Data-Driven Switchback Experiments: Theoretical Tradeoffs and Empirical Bayes Designs

Ruoxuan Xiong*, Alex Chin†, and Sean Taylor†

Emory University* and Motif Analytics†

Motivation

Suppose we run an experiment in Atlanta to test the effect of a new algorithm, e.g., pricing or matching, on a ride-sharing platform

By the end of the experiment, estimate global average treatment effect (GATE)

- Difference in outcomes between when the treatment is employed indefinitely versus when it is absent
- Primary outcome of interest: average conversion rate (fraction of riders requesting ride after checking price)

Question: How to run this experiment?

- Maybe conventional A/B testing, 50% users in old algorithm and 50% in new algorithm?
- Any issue?

Motivation

Suppose we run an experiment in Atlanta to test the effect of a new algorithm, e.g., pricing or matching, on a ride-sharing platform

By the end of the experiment, estimate global average treatment effect (GATE)

- Difference in outcomes between when the treatment is employed indefinitely versus when it is absent
- Primary outcome of interest: average conversion rate (fraction of riders requesting ride after checking price)

Question: How to run this experiment?

- Maybe conventional A/B testing, 50% users in old algorithm and 50% in new algorithm?
- Any issue?

Motivation

Suppose we run an experiment in Atlanta to test the effect of a new algorithm, e.g., pricing or matching, on a ride-sharing platform

By the end of the experiment, estimate global average treatment effect (GATE)

- Difference in outcomes between when the treatment is employed indefinitely versus when it is absent
- Primary outcome of interest: average conversion rate (fraction of riders requesting ride after checking price)

Question: How to run this experiment?

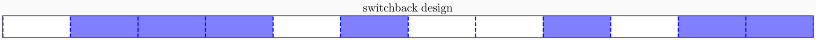
- Maybe conventional A/B testing, 50% users in old algorithm and 50% in new algorithm?
- Any issue?

$\Rightarrow$ Interference between treated and control users!

Solution: Switchback experiment

Use the switchback design to control spillovers from treated to control users

- Randomize at the city level (all users either in treated or control state)
- Experiment lasts for, e.g., two weeks
- Flip a coin, e.g., every 56 minutes, to determine whether the next 56 minutes are treated



dash lines are switching points, and treated intervals are shaded

This paper studies the design of switchback experiments, with the objective of more precisely estimating GATE

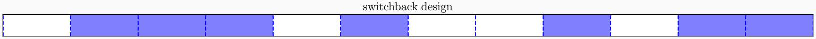
- Flip a coin every 56 minutes, or longer or shorter?
- Randomize coin flipping frequency (interval length)?
- Other considerations?

Question: What factor may affect the design?

Solution: Switchback experiment

Use the switchback design to control spillovers from treated to control users

- Randomize at the city level (all users either in treated or control state)
- Experiment lasts for, e.g., two weeks
- Flip a coin, e.g., every 56 minutes, to determine whether the next 56 minutes are treated



dash lines are switching points, and treated intervals are shaded

This paper studies the design of switchback experiments, with the objective of more precisely estimating GATE

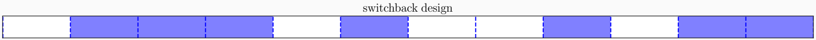
- Flip a coin every 56 minutes, or longer or shorter?
- Randomize coin flipping frequency (interval length)?
- Other considerations?

Question: What factor may affect the design?

Solution: Switchback experiment

Use the switchback design to control spillovers from treated to control users

- Randomize at the city level (all users either in treated or control state)
- Experiment lasts for, e.g., two weeks
- Flip a coin, e.g., every 56 minutes, to determine whether the next 56 minutes are treated



dash lines are switching points, and treated intervals are shaded

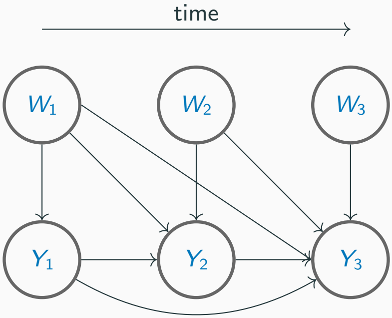
This paper studies the design of switchback experiments, with the objective of more precisely estimating GATE

- Flip a coin every 56 minutes, or longer or shorter?
- Randomize coin flipping frequency (interval length)?
- Other considerations?

Question: What factor may affect the design?

A prominent factor: Carryover effect

Carryover effect: Treatment W_s at earlier times may affect outcomes Y_t at later times $t \geq s$

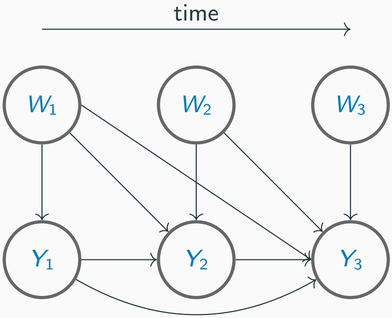


Cumulative effect: The total effect of current and prior treatments on current outcome. Converge to GATE as the treatment duration increases

If the carryover effect is more persistent, switch less frequently (Bojinov, Simchi-Levi, and Zhao 2023; Hu and Wager 2022)

A prominent factor: Carryover effect

Carryover effect: Treatment W_s at earlier times may affect outcomes Y_t at later times $t \geq s$

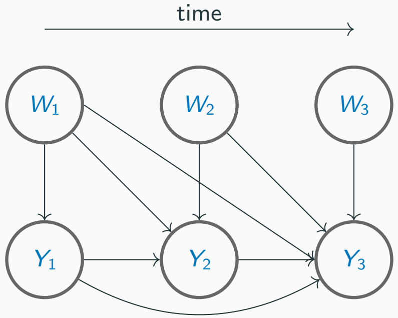


Cumulative effect: The total effect of current and prior treatments on current outcome. Converge to GATE as the treatment duration increases

If the carryover effect is more persistent, switch less frequently (Bojinov, Simchi-Levi, and Zhao 2023; Hu and Wager 2022)

A prominent factor: Carryover effect

Carryover effect: Treatment W_s at earlier times may affect outcomes Y_t at later times $t \geq s$



Cumulative effect: The total effect of current and prior treatments on current outcome. Converge to GATE as the treatment duration increases

If the carryover effect is more persistent, switch less frequently (Bojinov, Simchi-Levi, and Zhao 2023; Hu and Wager 2022)

This paper: Estimate the distribution of carryover effects

Assumptions on carryover effect (e.g., hypothesized maximum duration) are often made when designing switchback experiment

An issue: Carryover effect is unknown ex-ante

Solution: Prior experiments are commonly available and contain information about carryover effects

This paper proposes to

- Estimate cumulative effects from historical experiments
- Obtain empirical distribution as prior distribution of cumulative effects
- Design experiments assuming cumulative effects are drawn from this prior

Illustration of empirical distribution of cumulative effects

Historical experimental data: Data of 149 experiments run across 114 markets with 890 distinct experiment-market pairs between June 2021 and March 2023 on a ride-sharing platform

- These switchback experiments are run for two weeks with a fixed interval length of 56 minutes

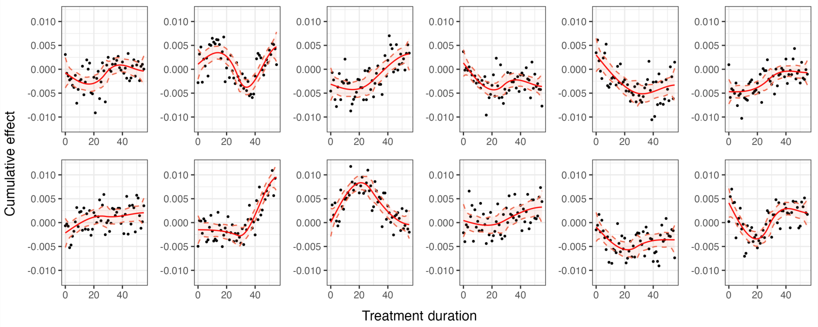
For each experiment-market pair, estimate a **cumulative effect curve** (CEC)

- The cumulative effect of treating one, two, three, ... minutes

Empirical distribution of CECs from prior experiments

12 representative estimated CECs for the treatment duration of $\{1, \cdots, 56\}$ minutes (in black dots) and their smooth curves by natural cubic splines (in red)

⇒ Most CECs are non-monotonic and change signs as treatment duration varies



Other realistic factors affecting the design

In addition to carryover effects, this paper considers **three other realistic factors** that affect the estimation error of GATE and the design, but not considered in the literature

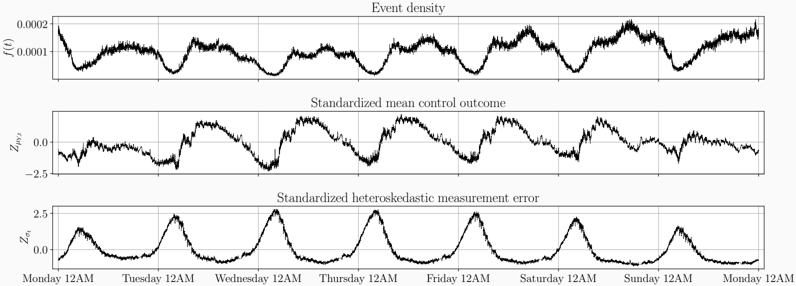
Factor 1: Periodicity in event density and outcome

- An event can be rider checking price
- The outcome can be whether the rider requests the ride

Periodicity in event density and outcome

Event density, mean control outcome, and variance of measurement errors all have a periodic pattern

Event density is higher during peak hours on weekends than during peak hours on weekdays



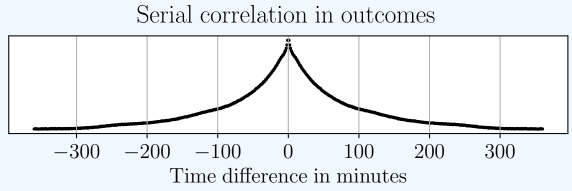
Other realistic factors affecting the design

In addition to carryover effects, this paper considers **three other realistic factors** that affect the estimation error of GATE and the design, but not considered in the literature

Factor 1: Periodicity in event density and outcome

- An event can be rider checking price
- The outcome can whether the rider requests the ride

Factor 2: Heteroskedastic and correlated measurement errors in event outcomes



Other realistic factors affecting the design

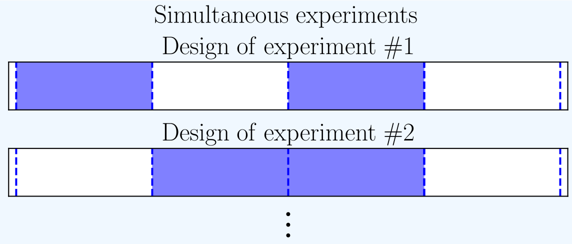
In addition to carryover effects, this paper considers **three other realistic factors** that affect the estimation error of GATE and the design, but have not been considered in the literature

Factor 1: Periodicity in event density and outcome

- An event can be rider checking price
- The outcome can whether the rider requests the ride

Factor 2: Heteroskedastic and correlated measurement errors in event outcomes

Factor 3: Impacts from simultaneous experiments



Other realistic factors affecting the design

In addition to carryover effects, this paper considers **three other realistic factors** that affect the estimation error of GATE and the design, but have not been considered in the literature

Factor 1: Periodicity in event density and outcome

- An event can be rider checking price
- The outcome can whether the rider requests the ride

Factor 2: Heteroskedastic and correlated measurement errors in event outcomes

Factor 3: Impacts from simultaneous experiments

Question: How to design a switchback experiment accounting for all factors?

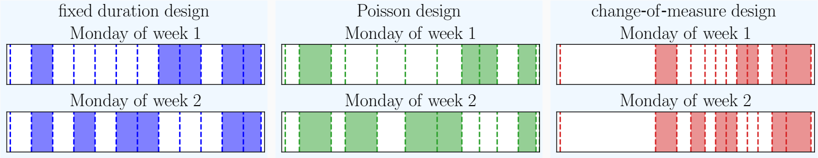
Empirical Bayes design approach to account for all factors

- Phase 1: Analyze historical data and model data generating process
- Phase 2: Estimate the empirical distribution of CECs from prior experiments
- Phase 3: Run synthetic experiments on historical data to compare candidate designs
- Phase 4: Choose the best-performed design in synthetic experiments

- In phase one, event density and periodic pattern are estimated from historical data and used as the input for the design

Various heuristic switchback designs

- Fixed duration design: Fixed interval lengths (status quo design)
- Poisson design: Interval lengths drawn from Poisson distribution
- Change-of-measure design: Fixed event occurrence probabilities



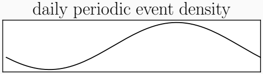


Illustration of various designs and daily periodic event density (dash lines are switching points, and treated intervals are shaded) on Mondays in a two-week experiment

Various heuristic switchback designs (cont.)

- Balanced designs: The treatment assignments in the second week mirror those in the first week, i.e., balance periodicity

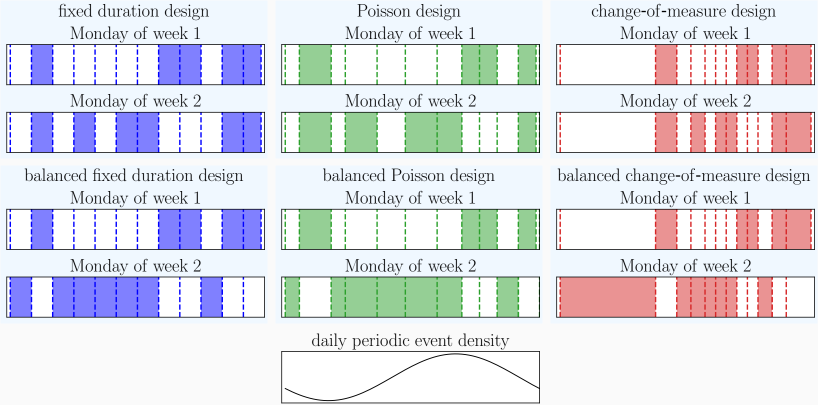


Illustration of various designs and daily periodic event density (dash lines are switching points, and treated intervals are shaded) on Mondays in a two-week experiment

Illustration of empirical Bayes design approach in the case study

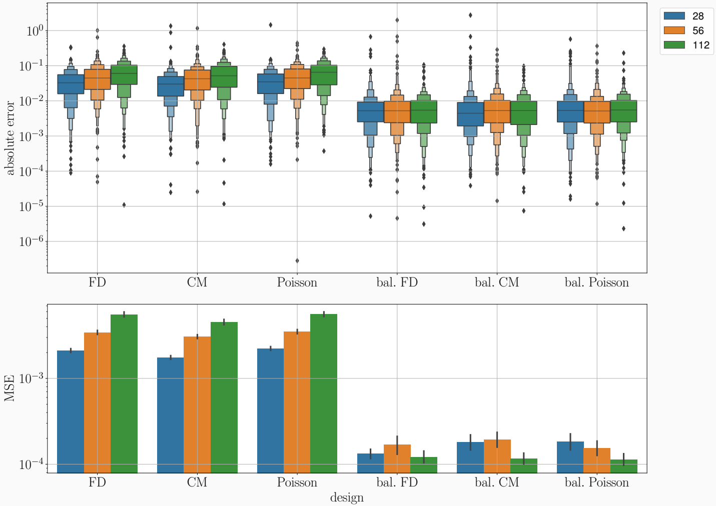
For each of the six switchback designs, consider three average interval lengths: 28, 56, and 112 minutes

Procedure of synthetic experiments

- Randomly select one experiment-market pair and use its two-week historical experimental data
- Randomly draw one CEC from the empirical distribution of CECs
- Given a design, use CEC to calculate the cumulative effect at every time point
- Add the cumulative effect to the two weeks of historical data to obtain synthetic experimental data
- Estimate GATE

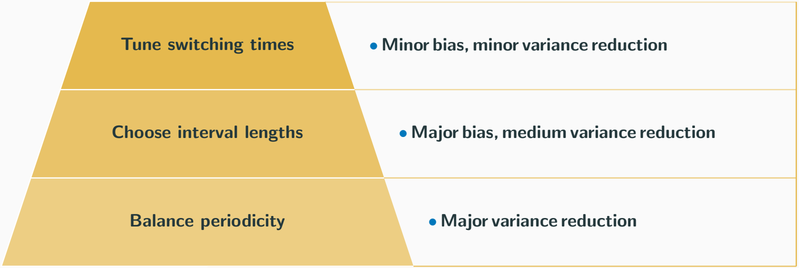
Repeat this procedure for 500 times

Estimation errors in synthetic experiments



Balanced Poisson duration switchback with an average interval length of 112 minutes reduces MSE by 33% compared to the status quo design

Pyramid of effectiveness of design principles



Hierarchical structure in the effectiveness of design principles in the reduction of MSE in the case study on ride-sharing platform

Related literature (incomplete list)

Most closely related to the recent literature on switchback designs

- Bojinov, Simchi-Levi, Zhao 2023; Hu and Wager 2022; Masoero et al. 2023; Ni, Bojinov, Zhao 2023; Li, Johari, Wager, Xu 2023; Chen and Simchi-Levi 2023

We aggregate units to abstract away interference, but a growing literature directly tackles interference using novel experimental design ideas

- Ugander et al. 2013; Eckles et al. 2017; Holtz et al. 2023; Bajari et al. 2023; Johari, Li, Liskovich, Weintraub 2022; Wager and Xu 2021

Related to some other designs in time-based experiments

- Doudchenko et al. 2019, 2021; Abadie and Zhao 2021; Xiong, Athey, Bayati, Imbens 2023; Basse, Ding, Toulis 2023; Wu et al. 2022

Analysis of Switchback Design

Setup

Directly analyze event-level data: n events in total between time 0 and T

Event density: $f(t) : [0, T] \to \mathbb{R}^+$

- $f(t)$ can be periodic, capturing the seasonality of human behavior

Marketplace outcome: Y_t at time t, can be viewed as the average outcome of all users in the marketplace

Event outcome: $Y^{(i)} \in \{0, 1\}$ for event i occurred at time t_i, a noisy measurement of the marketplace outcome

$$Y^{(i)} = Y_{t_i} + \varepsilon^{(i)},$$

where the measurement error $\varepsilon^{(i)}$ has mean zero

Serial correlation in measurement errors of events close in time

$$\mathrm{Cov}(\varepsilon^{(i)}, \varepsilon^{(j)}) \neq 0 \qquad \text{for} \quad t_i \neq t_j ,$$

caused by external factors like weather, supply conditions, and traffic

Potential outcomes and GATE

With non-anticipating outcomes, the potential outcomes of the marketplace at time t is

$$Y_t(\boldsymbol{w}_t, \boldsymbol{w}_t^{\mathrm{s}})$$

- $\boldsymbol{w}_t \in \{w_u \in \{0,1\}, \forall u \in [0,t]\}$: a realization of $\boldsymbol{W}_t$, design of primary experiment from time 0 to t
- $\boldsymbol{w}_t^{\mathrm{s}} \in \{w_u^s \in \{0,1\}, \forall u \in [0,t]\}$: a realization of $\boldsymbol{W}_t^{\mathrm{s}}$, design of simultaneous experiment from time 0 to t
- One experiment is run simultaneously with the primary experiment, but generalizable to many simultaneous experiments

Estimand of primary interest: Global average treatment effect (GATE)

$$\delta^{\text{gate}} = \int \delta_t^{\text{gate}} f(t) dt$$

- $\delta_t^{\text{gate}} = Y_t(\boldsymbol{W}_t = 1_t, \boldsymbol{W}_t^{\mathrm{s}} = 0_t) - Y_t(\boldsymbol{W}_t = 0_t, \boldsymbol{W}_t^{\mathrm{s}} = 0_t)$
- $1_t \in \{w_u = 1, \forall u \in [0,t]\}$, 0_t defined analogously
- difference in average outcomes between when an intervention is deployed indefinitely versus when it is absent

Potential outcomes and GATE

With non-anticipating outcomes, the potential outcomes of the marketplace at time t is

$$Y_t(\boldsymbol{w}_t, \boldsymbol{w}_t^{\mathrm{s}})$$

- $\boldsymbol{w}_t \in \{w_u \in \{0,1\}, \forall u \in [0,t]\}$: a realization of $\boldsymbol{W}_t$, design of primary experiment from time 0 to t
- $\boldsymbol{w}_t^{\mathrm{s}} \in \{w_u^s \in \{0,1\}, \forall u \in [0,t]\}$: a realization of $\boldsymbol{W}_t^{\mathrm{s}}$, design of simultaneous experiment from time 0 to t
- One experiment is run simultaneously with the primary experiment, but generalizable to many simultaneous experiments

Estimand of primary interest: Global average treatment effect (GATE)

$$\delta^{\mathrm{gate}} = \int \delta_t^{\mathrm{gate}} f(t) dt$$

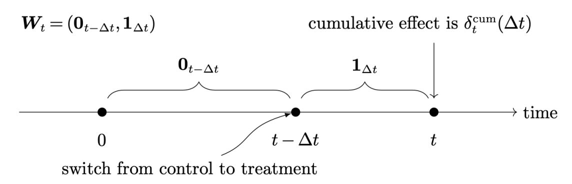
- $\delta_t^{\mathrm{gate}} = Y_t(\boldsymbol{W}_t = 1_t, \boldsymbol{W}_t^{\mathrm{s}} = 0_t) - Y_t(\boldsymbol{W}_t = 0_t, \boldsymbol{W}_t^{\mathrm{s}} = 0_t)$
- $1_t \in \{w_u = 1, \forall u \in [0,t]\}$, 0_t defined analogously
- difference in average outcomes between when an intervention is deployed indefinitely versus when it is absent

Cumulative effect

Cumulative effect at time t given the treatment is employed from time $t - \Delta t$ to t

$$\delta_t^{\text{cum}}(\Delta t) = Y_t(\boldsymbol{W}_t = (0_{t-\Delta t}, 1_{\Delta t}), \boldsymbol{W}_t^{\text{s}} = 0_t) - Y_t(\boldsymbol{W}_t = 0_t, \boldsymbol{W}_t^{\text{s}} = 0_t)$$

where $(0_{t-\Delta t}, 1_{\Delta t})$ concatenates $0_{t-\Delta t}$ and $1_{\Delta t}$



Post-experiment estimation

Horvitz-Thompson (HT) estimator for GATE

$$\hat{\delta}^{\text{gate}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{W_{t_i}Y^{(i)}}{\pi} - \frac{(1-W_{t_i})Y^{(i)}}{1-\pi}\right]$$

where $W_{t_i} \in \{0,1\}$ is the treatment status of event i occurred at time t_i. In this paper, $\pi = 1/2$ is the probability of being treated

- Use outcomes of events in treated intervals to estimate the average outcome under global treatment
- Use outcomes of events in control intervals to estimate the average outcome under global control

Design problem

Decision maker chooses the number of intervals M and the interval switching points, aiming to reduce the MSE of GATE

$$\mathbb{E}\left[\left(\hat{\delta}^{\text{gate}} - \delta^{\text{gate}}\right)^2\right]$$

Decomposition of bias and MSE

Assumptions

- Events are sampled i.i.d. from density function $f(t)$
- Carryover effect can be parametrized by a kernel function
- Each interval is treated with probability $1/2$, independently of other intervals, for both primary and simultaneous interventions

Decomposition of bias

$$\mathbb{E}\left[\hat{\delta}^{\text{gate}} - \delta^{\text{gate}}\right] = \text{Bias}(\mathcal{E}_{\text{carryover}}) + \text{Bias}(\mathcal{E}_{\text{simul}})$$

Decomposition of MSE

$$\mathbb{E}\left[(\hat{\delta}^{\text{gate}} - \delta^{\text{gate}})^2\right] = [\text{Bias}(\mathcal{E}_{\text{carryover}})]^2 + \text{Var}(\mathcal{E}_{\text{meas}})$$
$$+ \text{Var}(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}}) + \mathbb{E}[\mathcal{E}_{\text{simul}}^2] + 2\mathbb{E}[\mathcal{E}_{\text{simul}} \cdot (\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}})]$$

Role of balancing

Without balancing, $\mathrm{Var}(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}})$ scales with $\left(\Xi^{(m)} + 2\mu_{Y\text{ctrl}}^{(m)}\right)^2$

- $\mathrm{Var}(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}})$ arises from the randomness in treatment assignments
- $\Xi^{(m)}$: average treatment effect δ_t^{gate} in m-th interval
- $\mu_{Y\text{ctrl}}^{(m)}$: average control outcome $Y_t(0_t, 0_t)$ in m-th interval

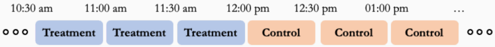
With balancing, $\mathrm{Var}(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}})$ scales with $\left(\Xi^{(m)}\right)^2$

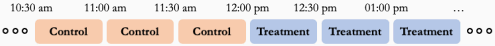
$\Rightarrow$ Balancing is particular effective when signal-to-noise ratio is low (the case in the ride-sharing platform)

Role of selecting switching lengths

Lengthening switching lengths reduces $\text{Bias}(\mathcal{E}_{\text{carryover}})$

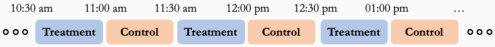
- $\text{Bias}(\mathcal{E}_{\text{carryover}})$ arises from using outcomes of events in treated and control intervals to approximate global treated and control outcomes

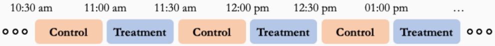




Shortening switching lengths reduces $\text{Var}(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}})$ and $\text{Var}(\mathcal{E}_{\text{meas}})$

- $\text{Var}(\mathcal{E}_{\text{meas}})$ arises from variance and covariance of measurement errors

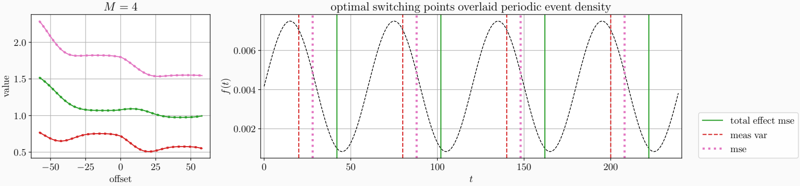




Role of tuning switching endpoints

Switches at high event density times reduces $\mathrm{Var}(\mathcal{E}_{\mathrm{meas}})$

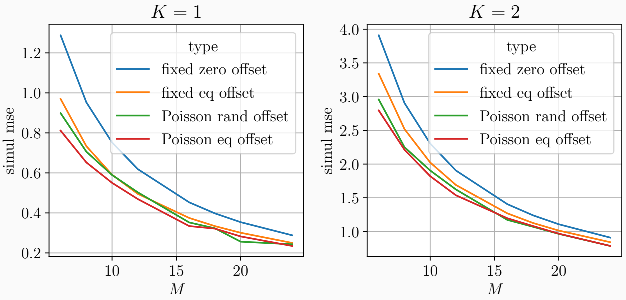
Switches at low event density times reduces $\mathrm{Bias}(\mathcal{E}_{\mathrm{carryover}})$



Role of offset parameter q in periodic event density (number of switches $M = 4$)

Role of tuning switching endpoints (cont.)

Poisson switchback with proper staggering switching points can be more effective than the fixed duration switchback



K denotes the total number of experiments running simultaneously; M denotes the number of switches

Conclusion

This paper studies the analysis and design of simultaneous switchback experiments in a highly generic setting

- Simultaneously capture the four realistic properties (carryovers, outcome covariance, event density, and simultaneous interventions)
- Continuous time framework and analysis of event outcomes
- Empirical Bayes design using historical data: A case study on a ride-sharing platform
- A decomposition of bias and MSE of the estimated GATE from any design
- A simulation study to explore the role of four realistic properties in affecting the MSE of heuristic designs

Supplementary

Expression of $\text{Bias}(\mathcal{E}_{\text{carryover}})$

The bias term $\text{Bias}(\mathcal{E}_{\text{carryover}})$ in the bias decomposition equals to

$$\text{Bias}(\mathcal{E}_{\text{carryover}}) = \sum_{m=1}^{M} I^{(m)} - \delta^{\text{co}},$$

where $I^{(m)}$ is the integrated carryover effect of treatments in $\mathcal{I}_m$ on outcomes in the same interval

$$I^{(m)} = \int_{t \in \mathcal{I}_m} \left[\delta_t^{\text{co}} \int_{t' \in \mathcal{I}_m} d_t^{\text{co}}(t') f(t') dt' \right] f(t) dt$$

$\delta_t^{\text{co}}(\boldsymbol{w}_t)$ is a carryover kernel that measures the intensity of the effect of intervention ℓ at time t' on the outcome at time t, so that the average carryover effect equals

$$\delta_t^{\text{co}}(\boldsymbol{w}_t) = \delta_t^{\text{co}} \cdot \int w_{t'} \cdot d_t^{\text{co}}(t') f(t') dt'$$

$\Rightarrow$ $\text{Bias}(\mathcal{E}_{\text{carryover}})$ increases with M

Expression of $\text{Var}(\mathcal{E}_{\text{meas}})$

The variance term $\text{Var}(\mathcal{E}_{\text{meas}})$ in the MSE decomposition equals to

$$\text{Var}(\mathcal{E}_{\text{meas}}) = 4 \sum_{m=1}^{M} \left(V^{(m)}/n + C^{(m)} \cdot (n-1)/n \right) ,$$

where $V^{(m)}$ measures the variance of measurement error of any event in $\mathcal{I}_m$ and is defined as

$$V^{(m)} = \int_{t_i \in \mathcal{I}_m} \mathbb{E}_\varepsilon \left[(\varepsilon^{(i)})^2 \mid t_i \right] f(t_i) dt_i ,$$

and $C^{(m)}$ measures the covariance of measurement errors of any two events in $\mathcal{I}_m$ defined as

$$C^{(m)} = \int_{t_i, t_j \in \mathcal{I}_m} \mathbb{E}_\varepsilon \left[\varepsilon^{(i)} \varepsilon^{(j)} \mid t_i, t_j \right] f(t_i) f(t_j) dt_i dt_j$$

⇒ As $n \to \infty$, $C^{(m)}$ dominates, $C^{(m)} = O(1/M^2)$ and $\text{Var}(\mathcal{E}_{\text{meas}}) = O(1/M)$ that decreases with M

Expression of $\mathsf{Var}(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}})$

The variance term $\mathsf{Var}(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}})$ in the MSE decomposition equals to

$$\mathsf{Var}(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}}) = \sum_{m=1}^{M} \left(\Xi^{(m)} + 2\mu_{Y^{\text{ctrl}}}^{(m)} \right)^2 + \sum_{m=1}^{M} \sum_{m' \neq m} \left(\left[I^{(m,m')} \right]^2 + I^{(m,m')} I^{(m',m)} \right),$$

where $\Xi^{(m)}$ is integrated total treatment effect in $\mathcal{I}_m$ and defined as

$$\Xi^{(m)} = \int_{t \in \mathcal{I}_m} \delta_t^{\text{gate}} f(t) dt$$

$\mu_{Y^{\text{ctrl}}}^{(m)}$ is the integrated global control outcome in $\mathcal{I}_m$ and defined as

$$\mu_{Y^{\text{ctrl}}}^{(m)} = \int_{t \in \mathcal{I}_m} Y_t(0_t, 0_t) f(t) dt$$

$I^{(m,k)}$ is the integrated carryover effect of treatments in $\mathcal{I}_k$ on outcomes in $\mathcal{I}_m$

$$I^{(m,k)} = \int_{t \in \mathcal{I}_m} \left[\delta_t^{\text{co}} \int_{t' \in \mathcal{I}_k} d_t^{\text{co}}(t') f(t') dt' \right] f(t) dt$$

$\Rightarrow$ The first part decreases with M while the second part increases with M

Expression of $\mathrm{Bias}(\mathcal{E}_{\mathrm{simul}})$

The bias term $\mathrm{Bias}(\mathcal{E}_{\mathrm{simul}})$ in the bias decomposition equals to

$$\mathrm{Bias}(\mathcal{E}_{\mathrm{simul}}) = \sum_{m=1}^{M} S^{(m)}$$

where $S^{(m)}$ measures the integrated bias from the treatment effects of simultaneous intervention

$$S^{(m)} = \int_{t \in \mathcal{I}_m} \Phi_t^{\mathrm{simul}} f(t) dt \,,$$

Φ_t^{simul} is defined as

$$\Phi_t^{\mathrm{simul}} = \mathbb{E}_{\boldsymbol{W}^{(-m)}} \left[\delta_t^{\mathrm{simul}}(\boldsymbol{W}^{(-m)}, W^{(m)} = 1) - \delta_t^{\mathrm{simul}}(\boldsymbol{W}^{(-m)}, W^{(m)} = 0) \right]$$

and $\delta_t^{\mathrm{simul}}(\boldsymbol{W}^{(-m)}, W^{(m)})$ is the expected treatment effects from the simultaneous intervention at time t, conditional on $\boldsymbol{W}^{(-m)}$ and $W^{(m)}$, formally defined as

$$\delta_t^{\mathrm{simul}}(\boldsymbol{W}^{(-m)}, W^{(m)}) = \delta_t^{\mathrm{simul}}(\boldsymbol{W}_t) = \mathbb{E}_{\boldsymbol{W}_t^{\mathrm{s}}} \left[Y_t(\boldsymbol{W}_t, \boldsymbol{W}_t^{\mathrm{s}}) - Y_t(\boldsymbol{W}_t, 0_t) \mid \boldsymbol{W}_t, t \right]$$

$\Rightarrow$ For the special case where the effects of main and simultaneous interventions are additive, $\mathrm{Bias}(\mathcal{E}_{\mathrm{simul}}) = 0$

Expression of $\mathbb{E}[\mathcal{E}^2_{\text{simul}}]$

The second-moment term $\mathbb{E}[\mathcal{E}^2_{\text{simul}}]$ in the MSE decomposition equals to

$$\mathbb{E}[\mathcal{E}^2_{\text{simul}}] = \sum_{m=1}^{M} \sum_{m'=1}^{M} S_{\text{var}}^{(m,m')}$$

For the special case where the effects of main and simultaneous interventions are additive,

$$\mathbb{E}[\mathcal{E}^2_{\text{simul}}] = \sum_{m=1}^{M} \left(\int_{t \in \mathcal{I}_m} \delta_t^{\text{s.gate}} f(t) dt \right)^2$$
$$+ \sum_{m=1}^{M} \sum_{m'=1}^{M} \left(\int_{t \in \mathcal{I}_m \cap \mathcal{I}^{\text{s}}_{m'}} \delta_t^{\text{s.inst}} f(t) dt + \int_{t \in \mathcal{I}_m, t' \in \mathcal{I}^{\text{s}}_{m'}} \delta_t^{\text{s.co}} d_t^{\text{s.co}}(t') f(t) f(t') dt dt' \right)^2$$

⇒ The first part decreases with M

⇒ The second part varies with how much $\mathcal{I}_m$ overlaps with $\mathcal{I}^{\text{s}}_{m'}$; staggering the switching times of different interventions reduces the second part

Expression of $\mathbb{E}[(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}}) \cdot \mathcal{E}_{\text{simul}}]$

The cross-term term $\mathbb{E}[(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}}) \cdot \mathcal{E}_{\text{simul}}]$ in the MSE decomposition equals to

$$\mathbb{E}[(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}}) \cdot \mathcal{E}_{\text{simul}}] = \sum_{m=1}^{M} \sum_{m'=1}^{M} S_{\text{cov}}^{(m,m')}$$

For the special case where the effects of main and simultaneous interventions are additive,

$$\mathbb{E}[(\mathcal{E}_{\text{inst}} + \mathcal{E}_{\text{carryover}}) \cdot \mathcal{E}_{\text{simul}}] = \sum_{m=1}^{M} \left(\Xi^{(m)} + 2\mu_{Y^{\text{ctrl}}}^{(m)} \right) \left(\int_{t \in \mathcal{I}_m} \delta_t^{\text{s.gate}} f(t) dt \right)$$

$\Rightarrow$ The covariance term decreases with M

Setup: Treatment design

A switchback experiment is run a geographically determined market between time 0 and T

Continuous time framework: naturally capture the stream of event data

Ex-ante, a treatment design for the whole experiment horizon is chosen $\boldsymbol{W} = \{W_t \in \{0,1\}, \forall t \in [0, T]\}$:

- $W_t = 1$: All users in the market are treated
- $W_t = 0$: Otherwise

Two-step design procedure:

- Partition the experimental horizon $[0, T]$ into M disjoint intervals
    - Endpoints of M intervals: $0 \leq t_0 \leq t_1 \leq \cdots \leq t_{M-1} \leq t_M = T$
    - m-th interval: $\mathcal{I}_m = [t_{m-1}, t_m]$
- Randomly choose the treatment assignment of each interval
    - Primarily focus on the case with $1/2$ treated probability

Summary of contributions

We study the design and analysis of switchback experiments, accounting for all four factors. For the design,

- Propose an empirical Bayes approach that uses knowledge from prior experiments to inform the design of new experiments
- Illustrate this approach through a case study on a ride-sharing platform
- The best-perform design randomizes and, on average, doubles the switching lengths than the status quo design with fixed lengths, yielding a 33% reduction in MSE

Summary of contributions

We study the design and analysis of switchback experiments, accounting for all four factors. For the analysis,

- A rigorous decomposition of bias and MSE of the estimated GATE
  Two sources of bias:
  - carryover effects from treatment at earlier times
  - confounding effects from simultaneous interventions

  Three sources of variance:
  - measurement errors of outcomes and their covariance
  - randomness of treatment assignments
  - randomness in event occurrence times
- Three design insights:
  - balancing periodicity reduces variance
  - switching less frequently reduces bias from carryovers, but increases variance from correlated outcomes
  - randomizing interval start and end points reduces both bias and variance from simultaneous experiments