Selective Randomization Inference for Adaptive Studies

Tobias Freidling, Zijun Gao, Qingyuan Zhao

University of Chicago 27/10/2023

Overview

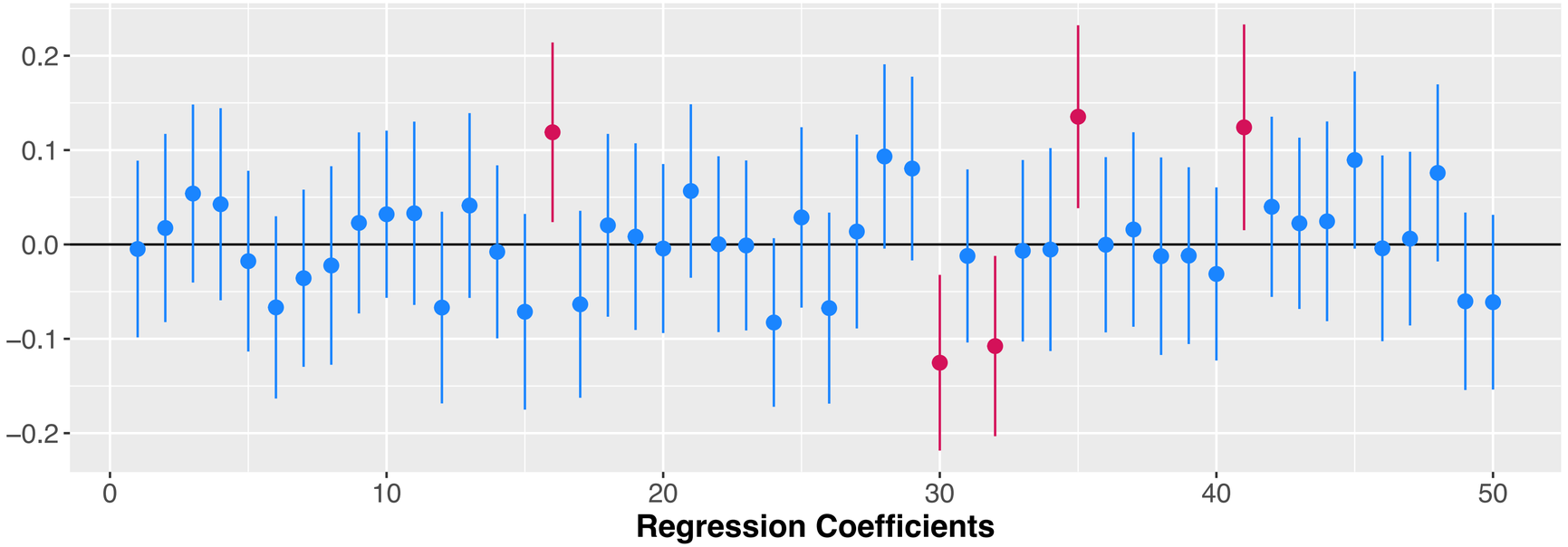
1. Background
    1. Post-selection Inference
    2. Adaptive Clinical Trials
    3. Randomization Inference
2. Adaptive Studies - A Graphical Model
3. Selective Randomization Inference
4. Confidence Intervals and Computation
5. Questions

Post-selection Inference - Toy Example

Linear regression with 50 covariates and $n = 300$: $Y_i = \sum_{i=1}^{50} X_{ij}\,\beta_j + \varepsilon_i, \quad \varepsilon_i \sim N(0,1)$

Task: Find the 5 most influential features and construct 90%-confidence intervals for their regression coefficients.

Ground truth: $\beta_j = 0$ for all $j \in \{1,\ldots,50\}$



Post-selection Inference

- Traditional statistics: model & null hypothesis → data → inference
- Yet, in practice: data → model & null hypothesis → inference
- Solutions:
  - Data splitting (Cox, 1975): potential loss of power, arbitrary splits
  - Selective inference (Lee et al., 2016; Fithian et al., 2017):
    - null hypothesis and model are chosen based on data D via selection rule $S(D)$
    - **Condition on selection event:** $D \mid S(D) = s$ selective distribution
    - Selective p-value: $\mathbb{P}\left(T_s(D) \leq t_{s,obs} \mid S(D) = s\right)$

Adaptive Clinical Trials

- Objectives: determine safe dosage, most effective treatment, responsive subpopulation
- Multiple stages
- Adaptive designs: response adaptive randomisation (RAR), multi-arm multi-stage (MAMS), enrichment trials etc. → **selective recruitment and treatment assignment**
- Analysis of adaptive studies:
  - Well-known problem, e.g. Armitage (1960), Pocock (1977) etc.
  - Specific to a certain parametric model or aggregation of p-values
  - Unaware of post-selection inference literature
- Remark: connection to bandit literature

Randomization Inference - Example

- Study with n participants testing 2 treatments
- Treatment assignment: $W \in \{0,1\}^n$; $\mathbb{P}(W = w)$ chosen by experimenter
- Potential outcomes of participants: $Y(\cdot) = (Y_i(0), Y_i(1))_{i=1}^n$
- Observed outcomes: $Y = Y(W)$ consistency
- Sharp null hypothesis: $Y_i(0) = Y_i(1) \quad \forall i \in [n] \quad \rightarrow$ all potential outcomes $Y(\cdot)$ are known
- Randomization distribution of test statistic T: $T(W, Y(\cdot)) \mid Y(\cdot)$
- Randomization p-value:

$$p = \mathbb{P}^*(\ T(W^*, Y(\cdot)) \leq T(W, Y(\cdot)) \mid W, Y(\cdot)), \quad W^* \overset{D}{=} W \text{ and } W^* \perp\!\!\!\perp W \mid Y(\cdot)$$

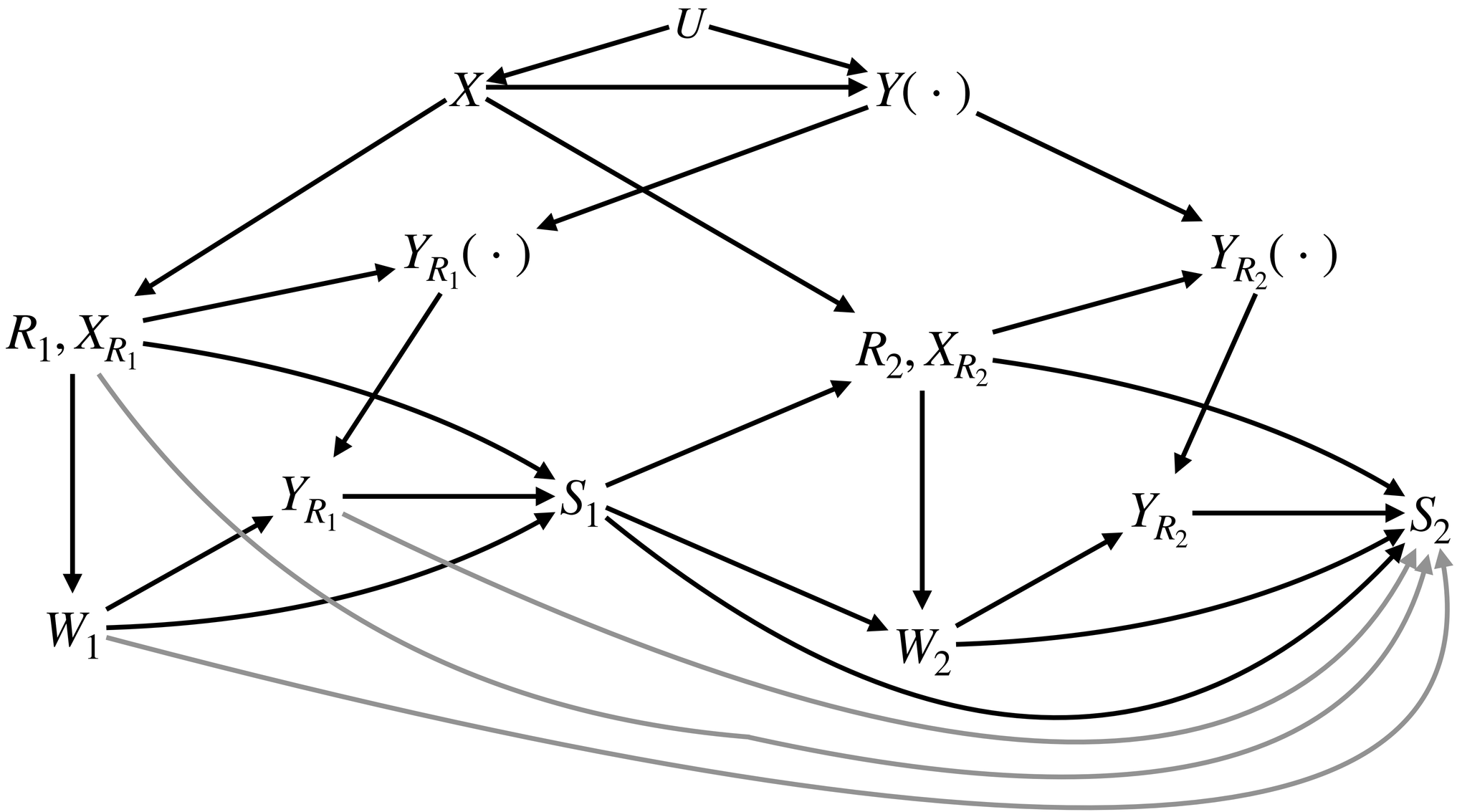
Randomization Inference

- Revival (Zhang & Zhao, 2023) of an old idea (Fisher, 1935)
- Leveraging known treatment assignment probability
- Extensions: conditional randomization test, partially sharp null hypotheses
- **Pros:** no modelling assumptions, arbitrary dependence between units
- **Cons:** uninteresting null hypothesis, computation of p-value

Adaptive Trials - Set-up

- Two-stage adaptive trial
- Potential outcomes $Y(\cdot) = Y_{[n]}(\cdot)$ and covariates $X = X_{[n]}$
- Recruitment for stages I and II: $R_1 \subseteq [n], \quad R_2 \subseteq [n] \backslash R_1, \quad R = R_1 \cup R_2$
- Observed outcomes: Y_{R_1} and Y_{R_2}
- Treatment assignments for stages I and II:
  $W_1 \in \{0, \ldots, L\}^{|R_1|}, \quad W_2 \in \{0, \ldots, L\}^{|R_2|}, \quad W = (W_1, W_2)$
- Summary statistics after stage I and II: $S = (S_1, S_2)$ S_2 models H_0
- Assumptions: Consistency, No interference

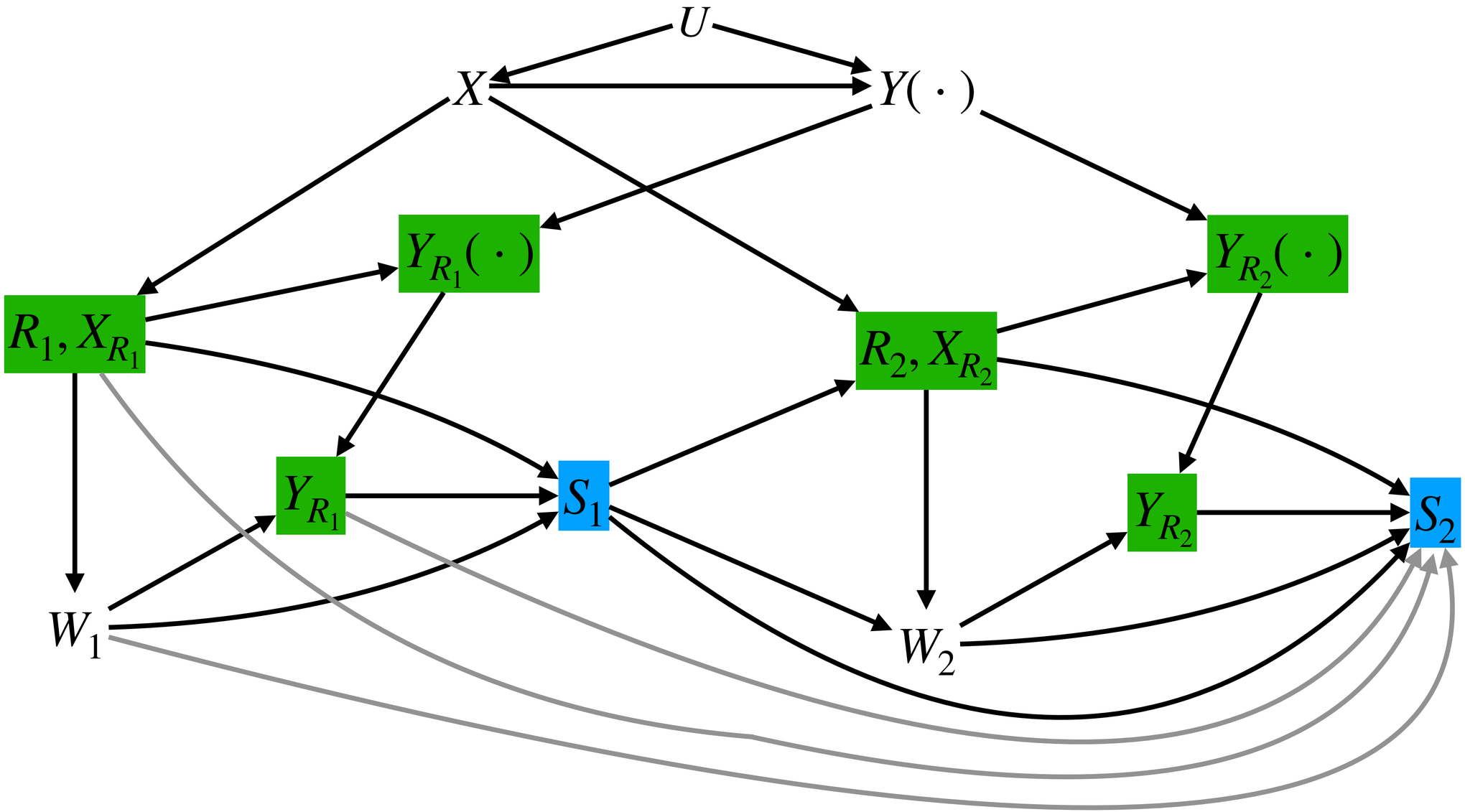
Adaptive Studies - Graphical Model



Selective Randomization Inference

- Selective randomization distribution: $W_1, W_2 \mid R, X_R, Y_R(\cdot), S$

Adaptive Studies - Graphical Model



Selective Randomization Inference

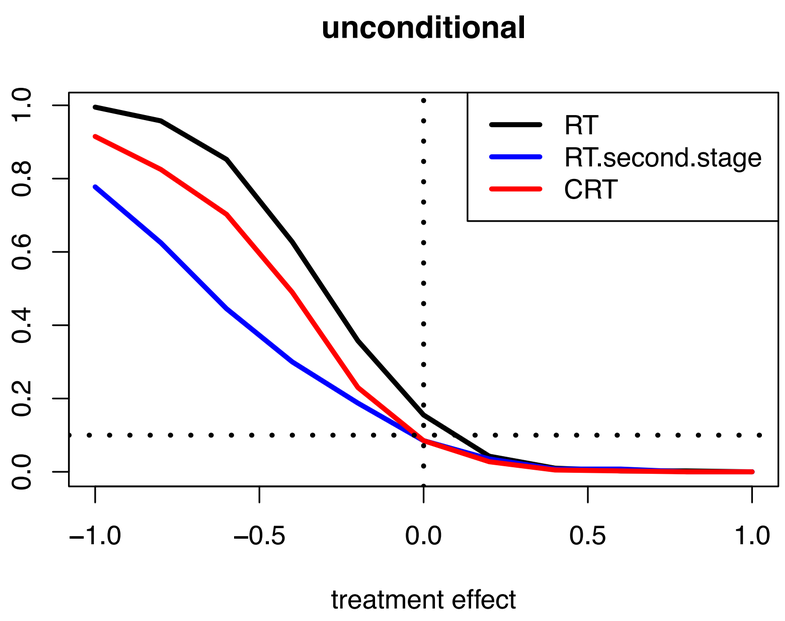
- Selective randomization distribution: $W_1, W_2 \mid R, X_R, Y_R(\cdot), S(W)$
- Selective randomization p-value:
  - Test statistic: $T(W) := T(W, X_R, R, Y_R(\cdot))$
  - $W^* \overset{D}{=} W$ and $W^* \perp\!\!\!\perp W \mid R, X_R, Y_R(\cdot)$
  - $p = \mathbb{P}^*(T(W^*) \leq T(W) \mid W, R, X_R, Y_R(\cdot), S(W^*) = S(W))$
- **Lemma:** p controls the selective type-I error. Simon & Simon (2011): special case
- Remark: Factorization without gray arrows
  $p(w \mid r, x_r, y_r(\cdot), s, h) = p(w_1 \mid r_1, x_{r_1}, y_{r_1}(\cdot), s) \cdot p(w_2 \mid r_2, x_{r_2}, y_{r_2}(\cdot), s, h)$

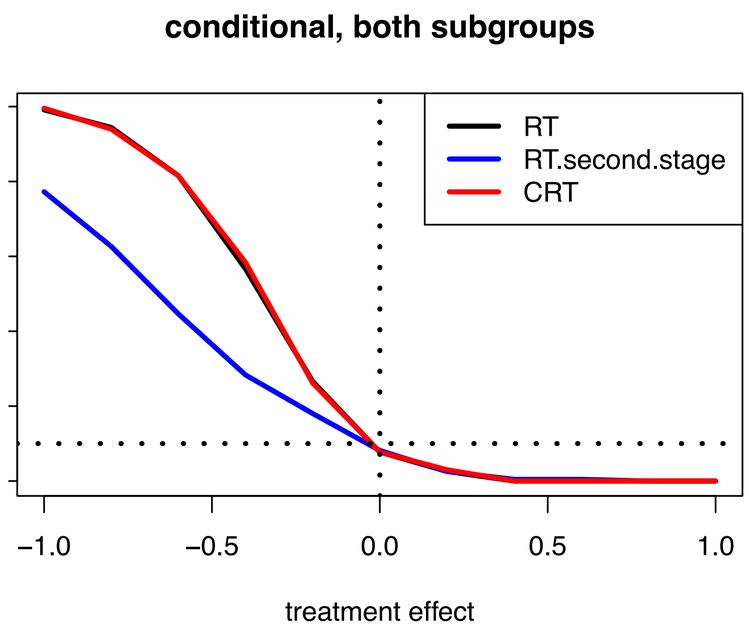
Simulation Study

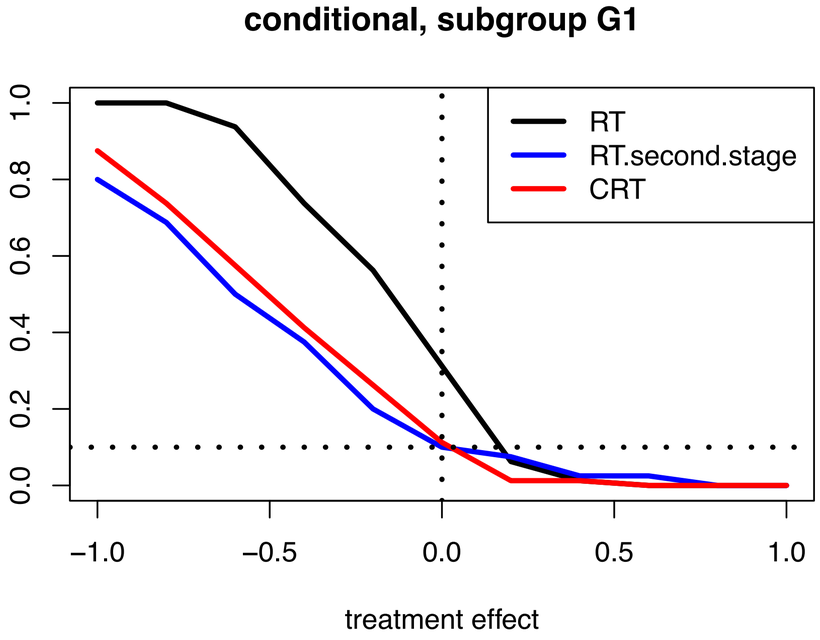
- 2 stages, 2 treatments, 2 groups G_1, G_2
- Potential outcomes: $Y_i(0) = Y_i(1) \sim N(0,1)$ i.i.d.
- First stage: 20 patients per group, $\Delta = \left[\widehat{\text{ATE}}\,(G_1) - \widehat{\text{ATE}}\,(G_2)\right] / \sqrt{2}$
- Selection variable: $S = \begin{cases} \text{recruit 20 from } G_2, & \Delta < \Phi(0.2) \\ \text{recruit 10/10 from } G_1 \text{ and } G_2, & \Phi(0.2) \leq \Delta \leq \Phi(0.8) \\ \text{recruit 20 from } G_1, & \Delta > \Phi(0.8) \end{cases}$

Simulation Study

- Null hypothesis: $Y_i(1) - Y_i(0) = c$ for the selected group(s)
- Test statistic T: Difference in means in selected group(s)

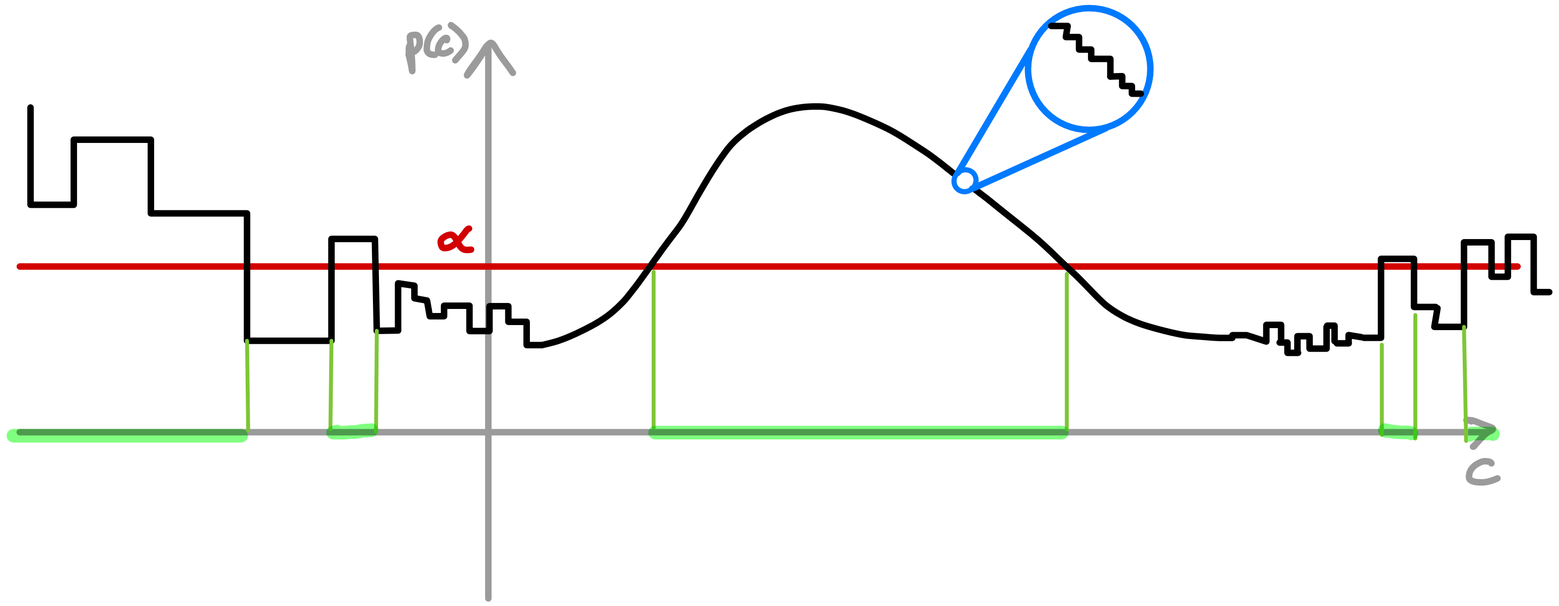






Confidence intervals

- Test collection of null hypotheses $H_0^c : Y_i(0) - Y_i(1) = c$
- P-value curve: $p(c)$; **possibly not uni-modal because of conditioning**
- For large or small effect c: very few feasible treatment assignments
- Remedy: hold-out set of patients that are not used for selection



Computation of p-value

- Monte Carlo approximation: Generate m feasible samples $(w_j^*)_{j=1}^m$, i.e. $S(w) = S(w_j^*)$, and compute

$$\hat{p}_m = \frac{\sum_{i=1}^m \mathbf{1}_{\{T(w_j^*) \leq T(w)\}} \mathbb{P}^*(W^* = w^* \mid R, X_R, Y_R(\,\cdot\,))}{\sum_{i=1}^m \mathbb{P}^*(W^* = w^* \mid R, X_R, Y_R(\,\cdot\,))}$$

- Two methods: Rejection sampling and Markov Chain Monte Carlo (MCMC)
- Ongoing work on convergence guarantees

Summary

- Intersection of post-selection inference, adaptive (clinical) trials and randomization inference
- Graphical model
- Selective randomization p-value
- Construction of selective confidence intervals
- Monte Carlo approximation

Thanks for your attention!

Any Questions?

taf40@cam.ac.uk

References

Armitage, P.M.A. (1960). ‘Sequential medical trials’, Blackwell Scientific Publications.

Cox, D.R. (1975) ‘A note on data-splitting for the evaluation of significance levels’, Biometrika, 62(2), pp. 441–444.

Fisher, R. A. (1935). ‘The design of experiments’, Edinburgh: Oliver & Boyd.

Fithian, W., Sun, D. and Taylor, J. (2017) ‘Optimal Inference After Model Selection’, arXiv:1410.2597

Lee, J.D., Sun, D.L., Sun, Y., Taylor J. (2016) ‘Exact post-selection inference, with application to the lasso’, The Annals of Statistics, 44(3).

Pocock, S. J. (1977). ‘Group sequential methods in the design and analysis of clinical trials’ Biometrika 64(2), pp. 191–199.

Simon, R. and Simon, N.R. (2011) ‘Using randomization tests to preserve type I error with response adaptive and covariate adaptive randomization’, Statistics & Probability Letters, 81(7), pp. 767–772.

Zhang, Y. and Zhao, Q. (2023) ‘What is a Randomization Test?’, Journal of the American Statistical Association, 0(0), pp. 1–15.