Nonlinear	Nonlinear	Yes	0.217	0.040	0.056	4.85	3.50	1.89

Table 1. Simulation results comparing matched and unmatched studies employing covariate-adaptive randomization inference ($n = 100, p = 2$, 5,000 iterations, strong PS signal)

Simulation 2: The role of matched design

Observations:

- When well-specified regression models are fit, the matched and unmatched studies show similar performance in the absence of Z-dependence.
    - In this case, the CIs are more precise in the unmatched setting, due to availability of more data to estimate the outcome model.
- However, when regression adjustment is absent or based on an incorrect model, the matched design has substantially increased precision relative to the unmatched designs.
- Type I error control may fail for matched designs with Z-dependence, whereas the inference in unmatched studies are punishingly conservative.

Simulation 2: The role of matched design

Misspecified PS model: $\text{logit}[P(Z = 1 \mid X)] = \log\left(\frac{0.3}{0.7}\right) + \mathbf{I}(X_1 X_2 \geq 0)$

- Two very distinct values of the propensity score are present, with subjects in different quadrants of (X_1, X_2).
- Matching confers robustness benefits not enjoyed by unmatched studies.

Regression?	**Unmatched**	**Matched**	
		with Z-dependence	**no Z-dependence**
No	0.135	0.072	0.064
Yes	0.139	0.059	0.058

Table 2. Type I errors of matched and unmatched studies under covariate-adaptive randomization inference under misspecification of the propensity score model ($n = 100, p = 2$, 5,000 iterations, strong PS signal)

Table of contents

Real data example - Welders dataset

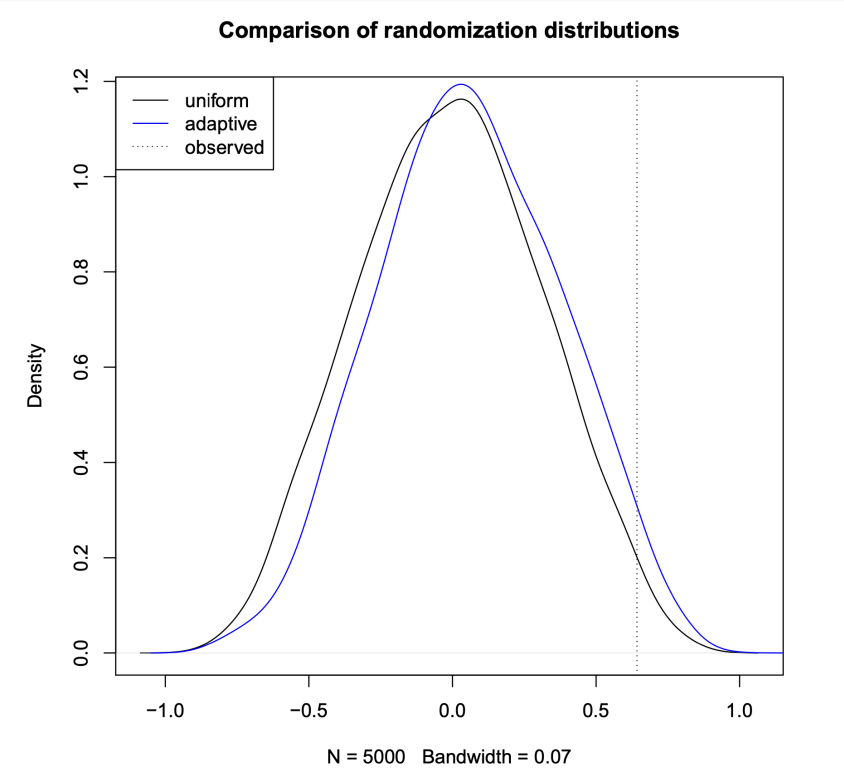
We now apply covariate-adaptive randomization inference to a real dataset originally due to Costa et al. (1993). This dataset compares genetic damage among welders and controls in other occupations.

- 21 treated units (welders), 26 control units (non-welders)
- Outcome: Genetic damage
- Covariates: Race, age, smoking status
- Robust Mahalanobis distance matching + soft PS calipers ($0.5 \times$ standard deviation of fitted PS)
- Test statistic: Difference-in-means estimator
- Summary of treated-control differences in covariates and PS:

	African-American	**Age**	**Smoker**	**PS**
Average	-0.05	-1.48	0.00	0.06

Real data example - Welders dataset

We constrast uniform randomization inference structured by matched pairs with covariate-adaptive randomization inference.



- The one-sided p-values are 0.015 and 0.029 for the uniform and covariate-adaptive tests, respectively.
- The respective threshold Γ for sensitivity analysis are 1.09 and 1.05.
- This is due to the right-shifted covariate-adaptive distribution, which accounts for positive residual PS differences leading to larger values under the null.

Table of contents

Discussion

Key contributions of the paper:

(i) Covariate-adaptive randomization inference is a valid method that closely attains the nominal Type I error rate and coverage probability.

- It can serve as an attractive alternative alongside approaches based on regression adjustments and matching calipers.

(ii) Combining matching and covariate-adaptive randomization inference offers an improvement over nonuniform inference in unmatched designs.

- It provides more accurate inference and robustness to propensity score misspecification, although it comes at the cost of introducing Z-dependence.
- Sensitivity analysis is still feasible.

Discussion

Potential future directions:

(i) Relaxing the strong assumptions for the theroetical guarantees, including restrictions on the relative sample sizes and the role of a hypothetical pilot sample used to fit the propensity score model.

(ii) Developing stronger model-based guidance about the imporatance of the propensity score and the construction of matched designs.

(iii) Exploring the nontrivial role of Z-dependence in matched randomized tests.

(iv) Extending the framework to test weak null hypotheses, which allow for unknown heterogeneous causal effects.