How to Design Big Comparative Studies?

Feifang Hu

George Washington University

Joint work with Yichen Qin, Fan Wang, Wei Ma and Yang Li
France, May1, 2018

Outline

1. Introduction and Background
2. Covariate-Adaptive Randomization via Mahalanobis Distance (CAM)
3. Causal Inference under CAM and Its Properties
4. Real Data Examples
5. Conclusions

Outline

1. Introduction and Background

2. Covariate-Adaptive Randomization via Mahalanobis Distance (CAM)

3. Causal Inference under CAM and Its Properties

4. Real Data Examples

5. Conclusions

Randomization and Covariate Imbalance

- Randomization: an essential tool for evaluating treatment effect.
- Traditional randomization methods (e.g., complete randomization (CR)): unsatisfactory, **unbalanced** prognostic or baseline covariates.

"Most of experimenters on carrying out a random assignment of plots will be shocked to find out how far from equally the plots distribute themselves." —Fisher (1926)

Why Covariate Balance?

Advantages of covariate balance:

- Improve accuracy and efficiency of inference.
- Remove the bias and increase the power.
- Increases the interpretability of results by making the units more comparable, enhance the credibility.
- More robust against model misspecification.

Covariate Balance in Causal Inference

When there exists covariate imbalance,

- Difficult to compare across treatment groups.
- Although ex-post adjustments are available (e.g., regression and matching), they are much less efficient than achieving an ex-ante balance.
- Adjustments rely on a nearly correct model, which is difficult to verify.
- Rubin (2008): the greatest possible efforts should be made during the design phase rather than the analysis stage.

Covariate Balance in Clinical Trial and Other Fields

Clinical trialists are often concerned that treatment arms will be unbalanced with respect to key covariates of interest. To prevent this, covariate-adaptive randomization is often employed. Over 50000 covariate-adaptive clinical trials had been reported from 1988-2008 (Taves, 2010).

Some Procedures in literature:

- Stratified permuted block design
- Minimizing procedures: Pocock and Simon's marginal procedure (1975) and Taves (1975).
- Hu and Hu's procedures (Hu and Hu, 2012), etc.

Covariate Balance in Clinical Trial and Other Fields

Some concerns of these methods:

- Only for discrete covariates.
- Not for many covariates.
- Theoretical properties?

What if large p and large n?

p: the number of covariates.
n: sample size, i.e., the number of units.

- The phenomenon of covariate imbalance is exacerbated as p and n increase.
- Ubiquitous in the era of big data.
- Example: the probability of one particular covariate being unbalanced is $\alpha = 5\%$. For a study with 10 covariates, the chance of at least one covariate exhibiting imbalance is $1 - (1 - \alpha)^p = 40\%$.

Example

The Project GATE (Growing America Through Entrepreneurship), sponsored by the U.S. Department of Labor, was designed to evaluate the impact of offering tuition-free entrepreneurship training services (GATE services) on helping clients create, sustain or expand their own business. (https://www.doleta.gov/reports/projectgate/)

The cornerstone is complete randomization. Members of the treatment group were offered GATE services; members of the control group were not.

- $n = 4,198$ participants
- $p = 105$ covariates

Rerandomization (RR)

Morgan and Rubin (2012) proposed rerandomization.

(1) Collect covariate data.

(2) Specify a balance criterion, $M < a$, i.e., threshold on the Mahalanobis distance,

$$M = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T [\mathrm{cov}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2),$$

where $\bar{\boldsymbol{x}}_1$ and $\bar{\boldsymbol{x}}_2$ are the sample means for treatment groups.

(3) Randomize the units using the complete randomization (CR).

(4) Check the balance criterion, $M < a$.

- If satisfied, go to Step (5); otherwise, return to Step (3).

(5) Perform the experiment using the final randomization obtained in Step (4).

Rerandomization (RR)

Pros:

- Desirable properties for causal inference:
  - Reduction in variance of estimated treatment effect.
- Work well with a few covariates.

Cons:

- No covariate information is used.
- Incapable to scale up for massive data.
- As p increases, the probability of acceptance $p_a = P(M < a)$ decreases, causing the RR to remain in the loop for a long time.

We Propose:

Covariate-adaptive randomization via Mahalanobis distance (CAM):

- Adaptive.
- Sequential.
- Capable for large p and large n.
- Better covariate balance.
- Less computational time.
- Optimality: the minimum asymptotic variance of estimated treatment effect in linear regressions.

Outline

1. Introduction and Background
2. Covariate-Adaptive Randomization via Mahalanobis Distance (CAM)
3. Causal Inference under CAM and Its Properties
4. Real Data Examples
5. Conclusions

Covariate-Adaptive Randomization via Mahalanobis Distance (CAM)

$\boldsymbol{x}_i \in \mathbb{R}^p$: covariate of the i-th unit.

$T_i \in \{1, 0\}$: treatment assignment of the i-th unit.

- $T_i = 1$: treatment 1.
- $T_i = 0$: treatment 2.

$i = 1, ..., n$

Covariate-Adaptive Randomization via Mahalanobis Distance (CAM)

(1) Use the new defined Mahalanobis distance

$$M(n) = 0.25(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^T [\mathrm{cov}(\bar{\boldsymbol{x}})]^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2).$$

(2) Randomly arrange units in a sequence

$$\underbrace{\boldsymbol{x}_1, \boldsymbol{x}_2}_{\textit{1st pair}}, \underbrace{\boldsymbol{x}_3, \boldsymbol{x}_4}_{\textit{2nd pair}}, \underbrace{\boldsymbol{x}_5, \boldsymbol{x}_6}_{\textit{3rd pair}}, \ldots, \boldsymbol{x}_n.$$

(3) Assign the 1st pair, $T_1 = 1,\ T_2 = 0$.

(4) For the next pair, i.e., $2i + 1$-th and $2i + 2$-th units, $(i > 1)$

- (4a) If $T_{2i+1} = 1$ and $T_{2i+2} = 0$, obtain the "potential" $M_i^{(1)}$.
- (4b) If $T_{2i+1} = 0$ and $T_{2i+2} = 1$, obtain the "potential" $M_i^{(2)}$.

Covariate-Adaptive Randomization via Mahalanobis Distance (CAM)

(5) Assign the $(2i+1)$-th and $(2i+2)$-th units by

$$P(T_{2i+1}=1, T_{2i+2}=0|\boldsymbol{x}_{2i}, T_{2i}\ldots) = \begin{cases} q & \text{if } M_i^{(1)} < M_i^{(2)}, \\ 1-q & \text{if } M_i^{(1)} > M_i^{(2)}, \\ 0.5 & \text{if } M_i^{(1)} = M_i^{(2)}, \end{cases}$$

$$P(T_{2i+1}=0, T_{2i+2}=1|\boldsymbol{x}_{2i}, T_{2i}\ldots) = 1 - P(T_{2i+1}=1, T_{2i+1}=0|\boldsymbol{x}_{2i}, T_{2i}\ldots),$$

where

- $0.5 < q < 1$.
- Note: $T_{2i+1} = T_{2i+2} = 0, 1$ is not allowed.

(6) Repeat Steps (4) and (5) until finish.

Covariate-Adaptive Randomization via Mahalanobis Distance (CAM)

- A **smaller** value of $M(n)$ indicates a **better** covariate balance.
- $q = 0.75$. More discussion in Hu and Hu (2012).
- Units are not observed sequentially; however, we allocate them sequentially (in pairs).
- Better covariate balance.
- $n!$ different possible sequences. Similar performance.

Properties of CAM

Theorem 1

Under CAM, suppose $\mathbf{x}_i$ is i.i.d. multivariate normal; then

$$M(n) = O_p(n^{-1}).$$

Note:

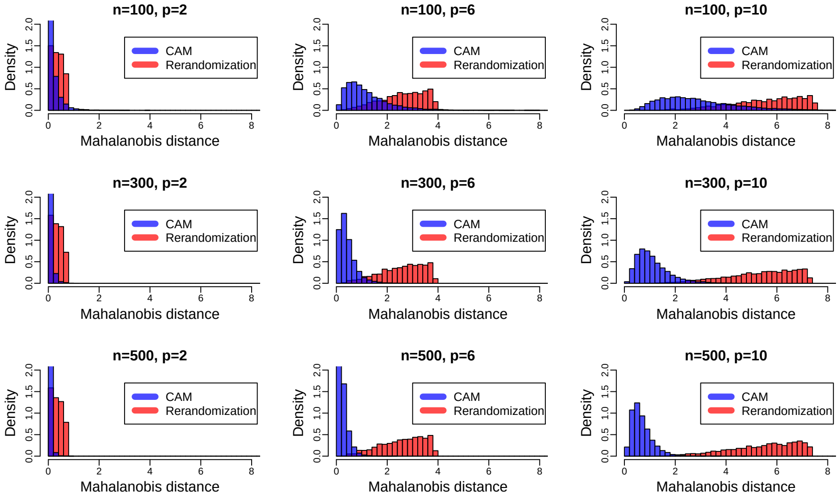
- Under CR, $M_{\text{CR}}(n) \sim \chi^2_{df=p}$, a stationary distribution of a Chi-square distribution with p degrees of freedom, regardless of n.
- Under RR, $M_{\text{RR}}(n) \sim \chi^2_{df=p} | \chi^2_{df=p} < a$, a stationary distribution of a Chi-square distribution with p degrees of freedom conditional on $M_{\text{RR}}(n) < a$, regardless of n.
- Under CAM, $M(n) \to 0$ at the rate of $1/n$.
  - More units, better balance.
  - Advantages of CAM in large n.

Properties of CAM

As p increases,

- Under CR, the stationary distribution becomes flatter, poorer covariate balance.
- Under RR, the stationary distribution becomes flatter, poorer covariate balance.
- Under CAM, $M(n) \to 0$ at the rate of $1/n$, regardless of p.
  - The effect of p on $M(n)$ is less severe than CR and RR.

CAM vs. Rerandomization: Mahalanobis Distance



n=100, p=2

n=100, p=6

n=100, p=10

n=300, p=2

n=300, p=6

n=300, p=10

n=500, p=2

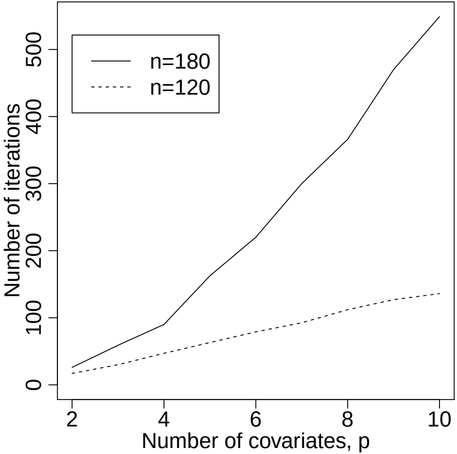
n=500, p=6

n=500, p=10

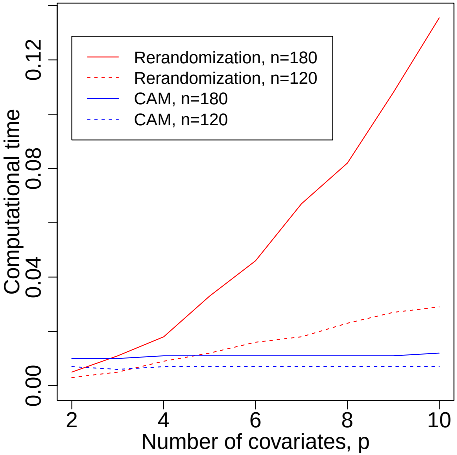
In the Figure

- At the fixed p.
  - Blue histogram shrinks to 0. Red is unchanged.
  - Advantage of CAM over RR.
- At the fixed n.
  - Overlap between the blue and red histograms becomes smaller as p increases.
  - Advantage of CAM over RR.

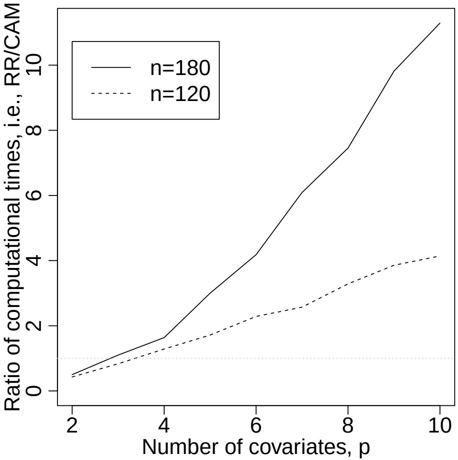
CAM vs. Rerandomization: Computational Iterations



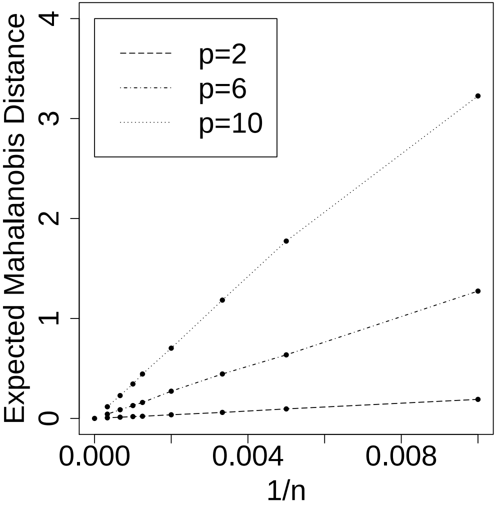
CAM vs. Rerandomization: Computational Time



CAM vs. Rerandomization: Ratio of Computational Times



Convergence Rate of $M(n)$ under CAM



Outline

1. Introduction and Background
2. Covariate-Adaptive Randomization via Mahalanobis Distance (CAM)
3. Causal Inference under CAM and Its Properties
4. Real Data Examples
5. Conclusions

Framework

A framework similar to Morgan and Rubin (2012).

- The observed outcome y_i, $i = 1, ..., n$, for each unit.
- Let $y_i(T_i)$ represents the potential outcome of the i-th unit under the treatment T_i.
- $y_i = y_i(1)T_i + y_i(0)(1 - T_i)$.
- The average treatment effect is

$$\tau = \frac{\sum_{i=1}^{n} y_i(1)}{n} - \frac{\sum_{i=1}^{n} y_i(0)}{n}.$$

- The fundamental problem in causal inference: only observe $y_i(T_i)$ for one particular T_i, therefore, τ cannot be calculated directly.

Estimate without Adjustment for Imbalance

A natural estimate, $\hat{\tau}$, **without** adjustment for imbalance:

$$\hat{\tau} = \frac{\sum_{i=1}^{n} T_i y_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n}(1 - T_i) y_i}{\sum_{i=1}^{n}(1 - T_i)},$$

- $\hat{\tau}$ cannot cope with imbalance in covariates.
- Example: estimate the drug effect using treatment groups with predominately male and female patients. Cannot remove the gender effect.

Estimate with Adjustment for Imbalance

Another estimate, $\tilde{\tau}$, **with** adjustment for imbalance:

- Outcome variable is assumed to follow

$$y_i = \mu_1 T_i + \mu_2(1 - T_i) + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i.$$

- Suppose

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \widetilde{\boldsymbol{T}} = \begin{bmatrix} T_1 & 1 - T_1 \\ \vdots & \vdots \\ T_n & 1 - T_n \end{bmatrix}, \boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \widetilde{\boldsymbol{X}} = [\widetilde{\boldsymbol{T}}; \boldsymbol{X}].$$

The OLS estimate of $\boldsymbol{\beta}^* = (\mu_1, \mu_2, \beta_1, \ldots, \beta_p)^T$ is

$$\hat{\boldsymbol{\beta}}^* = (\widetilde{\boldsymbol{X}}^T \widetilde{\boldsymbol{X}})^{-1} \widetilde{\boldsymbol{X}}^T \boldsymbol{Y}.$$

- Let $\boldsymbol{L} = (1, -1, 0, \ldots, 0)^T \in \mathbb{R}^{p+2}$ and define

$$\tilde{\tau} = \boldsymbol{L}^T \hat{\boldsymbol{\beta}}^*,$$

Framework

- $\hat{\tau}$ is equivalent to $\texttt{lm(}\boldsymbol{Y} \sim \widetilde{\boldsymbol{T}}\texttt{)}$.
- $\tilde{\tau}$ is equivalent to $\texttt{lm(}\boldsymbol{Y} \sim \widetilde{\boldsymbol{T}} + \boldsymbol{X}\texttt{)}$.

Randomization method	Working model for estimating τ	
	$\texttt{lm(}\boldsymbol{Y} \sim \widetilde{\boldsymbol{T}}\texttt{)}$	$\texttt{lm(}\boldsymbol{Y} \sim \widetilde{\boldsymbol{T}} + \boldsymbol{X}\texttt{)}$
Complete Randomization	$\hat{\tau}_{\text{CR}}$	$\tilde{\tau}_{\text{CR}}$
Rerandomization	$\hat{\tau}_{\text{RR}}$	$\tilde{\tau}_{\text{RR}}$
CAM	$\hat{\tau}_{\text{CAM}}$	$\tilde{\tau}_{\text{CAM}}$

Unbiasedness

Unbiasedness of $\hat{\tau}_{\text{CR}}$ and $\hat{\tau}_{\text{RR}}$ has been established.

Theorem 2

Under CAM, we have

$$\mathbb{E}[\hat{\tau}_{\text{CAM}}|\boldsymbol{X}, \text{CAM}] = \tau.$$

Balanced Covariates

In addition to unbiasedness, we have,

Theorem 3

Under CAM, suppose $\boldsymbol{x}_i$ is i.i.d. multivariate normal; then, we have

$$\mathrm{cov}[\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2 | \boldsymbol{X}, \mathrm{CAM}] = u_n \mathrm{cov}[\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2 | \boldsymbol{X}, \mathrm{CR}],$$

where $u_n = \mathbb{E}[M(n) | \boldsymbol{X}, \mathrm{CAM}]$ *and* $u_n = O(n^{-1})$.

- Recall that rerandomization has

$$\mathrm{cov}[\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2 | \boldsymbol{X}, \mathrm{RR}] = v_a \mathrm{cov}[\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2 | \boldsymbol{X}, \mathrm{CR}],$$

where $v_a = \mathbb{E}[M(n) | \boldsymbol{X}, \mathrm{RR}]$ and v_a does not depend on the sample size.

Percent Reduction in Variance (PRIV)

Consider the PRIV for the j-th covariate,

$$100\left(\frac{\mathrm{Var}[\bar{x}_{j,1}-\bar{x}_{j,2}|\boldsymbol{X},\mathrm{CR}]-\mathrm{Var}[\bar{x}_{j,1}-\bar{x}_{j,2}|\boldsymbol{X},\mathrm{CAM}]}{\mathrm{Var}[\bar{x}_{j,1}-\bar{x}_{j,2}|\boldsymbol{X},\mathrm{CR}]}\right),$$

- Rerandomization's PRIV is

  $$100(1-v_a)\%,$$

  which is a constant and independent of the sample size.
- CAM's PRIV is

  $$100(1-u_n)\%,$$

  which converges to 100% as $n \to \infty$.

PRIV for Estimated Treatment Effect

Theorem 4

Under CAM, suppose that y_i and $\boldsymbol{x}_i$ are normally distributed, and that the treatment effect is additive; then, the PRIV of $\hat{\tau}_{\text{CAM}}$ is

$$100\left(\frac{\text{Var}[\hat{\tau}_{\text{CR}}|\boldsymbol{X}, \text{CR}] - \text{Var}[\hat{\tau}_{\text{CAM}}|\boldsymbol{X}, \text{CAM}]}{\text{Var}[\hat{\tau}_{\text{CR}}|\boldsymbol{X}, \text{CR}]}\right) = 100(1 - u_n)R^2,$$

where R^2 is the squared multiple correlation between y_i and $\boldsymbol{x}_i$ within the treatment groups and $u_n = O(n^{-1})$.

- Rerandomization's PRIV is

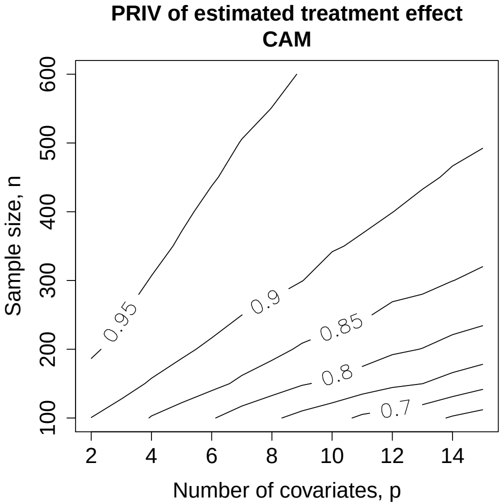
$$100(1 - v_a)R^2\%,$$

  which is a constant and independent of the sample size.
- CAM's PRIV is

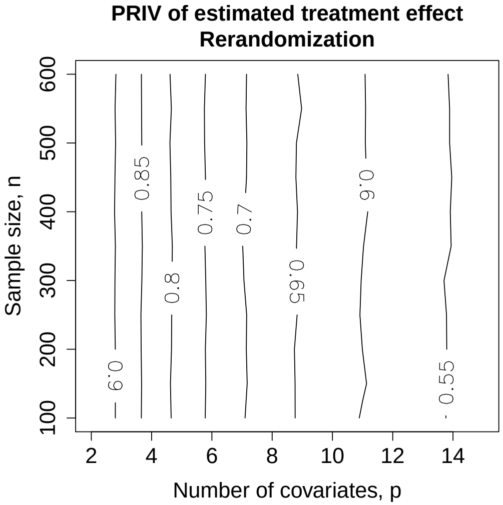
$$100(1 - u_n)R^2\%,$$

  which converges to 100% as $n \to \infty$.

CAM vs Rerandomization



CAM vs Rerandomization



Optimality of CAM

Theorem 5 (Optimal precision)

Suppose y_i truly follows the linear regression model; then, we have

$$\sqrt{n}\big(\hat{\tau}_{\text{CAM}} - (\mu_1 - \mu_2)\big) \xrightarrow{D} N(0, V_1)$$
$$\sqrt{n}\big(\tilde{\tau}_{\text{CAM}} - (\mu_1 - \mu_2)\big) \xrightarrow{D} N(0, V_2)$$
$$\sqrt{n}\big(\tilde{\tau}_{\text{CR}} - (\mu_1 - \mu_2)\big) \xrightarrow{D} N(0, V_3)$$
$$\sqrt{n}\big(\hat{\tau}_{\text{CR}} - (\mu_1 - \mu_2)\big) \xrightarrow{D} N(0, V_4).$$

where $4\sigma_\epsilon^2 = V_1 = V_2 = V_3 < V_4$.

Optimality of CAM

- The precision of $\hat{\tau}_{\mathrm{CAM}}$ is the same as $\tilde{\tau}_{\mathrm{CAM}}$
  - The regression adjustment would not be necessary under the proposed method, because the covariates already would have been balanced sufficiently well.
  - The regression adjustment does not provide any additional benefit.
- The precision of $\hat{\tau}_{\mathrm{CAM}}$ is the same as $\tilde{\tau}_{\mathrm{CR}}$.
  - $\tilde{\tau}_{\mathrm{CR}}$ is considered optimal, therefore, $\hat{\tau}_{\mathrm{CAM}}$ is optimal too.
  - $\tilde{\tau}_{\mathrm{CR}}$ requires to estimate all regression coefficients, whereas $\hat{\tau}_{\mathrm{CAM}}$ is simply the sample mean difference.

Optimality of CAM

Randomization method	Working model for estimating τ		
	$\texttt{lm}(\boldsymbol{Y} \sim \widetilde{\boldsymbol{T}})$		$\texttt{lm}(\boldsymbol{Y} \sim \widetilde{\boldsymbol{T}} + \boldsymbol{X})$
Complete Randomization	Asym. Var.	$>$	Asym. Var.
	$\vee$		$\Vert$
CAM	Asym. Var.	$=$	Asym. Var.

Optimality of CAM: Simulation

- Four continuous covariates
- Outcome:

$$y_i = \mu_1 T_i + \mu_2(1 - T_i) + 1 * x_{i1} + 1 * x_{i2} + 1 * x_{i3} + 1 * x_{i4} + \epsilon_i,$$

where $\mu_1 = 0$, $\mu_2 = 1$, $x_{ij} \sim N(0, 1)$, and $\epsilon_i \sim N(0, 36)$.

- Sample size was $n = 5000$.

Randomization method	Working model for estimating τ	
	$\texttt{lm(}\boldsymbol{Y} \sim \widetilde{\boldsymbol{T}}\texttt{)}$	$\texttt{lm(}\boldsymbol{Y} \sim \widetilde{\boldsymbol{T}} + \boldsymbol{X}\texttt{)}$
Complete randomization	161.1932	144.5853
CAM	145.5646	145.6051

Computational Time

- Why not let $v_a \to 0$ in rerandomization to match the performance of CAM?
- However, this option is infeasible in many cases.

Computational Time

Suppose that

- Cp: time to allocate one additional unit using CAM.
- R: time to allocate one unit using complete randomization.

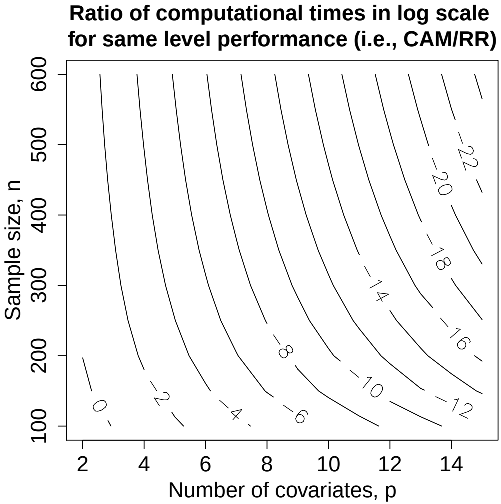
Theorem 6

To achieve the same level of performance, the ratio of average computational times (i.e., CAM/RR) is proportional to

$$\chi^2_{df=p}(a^*)Cp/R,$$

where $\chi^2_{df=p}(\cdot)$ *is the cumulative distribution function of a Chi-square distribution with* p *degrees of freedom, and* a^* *is the root of* $\gamma(p/2, a^*/2)Dp = 2\gamma(p/2+1, a^*/2)n$ *where* $D > 0$ *is a constant and* $\gamma(w,t) = \int_0^t x^{w-1}\exp\{-x\}dx$ *is the incomplete gamma function.*

CAM vs Rerandomization



CAM vs Rerandomization

Sample size n	$p = 2$	4	6	8	10
200	0.9830	0.1084	0.0094	7.492e-04	5.686e-05
400	0.4957	0.0275	0.0012	4.884e-05	1.876e-06
600	0.3312	0.0123	0.0003	9.748e-06	2.510e-07

Outline

1. Introduction and Background
2. Covariate-Adaptive Randomization via Mahalanobis Distance (CAM)
3. Causal Inference under CAM and Its Properties
4. Real Data Examples
5. Conclusions

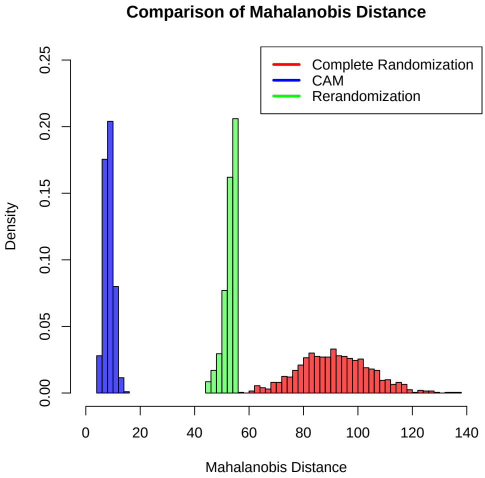
Real Data Example I - Project GATE

- Two treatment groups:
  Treatment: were offered GATE services; control: were not offered GATE services.
- $p = 105$ (covariates obtained from the application packages, 13 continuous and 92 categorical)
- Sample size $n = 3,448$ (out of 4,198 participants from who answered the evaluation survey 6 months after the assignment)
- Original allocation $M = 75.27$, moderate covariate imbalance.
- We repeat the allocation 1,000 times for these participants using CAM, complete randomization and rerandomization.

CAM vs Rerandomization

The Maximum of Malahanobis distances obtained from CAM is 12. If we set the balance criterion for rerandomization to $M < 12$, the probability of acceptance $P_a = P(\chi^2_{df=105} < 12) = 3.4 \times 10^{-31}$, which means nearly impossible for rerandomization to achieve a similar balance level as CAM.

We set $P_a = 2 \times 10^{-5}$ for Rerandomization to have similar computational time with CAM.


Comparison of Mahalanobis Distance
Complete Randomization
CAM
Rerandomization
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
20
40
60
80
100
120
140
Mahalanobis Distance

Estimation

- The outcome variable (0/1): has owned a business within 6 months after assignment or not.
- After the allocation, we simulate the outcome variable according to

$$\operatorname{logit}(P(y_i^{\text{sim}} = 1)) = \hat{\mu}_1 T_i^{\text{sim}} + \hat{\mu}_2(1 - T_i^{\text{sim}}) + x_i^T \hat{\boldsymbol{\beta}} + \epsilon^{\text{sim}},$$

where $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\boldsymbol{\beta}}$ are obtained from fitting regression to original data. ϵ^{sim} is drawn from the residuals of that regression.

Performance Comparison

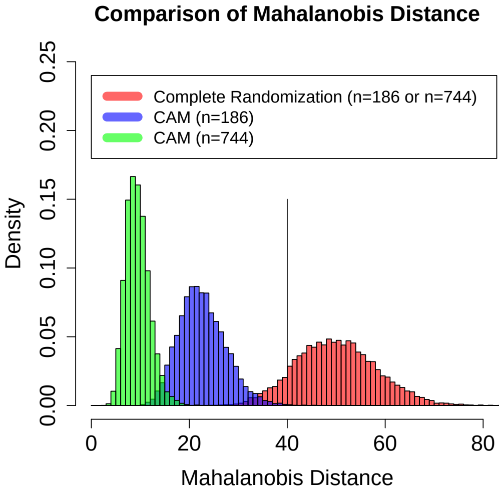
Compare the estimation performance (PRIV) of CAM and rerandomization.

Method	PRIV	u_n or v_a
CAM	17.7%	0.081
Rerandomization	10.5%	0.505

Real Data Example II

- A real data set obtained in a clinical study of a Ceragem massage (CGM) thermal therapy bed, a medical device for the treatment of lumbar disc disease.
- Number of covariates $p = 50$.
- 30 numerical covariates: age, measurements of the patient's current conditions, including lower back pain, leg pain, leg numbness, body examination scores, and magnitudes of pain in shoulders, neck, chest, hip and so on. All are measured on 0-10 scales.
- Sample size $n = 186$.
- Replicate the data four times to have a sample size of $n = 744$.
- Original allocation $M = 57.67$, moderate covariate imbalance.
- We repeat the allocation for these patients using CAM, complete randomization and rerandomization.

CAM vs Rerandomization



Estimation

- After the allocation, we simulate the outcome variable according to

$$y_i^{\text{sim}} = \hat{\mu}_1 T_i^{\text{sim}} + \hat{\mu}_2(1 - T_i^{\text{sim}}) + x_i^T \hat{\boldsymbol{\beta}} + \epsilon^{\text{sim}},$$

  where $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\boldsymbol{\beta}}$ are obtained from fitting regression to original data. ϵ^{sim} is drawn from the residuals of that regression.

- Compare the estimation performance (PRIV and MSE) of CAM, complete randomization, and rerandomization ($M < 30$, $M < 40$).

- Optimal PRIV is 0.33 (R^2 is 0.33).

Performance Comparison

Sample Size	Method	PRIV	MSE	u_n or v_a
$n = 186$	CAM	19.7%	0.081	0.502
	Rerandomization ($M < 30$)	15.1%	0.085	0.562
	Rerandomization ($M < 40$)	12.2%	0.090	0.730
	Complete Randomization	-	0.100	-
$n = 744$	CAM	27.4%	0.018	0.205
	Rerandomization ($M < 30$)	14.6%	0.021	0.556
	Rerandomization ($M < 40$)	10.9%	0.022	0.718
	Complete Randomization	-	0.025	-

Note: The optimal PRIV is 0.33 (i.e., $R^2 = 0.33$).

Outline

1. Introduction and Background
2. Covariate-Adaptive Randomization via Mahalanobis Distance (CAM)
3. Causal Inference under CAM and Its Properties
4. Real Data Examples
5. Conclusions

Conclusions

- A covariate-adaptive randomization (CAM) approach to generate a more balanced treatment allocation, and thus to improve the quality of the subsequent causal inference.
- Allocate units adaptively and sequentially.
- For cases with a large number of covariates or a large number of units, the proposed method exhibits superior performance, with a more balanced randomization and much less computational time.
- The proposed method is proven to be optimal, in that, the estimated treatment effect under the proposed method achieves the minimum asymptotic variance in the linear regression framework.

References I

Ronald A. Fisher. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503513, 1926.

Yanqing Hu and Feifang Hu. Asymptotic properties of covariate-adaptive randomization. *Annals of Statistics*, 40(3):17941815, 2012.

Kari L. Morgan and Donald B. Rubin. Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2): 12631282, 2012.

Donald B. Rubin. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484):13501353, 2008.

Thank You!