Improving experiments with
Threshold Blocking

Fredrik Sävje — UC Berkeley — fredriksavje.com

Get the paper!

Sävje (2015) On the advantages of threshold blocking. arXiv:1506.02824

[1] Imai, King, Nall (2009) The essential role of pair matching in cluster-randomized experiments. *Statistical Science* 24(1):29-53.

[2] Moore (2012) Multivariate continuous blocking to improve political science experiments. *Political Analysis* 20(4):460-479.

[3] Higgins, Sävje, Sekhon (2015) Improving massive experiments with threshold blocking. http://goo.gl/xhT6x9

At a glance

- Current multivariate blocking methods (e.g., paired-matching) create fixed-sized blocks: all blocks must be exactly of a certain size.
- Threshold blocking instead imposes a *minimum* size.
- **Relaxing the block structure in this way is guaranteed to yield better blockings for any sample and objective function.**
- When covariates are predictive of the outcome, the method will lead to large reductions in variance.
- Fixed-sized blocking can, however, lead to lower variance when covariates are uninformative.

Blocking

A randomized experiment is the best way to investigate causal effects. There are, however, problems:

- Conditional bias.
- High variance and low power.
- Subgroup analysis can be complicated.
- Fishing for significant results.

These problems can partly be solved with blocking. First construct groups of similar units – blocks – and then assign treatment randomly in fixed proportions within the blocks.

Blocking as optimization

One can rarely construct completely homogeneous blocks. Instead, blocking can be seen as an optimization problem where the objective is to minimize the within-block unit dissimilarity as measured by a distance metric. Current methods solve this problem by forming pairs [1] or groups of some other fixed size [2].

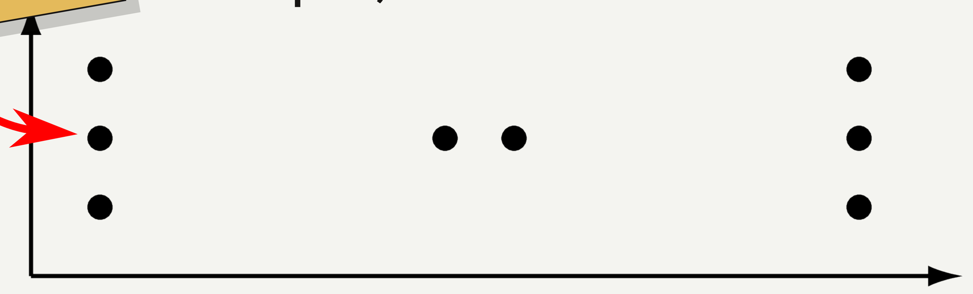
Threshold blocking

Relaxing the block structure

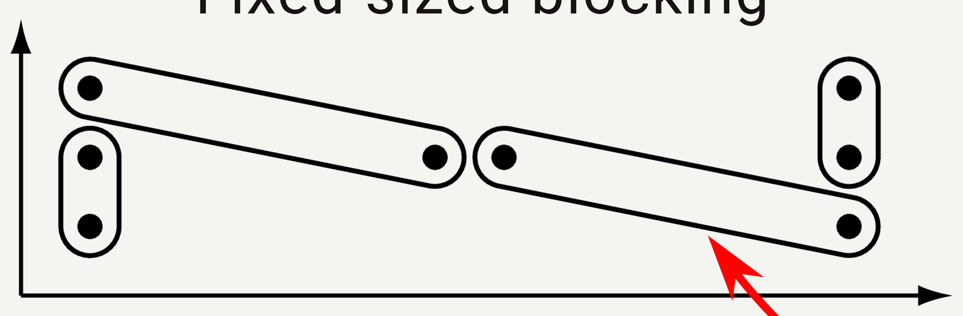
Instead of requiring that all blocks are of a certain size, threshold blocking only imposes a minimum block size. Blocks must contain at least a certain number of units, but they can be larger.

Units in 2D space

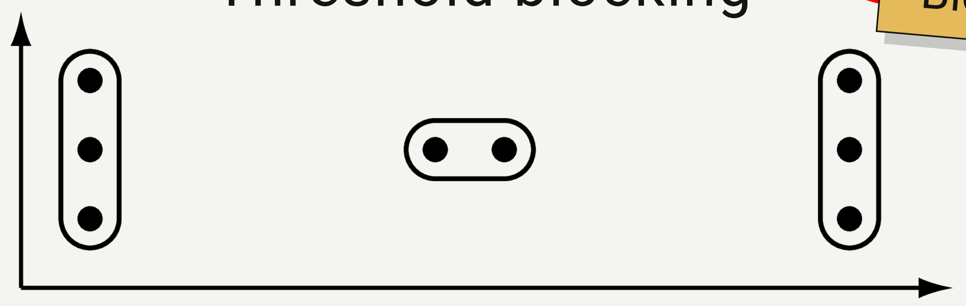
Sample, two covariates



Fixed-sized blocking



Threshold blocking



Blocks

The advantage

Threshold blocking avoids situations where units that are quite different are forced into the same block just to enforce a rigid block structure.

Added flexibility leads to better blockings!

Theorem 1

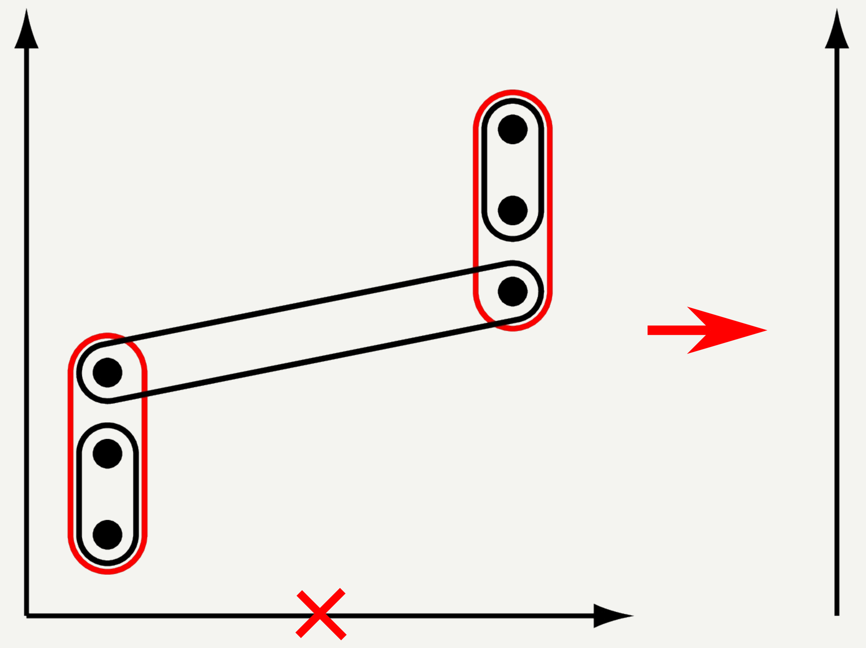
For all samples, all objective functions and all desired block sizes, the optimal threshold blocking is always weakly better than the optimal fixed-sized blocking.

Reducing variance

The goal of blocking cannot always be precisely quantified. Variance reduction of the estimator is one such case: we can never observe the variance that a blocking would produce in advance. Instead, we often construct blocks so to minimize covariate imbalances.

As the mapping between covariate imbalances and variance is unknown, the blocking with least imbalance might not lead to the lowest variance.

Blocking on two covariates when one is irrelevant



A variance trade-off

Variation in # of treated

$$n\,\mathrm{Var}(\hat{\delta}) = \mathrm{E}_{\mathbf{x}}\left[4 \sum_{\mathbf{b}\in\mathcal{B}(\mathbf{x})} \frac{n_{\mathbf{b}}}{n}\left[1 + \mathrm{Std}\left(\frac{1}{2\sum_{i\in\mathbf{b}} T_i}\middle| n_{\mathbf{b}}\right)\right] \mathrm{E}\left(s^2_{y\mathbf{b}}\middle|\mathbf{x}\right)\right]$$

Within-block imbalances

More balanced covariates tend to reduce estimator variance as it avoids imbalances in the potential outcomes. Theorem 1 implies that threshold blocking will perform better than fixed-sized blocking in this aspect, and thus lead to lower variance.

However, unlike previous methods, threshold blocking allows for blocks that are not divisible by the number of treatments. This will introduce variation in the number of treated which tends to increase variance.

With sufficiently informative covariates, the decrease in variance due to fewer imbalances offsets the increase due to variation in number of treated. Threshold blocking then leads to the lowest variance.

Finding optimum

The threshold blocking problem is harder than the fixed-sized problem. Specifically, it is NP-hard and optimal blockings cannot be found except for very small samples. **Near-optimal threshold blockings are, however, possible to derive quickly [3].**

Many more blockings!

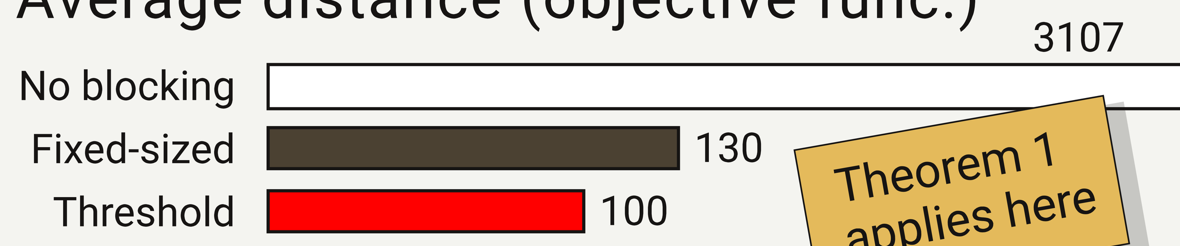
Number of unique blockings

Units	Fixed-sized	Threshold
10	945	17,722
12	10,395	580,317
14	135,135	24,011,157
16	2,027,025	1,216,070,380
18	34,459,425	73,600,798,037
20	654,729,075	5.2×10^{12}

Performance

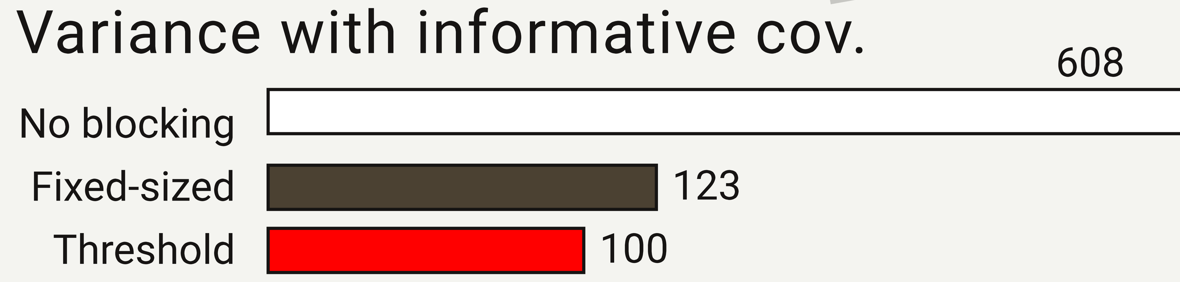
- Simulations with 32 units to illustrate the trade-off.
- One covariate, uniformly distributed on [-5, 5].
- Model 1: $E[y(1)|x] = 2x^2$, $E[y(0)|x] = 1.7x^2$.
- Model 2: $E[y(1)|x] = E[y(0)|x] = 0$.

Average distance (objective func.)



Theorem 1 applies here

Variance with informative cov.



Variance with uninformative cov.