Haphazard Intentional Allocation and Rerandomization to Improve Covariate Balance in Experiments

Marcelo S. Lauretto *, **Rafael B. Stern**†,
Kari L. Morgan‡, **Margaret H. Clark**$^+$, **Julio M. Stern***

* Universidade de São Paulo
‡ Penn State University
† Universidade Federal de São Carlos
$^+$ University of Central Florida

13th EBEB, Feb. 22-25/2016, Belo Horizonte, Brazil
36th MaxEnt, Jul. 10-15/2016, Ghent, Belgium
1st LACSC, Jul. 22-24/2016, Gramado, Brazil

Introduction I

- In randomized experiments, a simple random allocation can yield groups that differ meaningfully with respect to a given covariate. Furthermore, it is unfeasible to control the allocation with respect to more than a moderate number of covariates.
- Morgan and Rubin (2012, 2015) propose an approach based on *Rerandomization* (repeated randomization) to ensure that the final allocation obtained is well balanced.
- Levels of the Rerandomization method:
  1. Lower level: Random samplings for obtaining proposed allocations (Guarantees stochastic behavior of proposed allocations)
  2. Upper level: Rejection of proposals that do not satisfy balance criteria ("Optimizes" balance of final allocation)
- However, despite the benefits of the Rerandomization method, it has an exponential computational cost in the number of covariates (for fixed balance constraints).

Introduction II

- We propose the use of *Haphazard Intentional Allocation*, an alternative allocation method based on optimal balance of the covariates extended by random noise, see Lauretto et al. (2012).
- Similarly to the allocation process in Morgan and Rubin (2012), our method can be divided into a randomization and an optimization step.
  1. Randomization step: consists of creating new (artificial) covariates according to a specified distribution.
  2. Optimization step: consists of finding the allocation that (approximately) minimizes a linear combination of:
     - the imbalance in the original covariates; and
     - the imbalance in the artificial covariates.

Haphazard intentional allocation I

- Let X denote the covariates of interest.
  - X: matrix in $\mathbb{R}^{n \times d}$, where n is the number of individuals to be allocated and d is the number of covariates of interest.
- An allocation consists of assigning to each individual a group, treatment or arm index, $g \in \mathcal{G} = \{0, 1, 2, \ldots\}$.
- We represent an allocation by w, a $1 \times n$ vector in $\mathcal{G}^n$.
- Our goal is to generate an allocation with a low value for a specified inbalance loss function, $L(w, X)$.
- The Haphazard Intentional Allocation consists of finding the approximate minimum of $L(w, [X, Z])$, where Z is a matrix containing random noise.

Haphazard intentional allocation II

- Let Z be an artificially generated matrix in $\mathbb{R}^{n \times k}$, with elements that are independent and identically distributed according to the standard normal distribution.
- For a given tuning parameter, $\lambda \in [0, 1]$, the Haphazard Intentional Allocation finds a feasible allocation, w^* minimizing

$$\begin{aligned} w^* &= \arg\min_{w \in \mathcal{G}^n} L(\lambda, w, X, Z) \\ &= \arg\min_{w \in \mathcal{G}^n} (1 - \lambda) L(w, X) + \lambda L(w, Z). \end{aligned}$$

- λ: controls the amount of perturbation that is added to the original loss function, $L(w, X)$.
  - $\lambda = 0 \Rightarrow w^* =$ deterministic minimizer of $L(w, X)$;
  - $\lambda = 1 \Rightarrow w^* =$ minimizer of the unrelated random loss, $L(w, Z)$.
  - Intermediate values of λ render intermediary characteristics.
- From now on, we consider the case of two groups, $\mathcal{G} = \{0, 1\}$, and Normal distributed random variables.

Haphazard intentional allocation III

- Morgan and Rubin (2012) discusses the case in which the loss function is based on the Mahalanobis distance between the covariates of interest in each group.
- In order to define this loss function, let A be an arbitrary matrix in $\mathbb{R}^{n\times d}$. Furthermore, define $\widetilde{A} := A\,L$, where L is the lower triangular Cholesky factor: $\text{Cov}(A)^{-1} = L\,L^t$, see [1].
- For an allocation w, let a^1 and a^0 denote the averages of each column of $\widetilde{A}$ over individuals allocated to, respectively, groups 1 and 0. That is,

$$a^1 := \frac{w}{n_1}\widetilde{A} \quad \text{and} \quad a^0 := \frac{(\mathbb{1}-w)}{n_0}\widetilde{A}, \text{ where } \begin{cases} n_1 = w^t\,\mathbb{1} \\ n_0 = (\mathbb{1}-w)^t\,\mathbb{1} \end{cases}$$

- The Mahalanobis loss between the groups is computed as:

$$M(w, A) = \sqrt{n_1\,n_0/n}\;\|a^1 - a^0\|_2 \qquad (1)$$

Haphazard intentional allocation IV

- We want to allocate a fixed number of individuals to each group, that is, $w^t \mathbb{1} = n_1$ and $(\mathbb{1} - w)^t \mathbb{1} = n_0 = n - n_1$.
- We can take all these restrictions into consideration by choosing a haphazard intentional allocation with minimal Mahalanobis loss function according to the following optimization problem:

$$\begin{aligned} \text{minimize}(w) \quad & M(\lambda, w, X, Z) \\ & = \lambda M(w, Z) + (1-\lambda) M(w, X) \\ \text{such that} \quad & w^t \mathbb{1} = n_1 \\ & w \in \{0,1\}^n \end{aligned} \tag{2}$$

- This is a mixed-integer *Quadratic Programming* problem, that is difficult to solve relative to the mixed-integer *Linear Programming*.

Haphazard intentional allocation V

- Hence, we use the following *Linear Programming* approximation, based on the *hybrid norm*:

$$H(w, A) = \|a^1 - a^0\|_1 \; + \; \sqrt{d}\|a^1 - a^0\|_\infty .$$

  The hybrid norm is a surrogate loss function for the quadratic norm, based on the extreme cases of the L_p norms for $p = 1$ and $p = \infty$, see [12].

- Furthermore, the resulting optimization problem has the form of Linear Programming:

$$\begin{array}{ll} \text{minimize}(w) & H(\lambda, w, X, Z) \\ & = \lambda H(w, Z) + (1 - \lambda) H(w, X) \\ \text{such that} & w^t \, \mathbb{1} = n_1 \\ & w \in \{0, 1\}^n \end{array} \tag{3}$$

Numerical Experiments I

- In order to perform a haphazard intentional allocation, it is necessary to choose a tuning parameter, λ. We explore the trade-off between randomization and optimization into a grid chosen for callibration convenience:
  - $r = 0.1/0.9;\ \lambda_i^* = 2^{i-4} r / \left[1 + 2^{i-4} r\right],\ i = 1 \ldots 7;$
  - $\lambda_i = \lambda_i^* / \left[\lambda_i^* (1 - k/d) + k/d\right]$.
- This case study is based on the dataset of Shadish et al. (2008), the same dataset used in Morgan and Rubin (2012, 2015), consisting of 24 random covariates.
- The new Haphazard Intentional Allocation method and the Rerandomization method of Morgan and Rubin (2012) were implemented using the R programming language and Gurobi optimization solver [3]. These routines ran on a 12-core Intel i7-4930K 3.4GHz machine.

Numerical Experiments II

- Each method (Haphazard and Rerandomization) ran under a budget of 5, 10, 20, 60, 300 and 900 seconds per allocation, running alone on a single core.
- For each point of the exploration grid, λ_i and time budget, 500 allocations were generated, using different noise inputs, in order to obtain consistent performance measures.
- Table 1 presents the median of the Mahalanobis loss function (on the original data, that is, $M(w, X)$) for the resulting allocations yielded by:
  - The Haphazard Intentional Allocation method optimizing the hybrid norm on the extended data, $H(\lambda, w, X, Z)$;
  - The fixed-time Rerandomization method; and
  - Pure randomization.

Numerical Experiments III

Table 1: Median Mahalanobis loss function for each λ_i (Haphazard) and time budget for each method.

	5s	10s	20s	60s	300s	900s
Hap. $\lambda^* = 0.014$	0.036	0.034	0.033	0.030	0.026	0.024
Hap. $\lambda^* = 0.027$	0.037	0.034	0.033	0.031	0.027	0.024
Hap. $\lambda^* = 0.053$	0.038	0.035	0.034	0.032	0.027	0.025
Hap. $\lambda^* = 0.100$	0.039	0.036	0.035	0.033	0.028	0.026
Hap. $\lambda^* = 0.182$	0.041	0.038	0.037	0.035	0.030	0.028
Hap. $\lambda^* = 0.308$	0.044	0.042	0.040	0.038	0.033	0.030
Hap. $\lambda^* = 0.471$	0.048	0.045	0.044	0.041	0.035	0.032
Rerandomization	0.226	0.217	0.210	0.198	0.184	0.174
Pure randomization	0.458					

Numerical Experiments IV

- Table 1 suggests the following conclusions:
  - The larger the time budget, the smaller the median value of the loss function $M(w, X)$.
  - The smaller the value of λ, less noise is added to the optimization problem and, therefore, the smaller the median value of the loss function $M(w, X)$.
  - Choosing $\lambda^* = 0.1$, Haphazard Intentional Allocation obtains a median Mahalanobis loss that is at least 6 times smaller than when using the fixed-time Rerandomization method.
- Figures 1a, 1b illustrate the difference in covariate balance between Haphazard ($\lambda^* = 0.1$), Rerandomization and pure random allocations.

Numerical Experiments V

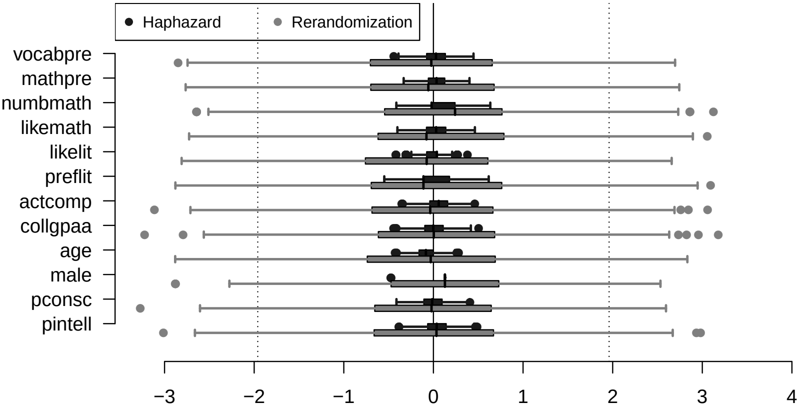

Figure 1a. Difference between covariate averages, 900 secs/allocation.

Numerical Experiments VI

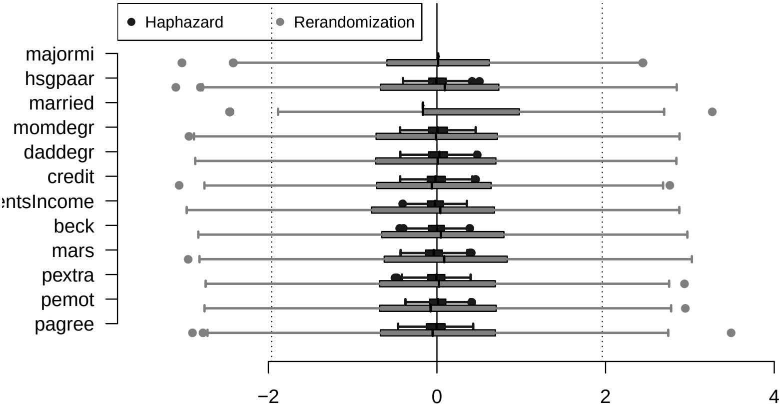

Figure 1b. Difference between covariate averages, 900 secs/allocation.

Numerical Experiments VII

- Table 2 presents the 95% percentile (over all $n^2/2 - n$ pairs of individuals) of the Yule coefficient (computed for each pair of individuals over the 500 allocations).
- Yule coefficient measures how often the individuals under consideration are allocated to the same group.
- Pure random allocation is the effective benchmark for lowest Yule coefficient.

Numerical Experiments VIII

Table 2: 95% percentile of the Yule correlation between allocations for each allocation procedure and time budget.

	5s	10s	20s	60s	300s	900s
Hap. $\lambda^* = 0.014$	0.315	0.254	0.227	0.176	0.153	0.151
Hap. $\lambda^* = 0.027$	0.313	0.258	0.216	0.172	0.152	0.151
Hap. $\lambda^* = 0.053$	0.350	0.280	0.224	0.171	0.151	0.150
Hap. $\lambda^* = 0.100$	0.203	0.192	0.182	0.161	0.152	0.150
Hap. $\lambda^* = 0.182$	0.229	0.190	0.178	0.158	0.151	0.150
Hap. $\lambda^* = 0.308$	0.230	0.194	0.176	0.159	0.150	0.150
Hap. $\lambda^* = 0.471$	0.266	0.224	0.187	0.158	0.150	0.150
Rerandomization	0.144	0.145	0.146	0.146	0.146	0.146
Pure randomization	.143					

Numerical Experiments IX

- Empirically, fixed-time Rerandomization attains a Yule coefficient comparable to the benchmark of pure random allocation.
- For Haphazard Intentional allocations:
  - In the scope of our experiments, the choice of λ doesn't play a preponderant role concerning the Yule coefficient.
  - Instead, time processing budget seems to be the preponderant factor to achieve low Yule coefficients.
  - With a time budget of $900s$, the Haphazard Intentional Allocation obtains a Yule coefficient 5% higher than simple random allocation.
- Hence, comparing the Haphazard Intentional Allocation method and the fixed-time Rerandomization method, we see that, using $\lambda^* = 0.1$, it is possible to obtain a balance on the covariates that is 500% better (measured by the Mahalanobis loss function), at a cost of only a 5% increase in nonrandom associations (measured by the Yule coefficient).

Numerical Experiments X

- An alternatively interpretation for our experiments is to see them as a proxy for other relevant statistical properties.
- For instance, one might be interested in testing the existence of a causal effect of the group assignment on a given response variable. Ex:
    - For each $j \in \{0, 1\}$, we simulate Y^j as the response variable when all individuals are assigned to group j.
    - We follow the procedure:
        1. $Y_i^0 = \epsilon_i + \sum_j \frac{X_{i,j} - \overline{X}_{\bullet,j}}{\mathrm{Var}(X_{\bullet,j})}$, where $\epsilon \sim N(0, \mathbb{I})$.
        2. $Y_i^1 = Y_i^0 + \tau$.

Numerical Experiments XI

- Figure 2 illustrates the difference of power in the allocations obtained by the Haphazard and the Rerandomization procedures for a permutation test for the hypothesis $\tau = 0$.
- The tests obtained using the Haphazard Intentional Allocation method are uniformly more powerful over τ than the ones obtained using the Rerandomization method.
- Figure 3 shows that the difference in power between these allocation procedures can be as high as 0.7 (at $\tau = .4$).

Numerical Experiments XII

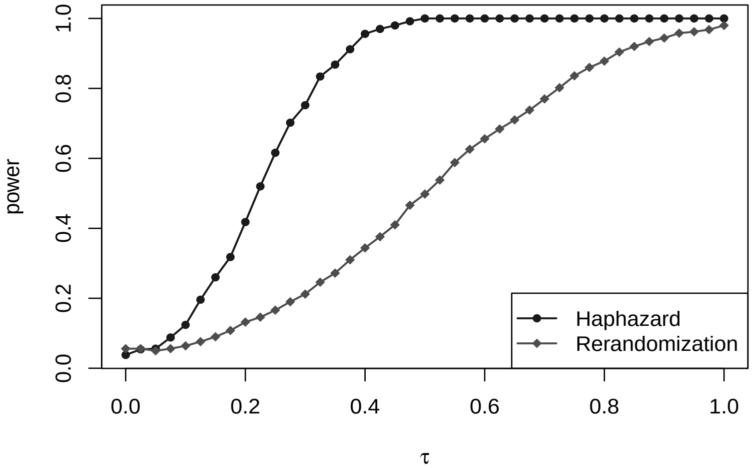

Figure 2. Power curves for each allocation procedure for testing $\tau = 0$ using a permutation test.

Numerical Experiments XIII

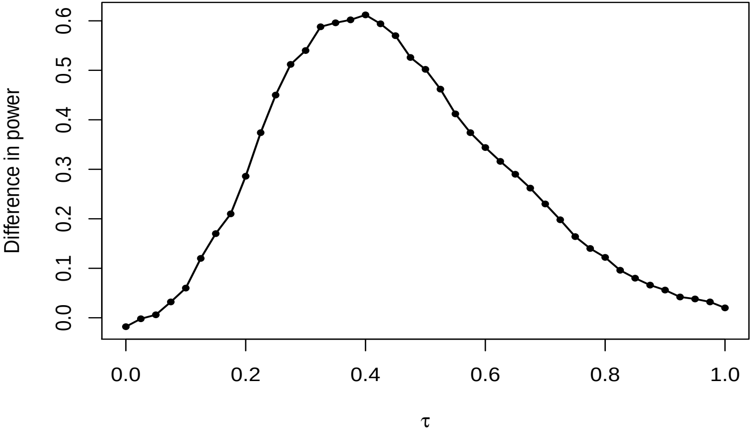

Figure 3. Difference between power curves of Haphazard and Rerandomization Allocations for testing $\tau = 0$ using a permutation test.

Future Research

- Explore the use of the Haphazard Intentional Allocation method and the Rerandomization method in applied problems in the field of:
    - Clinical trials;
    - Jurimetrics.
- Explore the use of alternative surrogate Loss functions for balance performance, such as CVaR norms, Deltoidal norms and Block norms [10, 2, 13].

References I

[1] G.H. Golub and C.F. Van Loan. Matrix Computations. JHU Press, 2012.

[2] J. Y. Gotoh and S. Uryasev. Two pairs of polyhedral norms versus l_p-norms: proximity and applications in optimization. Mathematical Programming A, 156, 391–431, 2016.

[3] Gurobi Optimization Inc. gurobi: Gurobi Optimizer 6.5 interface. URL http://www.gurobi.com, 2015.

[4] M. S. Lauretto, F. Nakano, C. A. B. Pereira, J. M. Stern Intentional Sampling by Goal Optimization with Decoupling by Stochastic Perturbation. *AIP Conference Proceedings, 1490, 189-201*, 2012.

[5] R. F. Love, J. G. Morris and G. O. Wesolowsky. Facilities location. Chapter 3:51–60, 1988.

[6] R. K. Martin. Large scale linear and integer optimization: a unified approach. Springer Science & Business Media, 2012.

[7] K. L. Morgan and D. B. Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics* 40 (2), 1263–1282, 2012.

References II

[8] K. L. Morgan and D. B. Rubin. Rerandomization to Balance Tiers of Covariates. *JASA*, 110, 512, 1412–1421, 2015.

[9] B. A. Murtag. Advanced linear programming: computation and practice. McGraw-Hill International Book, 1981

[10] K. Pavlikov and S. Uryasev. CVaR norm and applications in optimization. Optimization Letters 8, 1999–2020, 2014.

[11] W. R. Shadish, M. R. Clark, P. M. Steiner. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association* 103(484), 1334–1344, 2008.

[12] J. E. Ward and R. E. Wendell. Technical Note-A New Norm for Measuring Distance Which Yields Linear Location Problems. *Operations Research* 28 (3-part-ii), 836–844, 1980.

[13] J. E. Ward and R. E. Wendell. Using Block Norms for Location Modeling. *Operations Research*, 33(5), 1074–1090, 1985.