Optimality of Matched-Pair Designs in Randomized Controlled Trials

Yuehao Bai, University of Michigan

This paper studies design of randomized controlled trials (RCTs).

- AEA RCT Registry has $\sim$4,000 experiments.

In particular, I study stratified randomization.

- Partition the units into strata based on observed covariates.
- Common in RCTs: $>$600 experiments in AEA RCT Registry.

Why stratify?

Leading motivation: reducing bias after treatment assignment.

By ensuring that the treatment and comparison groups are balanced on key characteristics, stratification reduces the chance that, for example, all the children with high (baseline) test scores will happen to be in our treatment group and we will overestimate the treatment impact as a result.

— Running Randomized Evaluations: A Practical Guide

Glennerster and Takavarasha (2013)

How to stratify?

- How many strata?
- What covariates to stratify on?
- Matching or coarse partition?

This paper asks: What is the optimal way to stratify?

- Consider estimation of Average Treatment Effect (ATE).
- Today: treated fraction $= 1/2$ for each stratum (relaxed in paper).
- Natural to estimate ATE by difference-in-means of treated and control.
- Choose stratification to minimize mean-squared error (MSE).

Combinatorial problem: many ways to partition.

I show solution is a matched-pair design: 2 units in each stratum.

Take certain 1-dimensional index function of covariates.

(expected outcome given covariates, with $\frac{1}{2}$ treatment probability)

$\Downarrow$

Order units according to this function.

$\Downarrow$

Pair first 2 units, second 2 units, …

Results capture important motivation for stratifying.

- Concern about ex-post bias: bias conditional on realized treatment.
- I show minimizing MSE $\Leftrightarrow$ minimizing size of ex-post bias.

Optimal stratification relies on index function.

- With pilot data, could estimate index function.
- Pilot already frequently run for many reasons.
- Below provide feasible procedures with large/small/no pilot.

THIS PAPER (CONT.)

Inference for ATE complicated by large number of strata (sample size / 2).

- Provide computationally easy procedures for inference.

Run experiment with pilot, under proposed stratification on MTurk.

- Same treatment and outcome as in DellaVigna and Pope (2018).
- Standard error reduced by 29% $\Rightarrow$ require half sample size for same s.e.

Stratified randomization and matched-pair design

Ashraf, Karlan & Yin '06, Duflo, Glennerster & Kremer '07, Angrist & Lavy '09, Bruhn & McKenzie '09, Niehaus & Sukhtankar '12, Glennester & Takavarasha '13, Niehaus & Sukhtankar '13, Banerjee, Duflo, Glennester & Kinnan '15, Crépon, Devoto, Duflo & Parienté '15, Groh & McKenzie '16 …

Optimal design of experiments

Cox & Reid '00, Bailey '04, Pukelsheim '06, Hahn, Hirano & Karlan '11, Barrios '13, Tabord-Meehan '18, Banerjee, Chassang, Montero & Snowberg '19

Inference for matched-pair designs

Imbens & Rubin '15, Athey & Imbens '17, Fogarty '18(ab), Bai, Romano & Shaikh '19

Setup and notation

Optimal stratification

Feasible procedures

Inference

Empirical application

i.i.d. units $1 \leq i \leq n$. For the ith unit:

$$X_i = \text{observed covariates} \in \mathbf{R}^p$$

$$D_i = \text{treatment status}$$

$$\underbrace{Y_i}_{\text{observed outcome}} = \underbrace{Y_i(1)}_{\text{potential outcome if treated}} D_i + \underbrace{Y_i(0)}_{\text{potential outcome if not treated}} (1 - D_i)\,.$$

Parameter of interest: average treatment affect (ATE)

$$\theta = E[Y_i(1) - Y_i(0)]\,.$$

SETUP AND NOTATION (CONT.)

Stratification $\lambda = \{\lambda_s : 1 \leq s \leq S\}$ is partition of $\{1, \dots, n\}$.

Define $\Lambda =$ set of all stratifications.

Define $\tau_s =$ treated fraction in λ_s.

Assumption $\tau_s \equiv \tau = \frac{1}{2}$.

Implies $|\lambda_s|$ is even, $\forall s$.

Relaxed in paper.

Treated fraction τ_s same in each stratum.

$\Rightarrow$ natural to use difference-in-means estimator:

$$\hat{\theta}_n = \hat{\mu}_n(1) - \hat{\mu}_n(0) ,$$

where $\hat{\mu}_n(d) =$ average of outcomes among treated/control units

$$= \frac{1}{n/2} \sum_{1 \le i \le n : D_i = d} Y_i .$$

$\hat{\theta}_n =$ regression with strata fixed effects = "fully saturated" regression.

Shorthand:

$$X^{(n)} = (X_1, \dots, X_n)'$$
$$D^{(n)} = (D_1, \dots, D_n)' .$$

Problem

Consider

$$\min_{\lambda \in \Lambda} \ \mathrm{MSE}(\lambda | X^{(n)}) , \qquad \text{(MIN-MSE)}$$

where

$$\mathrm{MSE}(\lambda | X^{(n)}) = E_{\lambda}[(\hat{\theta}_n - \theta)^2 | X^{(n)}] .$$

E_λ: distribution of treatment status $D^{(n)}$ depends on λ.

To introduce intuition in solving (MIN-MSE), define

Ex-ante bias = $E_\lambda[\widehat{\theta}_n | X^{(n)}] - \theta$.

Ex-post bias = $E[\widehat{\theta}_n | X^{(n)}, D^{(n)}] - \theta$.

Ex-ante: before treatment status is realized.

Ex-post: after treatment status is realized.

Distribution of ex-post bias (over $D^{(n)}$) centered at ex-ante bias:

$$E_\lambda[\text{ex-post bias} | X^{(n)}] = \text{ex-ante bias} .$$

OPTIMAL STRATIFICATION (CONT.)

Ex-ante bias only depends on marginal treatment prob. of each unit.

- Difference-in-means estimator

$$\hat{\theta}_n = \frac{1}{n/2} \sum_{1 \le i \le n} \left[D_i Y_i(1) - (1 - D_i) Y_i(0) \right] .$$

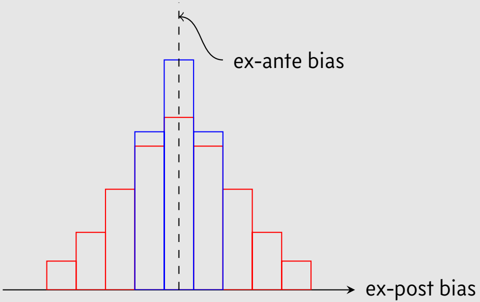
- Ex-ante bias doesn't depend on λ:

$$\frac{1}{n} \sum_{1 \le i \le n} \left(E[Y_i(1) | X_i] - E[Y_i(0) | X_i] \right) - \theta .$$

But distribution of ex-post bias does depend on λ.

- Ex-post bias depends on realized treatment status D_i.

OPTIMAL STRATIFICATION (CONT.)



Distributions of ex-post bias under two stratifications (red and blue).

First element of proof

Only second moment of ex-post bias matters.

Further equivalent to minimizing dispersion of ex-post bias.

Lemma Solutions to (MIN-MSE) are exactly those to

$$\min_{\lambda \in \Lambda_n} E_\lambda[\text{ex-post bias}^2 | X^{(n)}] ,$$

and those to

$$\min_{\lambda \in \Lambda_n} \text{Var}_\lambda[\text{ex-post bias} | X^{(n)}] .$$

First element of proof

Only second moment of ex-post bias matters.

Further equivalent to minimizing dispersion of ex-post bias.

Lemma Solutions to (MIN-MSE) are exactly those to

$$\min_{\lambda \in \Lambda_n} E_\lambda[\text{ex-post bias}^2 | X^{(n)}] ,$$

and those to

$$\min_{\lambda \in \Lambda_n} \operatorname{Var}_\lambda[\text{ex-post bias} | X^{(n)}] .$$

Reason: for any random variable Z,

$$\operatorname{Var}[Z] = E[Z^2] - E[Z]^2 .$$

First element of proof

Only second moment of ex-post bias matters.

Further equivalent to minimizing dispersion of ex-post bias.

Lemma Solutions to (MIN-MSE) are exactly those to

$$\min_{\lambda \in \Lambda_n} E_\lambda[\text{ex-post bias}^2 | X^{(n)}] ,$$

and those to

$$\min_{\lambda \in \Lambda_n} \operatorname{Var}_\lambda[\text{ex-post bias} | X^{(n)}] .$$

Reason:

$$\operatorname{Var}_\lambda[\text{ex-post bias} | X^{(n)}] = E_\lambda[\text{ex-post bias}^2 | X^{(n)}] - \text{ex-ante bias}^2 .$$

Ex-ante bias same across λ.

Second element of proof

Only need to consider matched-pair designs, i.e., $|\lambda_s| \equiv 2$.

Every stratification is a mixing of matched pair designs.

e.g.,

$$\{1, 2, 3, 4\}$$

equivalent to three matched-pair designs with $\frac{1}{3}$ probability each:

$$\{\{1, 2\}, \{3, 4\}\}, \{\{1, 3\}, \{2, 4\}\}, \{\{1, 4\}, \{2, 3\}\} \, .$$

Mixing cannot be give smaller MSE than optimal “pure strategies.”

To solve (MIN-MSE), combine previous results:

(a) Only need to consider pairs.

(b) Note that any matched-pair design λ can be written as

$$\lambda = \{\{\pi(2s-1), \pi(2s)\} : 1 \leq s \leq n/2\},$$

for some $\pi \in \Pi =$ set of all permutations on $\{1, \ldots, n\}$. Moreover,

$$\mathrm{Var}_\lambda[\text{ex-post bias}|X^{(n)}, \lambda] \propto \sum_{1 \leq s \leq n/2} (g_{\pi(2s-1)} - g_{\pi(2s)})^2,$$

where $g(x) = E[Y_i(1) + Y_i(0)|X_i = x]$ and $g_i = g(X_i)$.

Finally, use Hardy-Littlewood-Pólya rearrangement inequality.

oracle-rearrange-detail pen-intro not-half

OPTIMAL STRATIFICATION (CONT.)

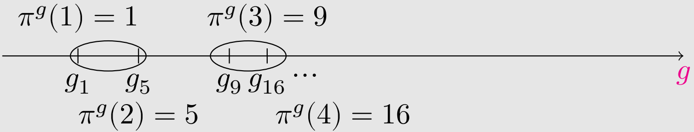
Define $g(x) = E[Y_i(1) + Y_i(0)|X_i = x]$.

Intuition: expected outcome given covariates, with $\frac{1}{2}$ treat. prob.

Let π^g be a permutation of $\{1, \ldots, n\}$ such that $g_{\pi^g(1)} \leq \ldots \leq g_{\pi^g(n)}$.

Theorem The solution to (MIN-MSE) is

$$\lambda^g(X^{(n)}) = \{\{\pi^g(2s-1), \pi^g(2s)\} : 1 \leq s \leq n/2\} .$$



Remark In test score example: match on baseline test scores.

Remark Matched-pair designs have been used in different forms.

$>$40 ongoing experiments in AEA RCT Registry.

Bruhn and McKenzie '09: 56% interviewed have used matched pairs.

Angrist & Lavy '09: baseline outcome.

Banerjee, Duflo, Glennester & Kinnan '15: match on two covariates.

Remark Results justify a certain matched-pair design.

Setup and notation

Optimal stratification

Feasible procedures

Inference

Empirical application

Index function g is unknown so needs to be estimated.

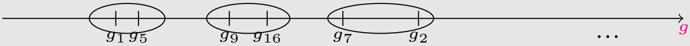
Suppose we have data from a pilot experiment of sample size m.

Pilot has same distr. as but is indep. of main experiment units.

Plug-in an estimate of g called $\hat{g}_m$:

$$\lambda^{\hat{g}_m}(X^{(n)}) = \{\{\pi^{\hat{g}_m}(2s-1), \pi^{\hat{g}_m}(2s)\} : 1 \leq s \leq n/2\},$$

where $\hat{g}_{m,\pi^{\hat{g}_m}(1)} \leq \ldots \leq \hat{g}_{m,\pi^{\hat{g}_m}(n)}$.



Index function g is unknown so needs to be estimated.

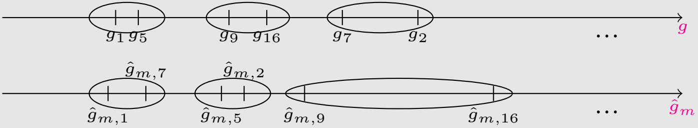
Suppose we have data from a pilot experiment of sample size m.

Pilot has same distr. as but is indep. of main experiment units.

Plug-in an estimate of g called $\hat{g}_m$:

$$\lambda^{\hat{g}_m}(X^{(n)}) = \{\{\pi^{\hat{g}_m}(2s-1), \pi^{\hat{g}_m}(2s)\} : 1 \le s \le n/2\},$$

where $\hat{g}_{m,\pi^{\hat{g}_m}(1)} \le \dots \le \hat{g}_{m,\pi^{\hat{g}_m}(n)}$.



Index function g is unknown so needs to be estimated.

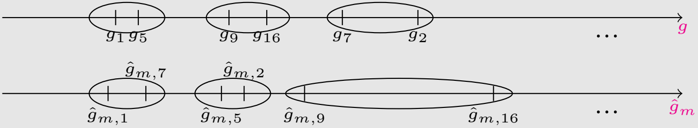
Suppose we have data from a pilot experiment of sample size m.

Pilot has same distr. as but is indep. of main experiment units.

Plug-in an estimate of g called $\hat{g}_m$:

$$\lambda^{\hat{g}_m}(X^{(n)}) = \{\{\pi^{\hat{g}_m}(2s-1), \pi^{\hat{g}_m}(2s)\} : 1 \leq s \leq n/2\},$$

where $\hat{g}_{m,\pi^{\hat{g}_m}(1)} \leq \ldots \leq \hat{g}_{m,\pi^{\hat{g}_m}(n)}$.



$\hat{g}_m$ could also come from previous observational studies.

Consider two settings separately according to the pilot:

Large pilot: nonparametric methods are suitable for $\hat{g}_m$.

Small pilot: nonparametric methods are not suitable for $\hat{g}_m$.

Technically:

Large pilot: pilot sample size m increases in asymptotic framework.

Small pilot: not necessarily.

Large pilot

First consider large pilot.

If $\hat{g}_m$ is L^2-consistent for g, plug-in $\approx$ optimal stratification.

L^2-consistency: as $m \to \infty$,

$$\int_{\mathbf{R}^p} |\hat{g}_m(x) - g(x)|^2 \underbrace{Q_X}_{\text{marginal distr. of} X_i}(dx) \xrightarrow{P} 0 \,. \qquad \text{(L2)}$$

Theorem Under (L2) and regularity conditions, limiting distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$ same as under optimal stratification.

Remark In addition, limits of $n \times$ MSE of $\widehat{\theta}_n$ same under two cases.

Large pilot (cont.)

No need for uniform consistency. No restriction on $\frac{m}{n}$.

Permit machine learning methods in high-dimensional settings.

Sufficient condition for (L2):

Series and sieves: Newey (1997), Chen (2007) …

Kernel: Li and Racine (2007) …

LASSO: Belloni et al. (2012, 2014) …

Random forests: Wager and Walther (2015) …

Neural nets: Chen and White (1999), Farrell et al. (2018) …

Support vector machines: Steinwart and Christmann (2008) …

⋮

Small pilot

Can't consistently estimate g.

Nonparametric methods no longer suitable, so use simple estimator $\hat{g}_m$.

For example, OLS.

For $d \in \{0, 1\}$, define $\hat{\beta}_m(d)$ = OLS estimate for units w/ $D_i = d$.

Define $\hat{\beta}_m = \hat{\beta}_m(1) + \hat{\beta}_m(0)$.

Recall $g(x) = E[Y_i(1)|X_i = x] + E[Y_i(0)|X_i = x]$.

A natural plug-in estimator $\hat{g}_m(x) = x'\hat{\beta}_m$.

Small pilot (cont.)

But under small pilot, concerned with accuracy of $\hat{g}_m$.

$\hat{g}_m$ approximates g poorly $\Rightarrow$ poor stratification.

Recall natural plug-in estimate is $\hat{g}_m(x) = x'\hat{\beta}_m$.

Similar as definition of π^g (optimal stratification), $\pi^{\hat{g}_m}$ solves

$$\min_{\pi \in \Pi_n} \sum_{1 \le s \le n/2} \left(\hat{g}_{m,\pi(2s-1)} - \hat{g}_{m,\pi(2s)}\right)^2 ,$$

or equivalently,

$$\min_{\pi \in \Pi_n} \sum_{1 \le s \le n/2} (X_{\pi(2s-1)} - X_{\pi(2s)})' \hat{\beta}_m \hat{\beta}_m' (X_{\pi(2s-1)} - X_{\pi(2s)}) .$$

Small pilot (cont.)

Regularize the above minimization problem.

Define $\hat{\Sigma}_m(d)$ = variance estimate.

Define $\hat{\Sigma}_m$ = $\hat{\Sigma}_m(1)$ + $\hat{\Sigma}_m(0)$.

Let $\pi^{\text{pen}}(X^{(n)})$ be the solution to

$$\min_{\pi \in \Pi_n} \sum_{1 \leq s \leq n/2} (X_{\pi(2s-1)} - X_{\pi(2s)})' (\hat{\beta}_m \hat{\beta}_m' + \hat{\Sigma}_m)(X_{\pi(2s-1)} - X_{\pi(2s)}) .$$

Define penalized stratification as

$$\lambda^{\text{pen}}(X^{(n)}) = \{\{\pi^{\text{pen}}(2s-1), \pi^{\text{pen}}(2s)\} : 1 \leq s \leq n\} .$$

Small pilot (cont.)

Further justification: $\lambda^{\text{pen}}(X^{(n)})$ is solution to a Bayesian problem!

Define $\tilde{W}_j = (Y_j, X_j, D_j)'$ and $\tilde{W}^{(m)} = (\tilde{W}_1, \dots \tilde{W}_m)'$ as pilot data.

Q_X^n: distribution of $X^{(n)}$. $Q_{\tilde{W}}^m$: distribution of $\tilde{W}^{(m)}$.

F is a prior on g.

$$\min_u \iiint \mathrm{MSE}(u(\tilde{w}^{(m)}, x^{(n)})|g, x^{(n)}) Q_X^n(dx^{(n)}) Q_{\tilde{W}}^m(d\tilde{w}^{(m)}) F(dg)\,.$$

Moreover, assume $g(x) = x'\beta$, so F becomes a prior on β.

Let F be normal, and drive it to diffuse prior (variance $\to \infty$).

Small pilot (cont.)

What does it look like?

$$\min_{\pi \in \Pi_n} \sum_{1 \le s \le n/2} (X_{\pi(2s-1)} - X_{\pi(2s)})'(\hat{\beta}_m \hat{\beta}_m' + \hat{\Sigma}_m)(X_{\pi(2s-1)} - X_{\pi(2s)}) \, .$$

$\hat{\beta}_m$ accurate $\Rightarrow$ $\hat{\Sigma}_m$ small $\Rightarrow$ solution close to $\pi^{\hat{g}_m}$.

$\hat{\beta}_m$ inaccurate $\Rightarrow$ $\hat{\Sigma}_m$ large

$\Rightarrow$ solution close to matching on weighted distance:

$$d(x, \tilde{x}) = (x - \tilde{x})' \hat{\Sigma}_m (x - \tilde{x}) \, .$$

simulation

Small pilot (cont.)

Procedure is computationally easy.

Define R_m from the Cholesky decomposition: $R_m' R_m = \hat{\beta}_m \hat{\beta}_m' + \hat{\Sigma}_m$.

Define $Z_i = R_m X_i$.

Above equivalent to

$$\min_{\pi \in \Pi_n} \sum_{1 \leq s \leq n/2} \left\| Z_{\pi(2s-1)} - Z_{\pi(2s)} \right\|^2 ,$$

Nonbipartite matching on Z, solved in `R` using `nbpMatching`.

Setup and notation

Optimal stratification

Feasible procedures

Inference

Empirical application

With an estimate of ATE, want to perform inference on ATE.

Q: How to test

$$H_0 : \theta = \theta_0 \text{ versus } \theta \neq \theta_0$$

at level $\alpha \in (0, 1)$?

Large pilot

First suppose g is known and we use infeasible optimal stratification.

Recall $\hat{\theta}_n = \hat{\mu}_n(1) - \hat{\mu}_n(0)$.

Natural to start with two-sample t-test

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0)}} \, ,$$

where

$$\hat{\sigma}_n^2(d) = \text{s.e. of treated (control) units} = \frac{1}{n/2} \sum_{1 \le i \le n : D_i = d} (Y_i - \hat{\mu}_n(d))^2 \, .$$

The test is conservative by Bai, Romano & Shaikh (2019).

Nominal level 5%, limiting rejection probability usually <5%.

Large pilot (cont.)

Limiting variance of $\hat{\theta}_n$ is

$$\mathrm{Var}[Y_i(1)]+\mathrm{Var}[Y_i(0)]-\frac{1}{2}E[E[Y_i(1)+Y_i(0)|g(X_i)]^2]+\frac{1}{2}E[Y_i(1)+Y_i(0)]^2\,.$$

All terms easy to estimate except highlighted term.

Provide consistent estimate $\hat{\rho}_n$ of highlighted term.

When $\hat{g}_m$ is used instead of g, details taken care of by L^2-consistency.

Large pilot (cont.)

Define new variance estimate

$$(\hat{\varsigma}_n^{\hat{g}_m})^2 = \hat{\sigma}_n^2(1) + \hat{\sigma}_n^2(0) - \frac{1}{2}\hat{\rho}_n + \frac{1}{2}(\hat{\mu}_n(1) + \hat{\mu}_n(0))^2 \,.$$

Test given by

$$\phi_n^{\hat{g}_m}(W^{(n)}) = I\{|T_n^{\hat{g}_m}(W^{(n)})| > z_{1-\frac{\alpha}{2}}\} \,,$$

where

$$T_n^{\hat{g}_m}(W^{(n)}) = \frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\hat{\varsigma}_n^{\hat{g}_m}} \,.$$

Theorem Under (L2) and regularity conditions, when the null holds,

$$\lim_{m,n\to\infty} E[\phi_n^{\hat{g}_m}(W^{(n)})] = \alpha \,.$$

Small pilot

Now consider small pilots.

For plug-in procedure, same test has correct size.

For penalized procedure, slight change because Z is vector.

Could combine pilot and main data for inference.

simulation

Take existing design in Dellavigna and Pope (2018):

- Large-scale experiment on how 18 incentives affect efforts.
- We copy only one treatment (incentive).
- Subjects asked to press a/b alternately in 10 minutes.
- Y = $\#$ of presses.
- All subjects receive some base payment.
- $D = 1$: 100 presses = $0.01. $D = 0$: no extra payment.

Run experiment on Amazon Mechanical Turk (MTurk).

Sample size $n = 176$. Pilot size $m = 44$.

Original paper: sample size = 1098. One stratum.

We use penalized stratification.

Compare four results:

Pen Penalized stratification.

Combined Penalized stratification, pilot + main data for inference.

Original (scaled) Original results with s.e. scaled to $n + m = 220$.

Original Original results.

EMPIRICAL APPLICATION (CONT.)

	Pen	**Combined**	**Original (scaled)**	**Original**
sample size	176	220	220	1098
$\widehat{\theta}_n$	644	624	-	499
standard error	108.16	92.05	129.95	58.70
t-statistic	5.95	6.78	-	8.50

Summary statistics from Dellavigna and Pope (2018) and our replication.

29% smaller s.e. at same sample size.

Penalized stratification needs only half sample size to attain same s.e.

Matching on almost anything is better than nothing.

Should use a large number of small strata.

MSE $\Leftrightarrow$ ex-post bias.

Small pilot—least squares + penalized stratification.

Large pilot—could use machine learning.

With logistical/political constraints ($\neq$ ½), use straightforward extension.